Expected Returns and Large Language Models *

Yifei Chen
Booth School of Business
University of Chicago

Bryan Kelly
Yale University, AQR Capital Management, and NBER

Dacheng Xiu
Booth School of Business
University of Chicago

Abstract

We leverage state-of-the-art large language models (LLMs) such as ChatGPT and LLaMA to extract contextualized representations of news text for predicting stock returns. Our results show that prices respond slowly to news reports indicative of market inefficiencies and limits-to-arbitrage. Predictions from LLM embeddings significantly improve over leading technical signals (such as past returns) or simpler NLP methods by understanding news text in light of the broader article context. For example, the benefits of LLM-based predictions are especially pronounced in articles where negation or complex narratives are more prominent. We present comprehensive evidence of the predictive power of news on market movements in 16 global equity markets and news articles in 13 languages.

Keywords: natural language processing (NLP), large language models, BERT, GPT, LLaMA, ChatGPT, Bag-of-Words, Word2vec, machine learning, return prediction

*We benefited tremendously from discussions with seminar and conference participants at NYU Courant Institute, PBC School of Finance at Tsinghua University, Zhejiang University, Central University of Finance and Economics, 2023 China Fintech Research Conference, 5th Annual NLP and Machine Learning in Investment Management Conference, 2023 China International Conference in Finance, WorldQuant 2023, Federal Reserve Bank of Philadelphia, Shandong University, AAAI 2023 Summer Symposium, Workshop on Frontiers of Statistics and Data Science, NLP SoDaS conference 2023, Wolfe Research, Risk Day at ETH 2023, GSU-RFS FinTech Conference, Macquarie University, Australian National University, SQA/CQA Joint Webinar, Brazilian Stock Exchange, UIC College of Business and Banco de Portugal. We gratefully acknowledge the computing support from the Research Computing Center and data support from the Fama-Miller Center at the University of Chicago. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein.

1 Introduction

Economic text is constantly generated by human writers striving to understand and make predictions about economic phenomena. In recent decades, the finance literature has begun to extract information from certain text sources like the financial news press, regulatory filings, and social media. The research agenda of improving economic models through text mining remains in its earliest stages. Research has thus far examined only a limited portion of market-relevant textual data, often focusing on a single specialized data source at a time (e.g., the front page of *The Wall Street Journal*, or the "risk factor" section of 10-K filings). And for each data source, text information is often represented in rudimentary ways (e.g., as a dictionary-based sentiment score or as a "bag of words").

There are good reasons for the limited use of text data to date. Its lack of regular structure makes it far more difficult to work with than standard numeric data sets. Language is an extremely nuanced information encoding scheme. As a result, highly complex models are necessary to faithfully unearth information contained in text. But complex models are prohibitive for many researchers. Technological barriers to entry exclude researchers who lack the specialized skill sets necessary to operate such models. The high computational cost of complex models excludes other researchers who may possess requisite skills but face research funding constraints.

This means that recent textual analysis in finance and economics is the tip of the iceberg. Text is an underexploited data source for understanding asset markets. The challenges of textual analysis today portend an exciting research agenda tomorrow, in which economists gradually expand sourced text corpora and increasingly refine their ability to elicit information from that text.

In this paper we aim to take a step in this direction by constructing refined news text representations derived from large language models (LLMs) and then using these to improve models of expected stock returns. To better understand the role of LLMs, it is helpful to first grasp the current landscape in financial text mining. The most prevalent methods to date are supervised machine learning models that are customized to specific tasks such as forecasting returns (Ke et al. (2019); Jegadeesh and Wu (2013)), volatility (Manela and Moreira (2017)), or macroeconomic conditions (Kelly et al. (2018); Bybee et al. (2020)).

These analyses proceed in two general steps: a text representation step and an econometric modeling step. Step 1 decides on the numerical representation of the text data that will be passed to the Step 2 econometric model. The most common choice in the literature is "bag of words" (BoW), which collapses each document observation into a high dimensional vector of counts spanning all unique terms in the full corpus of documents. In some cases, the numerical representation stops here (e.g., Jegadeesh and Wu (2013); Kelly et al. (2018)). In other cases, the numerical representation is refined further. For example, Ke et al. (2019) reduce the BoW dimensionality from several tens of thousand to a few hundred terms with a correlation screening procedure to filter out irrelevant terms, and Bybee et al. (2020) reduce the dimensionality of counts with an unsupervised topic

model.[1] The output of Step 1 is a numerical data matrix X of dimension $D \times P$. Rows correspond to the D documents in the text corpus, and each row contains the P-dimensional numerical vector representation of those documents (e.g., P can be the number of terms in a BoW or the number of topics in a topic model). Step 2 uses X as data in an econometric model to describe some economic phenomenon (e.g., return, volatility, and macroeconomic modeling in the references above).

The financial text representations referenced above have some limitations. First, all of these examples begin from a BoW representation, which is overly simplistic and only accesses the information in text that is conveyable by term usage frequency. It sacrifices nearly all information that is conveyed through word ordering or contextual relationships between terms. Second, the ultra-high dimensionality of BoW representations leads to statistical inefficiencies—Step 2 econometric models must include many parameters to process all these terms despite many of the terms conveying negligible information. Dimension reductions like LDA and correlation screening are beneficial because they mitigate the inefficiency of BoW. However, they are derived BoW and thus do not avoid the information loss from relying on term counts in the first place. Third, and more subtly, the dimension-reduced representations are *corpus specific*. For example, when Bybee et al. (2020) build their topic model, the topics are estimated only from *The Wall Street Journal*, despite the fact that topics are general language structures and can be better inferred by using additional text outside of their sample.

Enter the concept of an LLM. LLMs are trained on large text data sets that span many sources and themes. The idea of a LLM is for a specialized research team to perform the Herculean feat of estimating a general purpose language model with astronomical parameterization on truly big text data. LLMs have billions of parameters and are trained on many millions of documents (including huge corpora of complete ebooks, all entries of Wikipedia, and more). But for each LLM, this feat is performed *just once*, then the estimated model is made available for distribution to be deployed by non-specialized researchers in downstream tasks.

In other words, the LLM delegates Step 1 of the procedure above to the handful of people in the world that can best execute it. A Step 2 econometric model can then be built around LLM output. Like LDA (or even BoW), the output of an LLM is a numerical vector representation (or "embedding") of a document. A non-specialized researcher obtains this output by feeding the document of interest through software (which is open-source in many cases). Therefore, an LLM model in Step 1 delivers a numerical matrix X just like the examples above, making it seamless to integrate into Step 2 with little or no modification. The main benefit of an LLM in Step 1 is that it provides more sophisticated and well-trained text representations than used in the literature referenced above. This benefit comes from the expressivity of massive nonlinear model parameterizations and from training on extensive language examples across many domains and from throughout human history. The

[1]Specifically, they employ latent dirichlet allocation (LDA) which can be thought of as a multinomial principal components estimator. This collapses their roughly 20,000-dimensional term count representation for each document to a 180-dimensional topic attention representation.

transferability of LLMs make this unprecedented scale of knowledge available for finance research.

Our primary research contribution revolves around showcasing the advantages of LLM representations for effectively modeling stock returns. In addition, we compare the performance of LLMs with supervised machine learning models commonly used in the extant finance literature. To achieve this, we undertake two distinct econometric exercises that harness the power of text mining in understanding the financial market. The first exercise involves sentiment analysis, where we extract sentiment information from financial news text and examine how this information is incorporated into the dynamics of stock returns. In the second exercise, we directly leverage the predictive power of financial news text to model the short-term cross-section of expected stock returns.

We study three large-scale pre-trained LLMs: BERT (developed by Google), RoBERTa (by Meta), LLaMA(LLaMA2) (by Meta). Additionally, we also obtain embeddings from OpenAI embedding model "text-embedding-3-large" with API provided by OpenAI. We compare this with SESTM, a sentiment analyzer based on BoW representation and trained on task-specific text data (developed by Ke et al. (2019)). We also study two other word-based models, Word2vec (a word-vector representation framework developed by Google), and Loughran-McDonald Master Dictionary (LMMD). The inputs to our modeling framework are global news text data from Refinitiv in their Thomson Reuters Real-time News Feed (RTRS) and Third Party Archive (3PTY) databases from January 1996 to June 2019. We merge this with individual stock data from CRSP (for US stocks) and Datastream-EIKON (for international stocks).

We find the following main empirical results. First, econometric models that use pre-trained LLM embeddings outperform prevailing text-based machine learning return predictions. This is best summarized in terms of out-of-sample trading strategy performance. A quintile spread long-short strategy that buys stocks with high foundation-based return forecasts and sells those with low forecasts earns an annualized Sharpe ratios of 3.60, 3.75, 3.89 (4.16) and 4.62 based on BERT, RoBERTa, LLaMA (LLaMA2) and OpenAI-based (a.k.a ChatGPT in the rest of the paper) models, respectively (gross of trading costs). All of these significantly outperform the corresponding strategy based on word-embedding forecasts, which earns an annualized Sharpe ratio of 3.43, 3.06 and 2.29 for SESTM, Word2vec and LMMD, respectively.

Furthermore, we delve into the analysis of the impact of news recency on the relative performance of different models. By focusing on articles labeled as "news alerts" by Refinitiv, we observe that returns remain predictable for significantly longer horizons compared to unflagged articles. Surprisingly, despite the brevity of news alerts, which often consist of only headlines, we find that the distinction between LLMs and word-based models becomes less pronounced. This suggests that the advantages of speed can overshadow differences in language model capacity when it comes to predicting returns based on recent news. In essence, a simple representation of recent news can yield comparable performance to more sophisticated representations of older news. However, as time elapses and the predictive information within the text gradually diminishes, the benefits of employing sophisticated models become comparatively more crucial.

We subsequently demonstrate the relationship between the complexity of LLMs and their performance in predicting returns. By employing a series of LLaMA models characterized by an escalating number of parameters, we illustrate that larger models typically surpass their smaller counterparts in terms of investment performance. The Sharpe ratios yielded by LLMs exhibit greater magnitudes, yet this improvement reaches a saturation point once the number of parameters exceeds 13 billion. This result suggests that while more complex LLMs possess enhanced capabilities in processing text, there is a limit to the benefits gained from increasing their complexity in terms of return prediction.

Our primary findings are derived from the US equity market, focusing on news articles written in English. Additionally, we analyze 16 international stock markets using news articles written in 12 other languages, including Chinese, Japanese, German, Italian, French, Swedish, Danish, Spanish, Finnish, Portuguese, Greek, and Dutch. As a preliminary contribution, we extend the analysis of Ke et al. (2019) to international text. We find similar SESTM performance globally to that documented by Ke et al. (2019) for the US sample. We also find that the general LLMs can on average outperforms SESTM.

The rest of the paper is organized as follows. Section 2 introduces the LLMs and other approaches we compare. Section 3 presents an empirical analysis of stock-level news and return prediction in US and international markets using these methods. Section 4 concludes. The appendix provides additional tables and figures.

2 The Text Mining Framework

2.1 A Tale of Two Objectives

We employ a supervised approach to mine news text with two primary objectives. The first objective involves sentiment analysis, which entails assessing the tone of a news article. The second objective focuses on predicting the cross-section of returns within a short horizon.

While both sentiment analysis and return prediction illuminate the statistical correlation between news text and returns, they are but components of a broader narrative. We aim to develop trading strategies that can efficiently translate these statistical correlations into profitable investments. The economic impact of these gains serves as a formidable challenge to the efficient market hypothesis.

The efficient market hypothesis, our null hypothesis, posits that expected returns are primarily driven by unpredictable news which is rapidly, and in the most extreme cases instantaneously, assimilated into prices. On the other hand, our research presents an alternative hypothesis: the information contained within news text is not immediately and completely integrated into market prices. This delay might be attributed to factors such as limits-to-arbitrage and rational inattention, suggesting a predictive capacity of news text for future asset price trends.

While the adoption of this alternative hypothesis is largely uncontroversial, its profound significance cannot be overstated. Our predictive analysis provides novel insights into the velocity and

magnitude of deviations from the efficient market hypothesis, furnishing fresh evidence to the pool of empirical studies examining this alternative hypothesis.

Sentiment analysis is commonly treated as a classification problem in machine learning. The primary aim is to delineate the relationship between specific text-based features, denoted as $x_{i,t}$, and their associated sentiment labels such as positive or negative, denoted by a binary variable $y_{i,t}$, based on a set of training articles.

The equation below posits the relationship between these labels and features:

$$\mathrm{E}(y_{i,t}|x_{i,t}) = \sigma(x'_{i,t}\beta). \tag{1}$$

In this context, $\sigma(x)$ is a logistic link function, represented as $\sigma(x) = \exp(x)/(1+\exp(x))$. This function has been specifically designed to convert the features into a value ranging from $[0, 1]$, thereby standardizing the quantification of sentiment. This method enables us to derive a sentiment score for any article of the testing sample. The sentiment score quantifies the tone of an article: a score closer to one denotes a stronger positive sentiment.

To accomplish this, we require a sentiment label for each article in the training sample. Each of our news articles is tagged with a corresponding stock and includes a timestamp that records the timing of the news event. Drawing from the methodology presented in Ke et al. (2019), we employ the sign of the stock's return, registered in close temporal relation to a news event, to assign a binary sentiment label (either positive or negative) to the relevant article. Although this label is inherently noisy, it is a simple and convenient alternative to manual labeling by expert readers. The empirical analysis conducted by Ke et al. (2019) demonstrates the effectiveness of this approach in measuring news sentiment and its robustness across different return definitions.

Recognizing that news articles often report on events from previous days, we create sentiment labels based on three-day returns, following Ke et al. (2019). This process involves analyzing returns from the day before the article's publication up to the day after. This approach improves the signal-to-noise ratio in sentiment labeling, leading to greater accuracy in the sentiment score — a key goal in sentiment analysis which is to establish a meaningful connection between text and score. It is crucial to note that these three-day returns are utilized solely for in-sample training, thereby avoiding any look-ahead bias when generating sentiment scores for articles within the testing sample.

Sentiment analysis does not directly provide a measurement of expected returns; instead, it merely reflects the tone present in news articles. Hence, we turn to a distinct approach for modeling expected returns, or stated differently, for predicting returns, to examine the extent to which information in news drives the short-term cross-sectional variation of expected returns. The simplest prediction model involves a standard panel regression. The regression equation translates article features, $x_{i,t}$, directly into the corresponding stock's expected return, $\mathrm{E}(r_{i,t+1})$, for the next period:

$$\mathrm{E}(r_{i,t+1}|x_{i,t}) = x'_{i,t}\theta. \tag{2}$$

Inspired by the empirical analysis of Gu et al. (2020), we train this model by collectively considering the next-period returns across all stocks and time periods within our training sample. We can evaluate the effectiveness of our model by assessing its predictive performance in subsequent testing samples.

This pooled panel regression model allows us to represent expected returns as a linear combination of article level features. Similarly, our sentiment model represents the probability of a positive label through a sigmoid function of linear combinations of these features. It is important to note that linearity in model specification is not necessarily the optimal choice. Alternative approaches, such as incorporating neural networks or other nonlinear architectures on top of the $x_{i,t}$s, are certainly viable. However, in order to emphasize the significance of text-based representations and highlight the role they play, we intentionally refrain from introducing such complex nonlinear models. This decision allows us to focus on the simplest model and emphasize the impact of text-based representations.

In the subsequent sections, we elucidate the procedure involved in deriving textual features $x_{i,t}$ from news text through various methodologies, initiating with cutting-edge LLMs in the domain of NLP. This stage is recognized as feature engineering within the parlance of machine learning. Following this, we will present several alternative methods that were proposed prior to the advent of the LLM era.

2.2 Large Language Models

LLMs represent an innovative approach within the Artificial Intelligence (AI) sphere, first gaining prominence within NLP. This methodology comprises a set of deep learning models, characterized by extensive parameterization and training on expansive datasets.

Distinguishing features of this paradigm pivot around a unique training process devoid of labeled data. Instead, it relies on self-supervised learning techniques. This involves randomly masking words within a text and predicting the masked terms, or through unsupervised language modeling, where the model maximizes the probability of predicting the subsequent sentence based on the current one. Once trained, these LLMs exhibit a remarkable capacity for transfer learning, a process by which the "knowledge" acquired from one task is applied to different tasks. This characteristic enhances their versatility and broadens their applicability across diverse domains.

State-of-the-art LLMs have been dominating performance benchmarks across various NLP tasks, primarily due to their expansive scale. They are often pre-trained on enterprise-level platforms by Google, OpenAI, Meta, etc, some of which have made their pre-trained models publicly available. Our work incorporates three distinct LLMs as benchmarks — Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. (2018)), Robustly Optimized BERT Pre-training Approach (RoBERTa) (Liu et al. (2019)), and Large Language Model Meta AI (LLaMA) (Touvron et al. (2023))

BERT holds a historical significance in the annals of LLMs and NLP as it marked a crucial shift in the creation and application of language models. Before BERT, models predominantly relied on

unidirectional or superficially bidirectional understanding of text. BERT brought about a revolution with its deeply bidirectional model, allowing a contextual understanding from both preceding and succeeding words for prediction. This change sparked an influx of research and development in NLP, yielding more advanced models like GPT-2, GPT-3, and RoBERTa. As such, we adopt BERT as our initial benchmark model.

On the other hand, RoBERTa, an offshoot of BERT, was developed by Meta AI. The goal was to augment BERT's performance by modifying its training regimen. Although it shares the foundational architecture with BERT, the variations in the pre-training phase lend RoBERTa its distinct identity. These changes led to substantial performance enhancements, allowing RoBERTa to outdo BERT in numerous NLP benchmark tests. Yet, the question remains whether a model's proficiency in NLP tasks unequivocally translates to stronger performance in investment scenarios. This intriguing query is what our empirical analysis will shed light on.

LLaMA, developed by Meta AI, is another LLM that we consider in our analysis. It has been trained on a wide array of text data, including books, articles, and encyclopedias, and is designed to generate embeddings for various NLP tasks such as sentiment analysis and text classification. LLaMA is available in multiple versions with varying capacities, including LLaMA1 and LLaMA2, the latter being the more advanced iteration. These versions include models with 7 billion, 13 billion, and 33 billion parameters for LLaMA, and 7 billion, 13 billion, and 70 billion parameters for LLaMA2. We specifically utilize the LLaMA2 with 13 billion parameters and LLaMA with 13 billion parameters as benchmarks in our study.

One remarkable instance of an LLM is ChatGPT, a chatbot that swiftly garnered global recognition. ChatGPT was engineered based on the GPT-3.5 architecture. The initial breakthrough of the GPT model was made by Radford et al. (2018), who introduced a computational framework comprising 117 million parameters. This was subsequently enhanced by Radford et al. (2019) with the introduction of GPT-2, a more robust model featuring a staggering 1.5 billion parameters. Following this, GPT-3 was unveiled in Brown et al. (2020), which saw the model grow to more than tenfold the size of GPT-2. Although the most advanced GPT models have not been publicly released, OpenAI provides an API designed for generating embeddings using various models. We send our text strings to the embedding API endpoint and, in return, receive an embedding—a list of floating-point numbers that numerically represents the textual information. For our analysis, we used the "text-embedding-3-large" model from this API, referred to as ChatGPT in subsequent sections.

We now turn to details of our implementation from tokenization to feature construction for LLMs.

2.2.1 Tokenization

In any LLM framework, the starting point of a contextualized representation is tokenization. The smallest component of an article is known as a token. Tokens can manifest as characters, words, or subwords, each representing different forms of tokenization. Within LLMs, tokens typically take

the form of subwords. Word-based tokenization, which partitions text into individual words based on specific delimiters, is most prevalent. LLMs implement similar tokenization algorithms, which effectively divide rare words into smaller, meaningful subwords. This method of subword tokenization helps to alleviate data sparsity, enabling token reuse and subsequently boosting their frequency of occurrence. Furthermore, it allows for the maintenance of a manageable vocabulary size. This is particularly beneficial given the vast array of different words, or surface forms, present in most languages, especially those that are morphologically rich.[2]

Here's an example from a piece of news regarding Apple: *"The Company also admitted that in addition to macroeconomics in the Chinese market, the price cuts to battery replacements a year earlier to fix the Company's prior surreptitious conduct had hurt iPhone sales."* The BERT tokenizer, which utilizes WordPiece encoding, breaks down this sentence into a sequence of ordered tokens, totaling 43: *'the', 'company', 'also', 'admitted', 'that', 'in', 'addition', 'to', 'macro', '##economic', '##s', 'in', 'the', 'chinese', 'market', ',', 'the', 'price', 'cuts', 'to', 'battery', 'replacements', 'a', 'year', 'earlier', 'to', 'fix', 'the', 'company', '", 's', 'prior', 'sur', '##re', '##pt', '##iti', '##ous', 'conduct', 'had', 'hurt', 'iphone', 'sales', '.'*. In particular, the relatively rare word '*macroeconomics*' is broken down into three tokens, and '*surreptitious*' into five.

While RoBERTa employs the same architectural framework as BERT, it opts for byte-level Byte-Pair Encoding (BPE) for tokenization, akin to GPT-2 as presented by (Radford et al. (2019)). The use of byte-level tokenization enables RoBERTa to more effectively manage out-of-vocabulary words. Notably, LLaMA and LLaMA2 also adopt the same BPE tokenizer. Regarding the above example, the BPE tokenizer yields a total of 41 tokens, including punctuations:[3] *'The', 'ĠCompany', 'Ġalso', 'Ġadmitted', 'Ġthat', 'Ġin', 'Ġaddition', 'Ġto', 'Ġmacro', 'econom', 'ics', 'Ġin', 'Ġthe', 'ĠChinese', 'Ġmarket', ',', 'Ġthe', 'Ġprice', 'Ġcuts', 'Ġto', 'Ġbattery', 'Ġreplacements', 'Ġa', 'Ġyear', 'Ġearlier', 'Ġto', 'Ġfix', 'Ġthe', 'ĠCompany', ''s", 'Ġprior', 'Ġsur', 're', 'pt', 'itious', 'Ġconduct', 'Ġhad', 'Ġhurt', 'ĠiPhone', 'Ġsales', '.'*. Similarly, '*macroeconomics*' and '*surreptitious*' are split into multiple tokens.

Both WordPiece and Byte-Pair encoding methods can handle words that are not in their initial vocabulary by breaking them down into smaller, known pieces. In practice, the differences in performance between these two methods tend to be relatively small.

2.2.2 Transformer Architecture

The fundamental architecture of LLMs is rooted in a novel encoder design of deep neural networks, known as the transformer, which was introduced by Vaswani et al. (2017). The transformer encoder maps tokens into vector form, utilizing a series of attention layers, as conceptualized by Bahdanau et al. (2014) and Luong et al. (2015). This enables the modeling of token dependencies, irrespective

[2]BERT has 30K tokens, LLaMA(LLaMA2) has 32k tokens, and RoBERTa uses about 50K. In contrast, some delimiter based tokenization may result in a vocabulary size over 250K.

[3]A character 'Ġ' is automatically added to represent the space before a word in the original input sentence.

of their respective positions in the input sequence. By implementing this technique, the traditional recurrent structure, which plays a crucial role in Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) for sequence processing, is effectively eliminated. Notably, the transformer approach bypasses limitations associated with parallelization and memory constraints, hence enhancing the model's scalability.

Although LLMs share core principles, their deep learning architectures differ, each tailored for specific NLP tasks. BERT and RoBERTa employ a bidirectional encoder, which generates contextual representations of tokens by considering both preceding and succeeding instances of their appearances. This architecture proves highly effective for tasks such as sentiment analysis and natural language understanding. On the other hand, LLaMA is a family of autoregressive large language models, which solely incorporates a decoder that translates vector embeddings into tokens, making it ideal for applications involving human-like conversation and natural language generation. Consequently, LLaMA's contextualized embeddings are obtained from a unidirectional architecture.

In our specific context, the distinction between encoder and decoder networks has minimal significance. Both networks possess the capability to generate contextualized embeddings for each token in an input article. These embeddings consider not only the individual tokens themselves but also their respective positions within the article. These contextualized embeddings serve as the fundamental inputs to our procedure for modeling returns. Our empirical analysis reveals that, in terms of the architecture distinction, the impact is relatively minor compared to the complexity of the models, primarily determined by the number of parameters they possess.

2.2.3 Pre-training and Fine-tuning

The training step in this transfer learning context is often termed *pre-training*, serving as a means to an end. This step involves learning about a large number of model parameters (e.g., millions or even trillions) from an extremely large and diverse dataset (e.g., Wikipedia, Common Crawl, WebText). This process allows the model to understand the syntax and semantics of the language, including learning the meanings of words, how they are used in different contexts, and the general structure and grammar of the language.

BERT's pre-training process involves two parallel unsupervised tasks: the Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the MLM task, 15% of tokens within the input sequence are randomly hidden, with the model then working to predict these obscured tokens. This MLM goal allows token representation to integrate context from both left and right directions. As implied by its name, NSP attempts to determine if two sentences are consecutive or not.

RoBERTa refines BERT's pre-training by omitting the NSP task and modifying the hyper-parameters, including an extension in training duration, expansion of training data (from 16GB to 160GB of uncompressed texts), an increase in batch-training sizes (from 256 to 8k), and elongation of training sequences. Moreover, while BERT produces static masks just once during data

preprocessing, RoBERTa improves upon this by generating masking patterns every time a sequence is input during training. This enhancement supports training across more steps and accommodates larger datasets.

The specific task used for pre-training LLaMA is called "next token prediction" or "autoregressive language modeling." The model is trained to predict the next word in a sentence given all of the previous words. It does this by learning to understand the context provided by the previous words and using that context to predict the most likely next word. For example, given the input "The Fed raised interest", the model might learn to predict the word "rate" as the next word. The model's parameters are updated during pre-training to minimize the difference between the predicted words and the actual words in the training data.

The adaptation stage, also known as the fine-tuning stage, follows pre-training in the development of LLMs. After the LLM has been pre-trained on a massive corpus of data to understand language in a broad and general sense, it is then adapted or fine-tuned before its deployment to a specific task. This fine-tuning involves training the model on a smaller, task-specific dataset. The tasks can be diverse, such as text classification, sentiment analysis, question answering, summarization, and more. Unlike pre-training, which is unsupervised, fine-tuning is a supervised learning process, as it uses labeled data specific to the task at hand. During fine-tuning, the model's parameters, which were learned during pre-training, are updated to optimize the model's performance on the specific task. This process is usually faster and requires less data than pre-training because the model has already learned a lot of the necessary language understanding during pre-training. In this way, pre-training provides the model with general language understanding, while fine-tuning adapts that understanding to the specific requirements of a task. This two-stage training process has proven to be very effective for building LLMs that perform well across a wide range of NLP tasks.

Drawing inspiration from the work of Peters et al. (2018), we employ a feature extraction approach, also known as probing, by directly utilizing the pre-trained parameters to generate features associated with text data for downstream tasks. To be more specific, we input a new article into the pre-trained model, which results in each token within the article being represented as a vector. These vector representations effectively capture the contextual essence of the tokens. These representations are then utilized in subsequent downstream tasks for further exploration and application. While we can perceive the subsequent training stage as a form of fine-tuning, it is important to note that we do not update any parameters produced during the pre-training stage. This streamlined approach minimizes computational efforts, making it easier to replicate our empirical analysis. By adopting this feature extraction approach, we leverage the power of pre-trained models to extract meaningful features from text data without the need for extensive retraining. This efficient process allows us to focus on the downstream tasks at hand while benefiting from the comprehensive contextual understanding encapsulated in the pre-trained parameters.

2.2.4 Article-level Representations

Ultimately, we attempt to construct article level representations, $x_{i,t}$, for subsequent classification and regression tasks. LLMs like BERT and RoBERTa can process input sequences of up to 512 tokens, translating these tokens into a 1024-dimensional vector representation. In contrast, LLaMA(LLaMA2) can manage sequences as long as 2,048(4,096) tokens and embed each token into a 5,120-dimensional space.[4] In scenarios where an article exceeds the upper token limit, we focus solely on the initial segment up to that upper token boundary. Approximately 60% of US news articles comply with this length restriction. Subsequent empirical analyses suggest that the preliminary 512 tokens effectively encapsulate necessary information for return prediction. After acquiring vector representations for each token, we calculate the vector average across all tokens within an article. The resulting vector is then utilized to represent the entire article's information. Although we employ the mean of all token embeddings to derive an article-level embedding, alternate methods could be considered. For example, it's a common practice to use the embedding of the first token (often referred to as the CLS token) in BERT and RoBERTa, or the last token in LLaMA(LLaMA2), for downstream classification tasks. We have opted for the mean, as it constitutes a reasonable approach for other models like Word2vec, to which we draw comparisons.

2.2.5 Other Fine-Tuned BERT and Multi-Language BERT Models

Several open-source BERT models are available for various tasks. For instance, Araci (2019) fine-tuned a BERT model for a classification task based on the Financial PhraseBank dataset collected by Malo et al. (2014). This dataset includes roughly 5,000 labeled sentences, divided into three categories: positive, neutral, and negative. In a separate work, Yang et al. (2020) pre-trained a different BERT model based on financial communication text. This included Corporate Reports 10-K and 10-Q (comprising 2.5 billion tokens), Earnings Call Transcripts (1.3 billion tokens), and Analyst Reports (1.1 billion tokens), amounting to a total of 4.9 billion tokens in corpus size. This model was later fine-tuned using 10,000 manually annotated sentences (categorized as positive, negative, neutral) from analyst reports. Although this model was pre-trained with data highly relevant to the financial context, it does not leverage the expansive corpus the original BERT was trained on. Due to space constraints, we only provide comparison results based on Yang et al. (2020)'s FinBERT, as our empirical analysis suggests that it surpasses the performance of the model by Araci (2019) (detailed findings not reported).

Beyond English, BERT has been pre-trained with multilingual datasets, enabling its application in the analysis of other languages. Moreover, XLM-RoBERTa, as presented by Conneau et al. (2020) is a multilingual adaptation of RoBERTa. It was pre-trained on 2.5TB of filtered CommonCrawl data encompassing 100 languages. We utilize XLM-RoBERTa large as an extension of RoBERTa in

[4]Specifically, we chose BERT large, RoBERTa large, and LLaMA(LLaMA2) (13 billion) as our benchmark set of LLMs. Their total parameters are, respectively, 0.345B, 0.354B, and 13B for BERT, RoBERTa, and LLaMA(LLaMA2).

the analysis of non-English languages.

2.3 Word Embeddings

LLMs have evolved beyond an earlier text embedding paradigm that focused on learning morphological word representations as vectors. This progression is rooted in the principles of distributional semantics, as postulated by Harris (1954) and Firth (1957). According to their distributional hypothesis, a word is characterized by the context in which it appears. This idea has been utilized to represent word meanings as vectors, thereby encapsulating semantic similarity in terms of vector similarity. This approach allows for the creation of contextualized embeddings of words in a semantic vector space, capturing the nuanced meaning shifts induced by the contextual environment.

The concept of learning continuous representations of words has a deep-rooted history in NLP, tracing back to the work of Rumelhart et al. (1986). More recently, Mikolov et al. (2013) proposed a simplified approach called Word2Vec that generates high-dimensional vectors on very large corpora. In their work, Mikolov et al. (2013) presented two distinct neural network architectures: the Continuous Bag-Of-Words (CBOW) and the Skip-Gram models. The CBOW model predicts the current word based on its context, excluding the word itself. Conversely, the Skip-Gram model predicts the surrounding words given the current word. An illustrative example by Mikolov et al. (2013) that showcases the efficacy of this approach is the operation vector("King") - vector("Man") + vector("Woman"). Interestingly, the resulting vector from this operation aligns most closely with the vector representation of the word "Queen."

In contrast to LLMs that operate directly on input sequences of variable lengths (up to 512 tokens), Word2Vec employs a fixed-size context window for each word, typically encompassing 5 or 10 words around the current word. This approach limits its capacity to capture contextual information that extends beyond this window. Additionally, Word2Vec is built on a two-layer neural network architecture, a significantly less complex structure compared to the extensive deep neural network architectures employed by foundational models.

For our purposes, we downloaded pre-trained word vectors for English and other languages from fastText, an extension of Mikolov et al. (2013)'s Skip-gram model.[5] For English word vectors, we select the model *wiki-news-300d-1M* by Mikolov et al. (2018). This model contains 1 million word vectors trained on Wikipedia 2017, the UMBC webbase corpus, and the statmt.org news dataset, incorporating a total of 16 billion tokens. For non-English languages, the word vectors were trained on Common Crawl and Wikipedia data (Grave et al. (2018)). All these word vectors are 300-dimensional. As we do with LLMs, we calculate the average of all word vectors within a news article to derive the article-level embedding, which is subsequently fed into downstream regressions as features.

[5]FastText, developed by Bojanowski et al. (2017). was chosen due to its multilingual support. Although we adopted the fastText package, we consistently use the term "Word2Vec" for ease of understanding.

2.4 Bag-of-Words

The Bag-of-Words (BOW) model, initially proposed by Harris (1954), represents an article as a vector of word frequencies. This representation takes into account the occurrence and frequency of words, but neglects grammar, word order, and the broader context.

We adhere to the SESTM approach proposed by Ke et al. (2019), which utilizes a structured sentiment model for BOW representations. This method is comprised of three steps. The first step identifies a list of terms (either unigrams or bigrams) most closely correlated with sentiment through a screening process. The second step assigns weights to these words by estimating a topic model. Finally, the third step aggregates these terms into an article-level sentiment score through penalized likelihood estimation. The simplicity, transparency, and theoretical soundness of this approach make it an appropriate BOW benchmark for our purposes.

In addition to the SESTM approach, we incorporate the Loughran-McDonald Master Dictionary (LMMD) for financial sentiment analysis. LMMD, first proposed by LOUGHRAN and MCDONALD (2011), specifically designed for financial contexts, is instrumental in accurately identifying and scoring financial terms, thus enhancing the precision of sentiment analysis in financial news.

Among the methods we consider, the SESTM stands out for its simplicity and transparency, but it falls short in accounting for contextual information. On the other hand, LLMs offer the capability to model intricate token connections in natural languages, albeit at the cost of being relatively opaque, often likened to "black boxes." Word2Vec, in comparison, strikes a balance between complexity and capacity, providing context-sensitive embeddings with a simpler architecture. The comparative analysis of these methods offers insights into the degree of predictability derived from a broad spectrum of NLP techniques.

3 Empirical Analysis

3.1 Data and pre-processing

3.1.1 Stylized Facts

We have sourced our news text data from Refinitiv, a trusted global provider of financial market data. This dataset encompasses global news from both Thomson Reuters Real-time News Feed (RTRS) and the Third Party Archive (3PTY), spanning from January 1996 to June 2019. For US firms, the news falls into two distinct categories: *articles* and *alerts*. Regular news articles feature both a headline and a body of text, offering a comprehensive narrative of various firm events. In contrast, news alerts focus on delivering timely updates on emerging and unfolding news, and thus consist only of a headline. It is important to note that our US 3PTY database and our international news database do not include alerts. This news text data is then integrated with US equity data from the Center for Research in Security Prices (CRSP), and with international equity data obtained from EIKON (Datastream).

In preparing the news database for analysis, we have implemented several filters. First, we have retained only those news articles and alerts associated with a single stock for which three-day close-to-close returns are available. Furthermore, we have removed excessively short news articles with fewer than 100 characters, as well as extremely detailed reports exceeding 100,000 characters. Moreover, we have taken measures to remove redundant articles that essentially replicate the content of preceding stories. Redundancy has been assessed through the computation of a novelty score, derived from cosine similarity calculations based on the bag-of-words representations of any pair of articles. An article is classified as redundant if it attains a cosine similarity score of 0.8 or higher when compared with another article published within the preceding five business days. This process ensures the diversity of the dataset and also safeguards the novelty of the content, resulting in a substantial reduction of superfluous repetition. It is important to note that the removal of such repetition also enhances the signal-to-noise ratio, which is critical as we utilize firm returns as labels in our supervised learning tasks.

Table 1: Summary Statistics of US News Articles and Alerts

Raw Articles			Articles Tagged with Single Stock			Articles With	Filtering Short	Filtering
RTRS	3PTY	Total	RTRS	3PTY	Total	Returns	& Long Articles	Redundancy
6,366,019	4,843,867	11,209,886	2,863,166	4,123,823	6,986,989	4,755,247	4,123,279	3,038,025

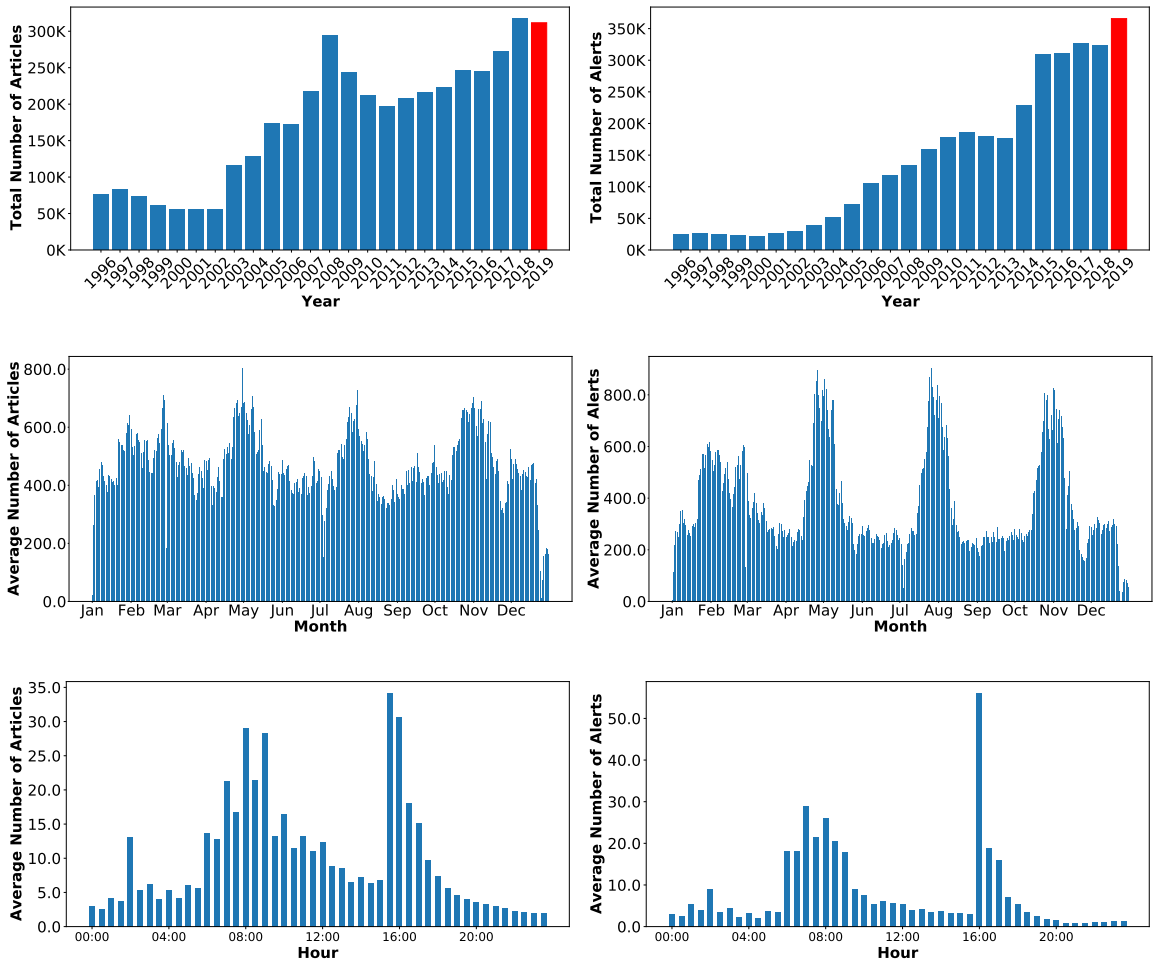
Raw Alerts	Alerts Tagged with Single Stock	Alerts With	Filtering	First In	Second In
RTRS	RTRS	Returns	Redundancy	Take Sequence	Take Sequence
4,976,374	4,054,683	3,286,003	2,935,852	1,296,733	522,258

Note: In this table, we report the remaining sample size after each filter applied on the news articles and news alerts on the top and bottom panels, respectively. Columns under "Raw Articles/Alerts" present the numbers of available articles/alerts separately from Thomson Reuters Real-time News Feed (RTRS) and Archive (3PTY). Columns under "Articles/alerts Tagged with Single Stock" presents the number of articles/alerts tagged with a single stock. Columns "Articles/Alerts with Returns" present the number of remaining articles/alerts after matching returns data. Column "Filtering Short & Long Articles" reports the number of articles with at least 100 characters and at most 1,000,000 characters. Columns "First/Second In Take Sequences" report the number of alerts tagged as the first/second in a sequence of developing alerts. Columns "Filtering Redundancy" report the number of remaining articles/alerts after removing those similar to existing ones (cosine similarity score > 0.8) published in the preceding five business days.

Table 1 provides the statistical breakdown of news articles and alerts associated with the US market after applying various filters. The dataset comprises a substantial volume of over 3 million news articles. Notably, a significant proportion of these articles is sourced from third-party news providers. In terms of alerts, the dataset contains approximately 3 million alerts in total. Among these alerts, 55.3% represent the initial news alerts within a sequence of unfolding alerts. 15.9% of the alerts constitute the second in the sequence, while the remaining alerts fall into the category of third or subsequent alerts.

Figure 1 presents the analysis of news articles and alerts' temporal distribution. Both categories share similar patterns, reflecting their intertwined nature. Annual data, displayed in the upper section, reveals a rising trend from 1996 to 2019, with a notable surge in 2008 during the global financial

Figure 1: US News Counts



Note: The top panel plots the annual time series of the total number of news articles/alerts, the middle plots the average numbers of news articles/alerts per half an hour (24 hour local time), and the bottom plots the average numbers of news articles/alerts per calendar day. Since our sample ends in June 2019, the number of articles/alerts in 2019 on the first panel is estimated and thus highlighted in red.

crisis. Monthly patterns, in the middle section, show cyclical peaks in February, May, August, and November, an occurrence likely attributed to concentrated earnings events. It is noticeable that the phenomenon gets especially prominent in alerts. Daily trends, in the lower section, depict increased news frequency around market opening and closing times, mirroring trading activity ebbs and flows.

Beyond the US market, our analysis also incorporates international markets including China (HK), UK, Australia, Canada, Japan, Germany, Italy, France, Sweden, Denmark, Spain, Finland, Portugal, Greece, and the Netherlands. Figure 2 exhibits annual time series depicting the number of news articles for each international market. A summary of the market information and the processed dataset for each country is encapsulated in Table 2. Table IA10 in the appendix provides a summary

of the sample size after undergoing a similar step-by-step filtering process for international markets.

Notably, there exists significant variation in the volume of news articles among different countries, ranging from a minimum of 3,751 articles (Netherlands) to a maximum of 571,285 (UK). While data acquisition for most countries commenced in 1996, aligning with the initiation year of US data, certain countries' datasets have later inception points due to gaps in news data provision. For instance, the Netherlands began data collection in September 2005. The monthly average of news-covered stocks varies from a minimum of 11 (Netherlands) to a maximum of 645 (Japan).

Table 2: Summary Statistics of International Markets

	Language	Market Hours	# of Articles	Initial Day	Avg. # of Stocks	# of Days	Avg. # of News
US (Alert)	English	09:30 – 16:00	2,935,852	1996-01-02	1,746	5,929	10,337
US	English	09:30 – 16:00	3,038,025	1996-01-02	2,593	5,933	10,697
UK	English	08:00 – 16:30	571,285	1996-01-02	454	6,087	2,011
Australia	English	10:00 – 16:00	249,190	1996-01-03	287	6,033	877
Canada	English	09:30 – 16:00	350,549	1996-01-03	406	6,032	1,234
China (HK)	Chinese	09:30 – 16:00	182,363	1996-01-03	247	5,768	642
Japan	Japanese	09:00 – 15:00	310,244	1996-01-05	645	5,875	1,092
Germany	German	09:00 – 17:30	178,039	1996-01-03	163	6,031	626
Italy	Italian	09:00 – 17:30	130,168	1996-01-05	97	5,778	458
France	French	09:00 – 17:30	153,779	1996-01-03	167	5,994	541
Sweden	Swedish	09:00 – 17:25	115,195	2001-06-07	170	4,629	526
Denmark	Danish	09:00 – 16:55	43,584	1996-01-22	37	4,559	156
Spain	Spanish	09:00 – 17:30	34,159	1996-01-05	37	5,520	120
Finland	Finnish	10:00 – 18:25	28,633	2003-01-03	50	4,025	143
Portugal	Portuguese	11:30 – 16:30	6,158	2005-05-13	11	2,616	36
Greece	Greek	10:15 – 05:20	7,710	2003-02-19	16	3,057	39
Netherlands	Dutch	09:00 – 17:30	3,751	2005-09-20	11	2,102	22

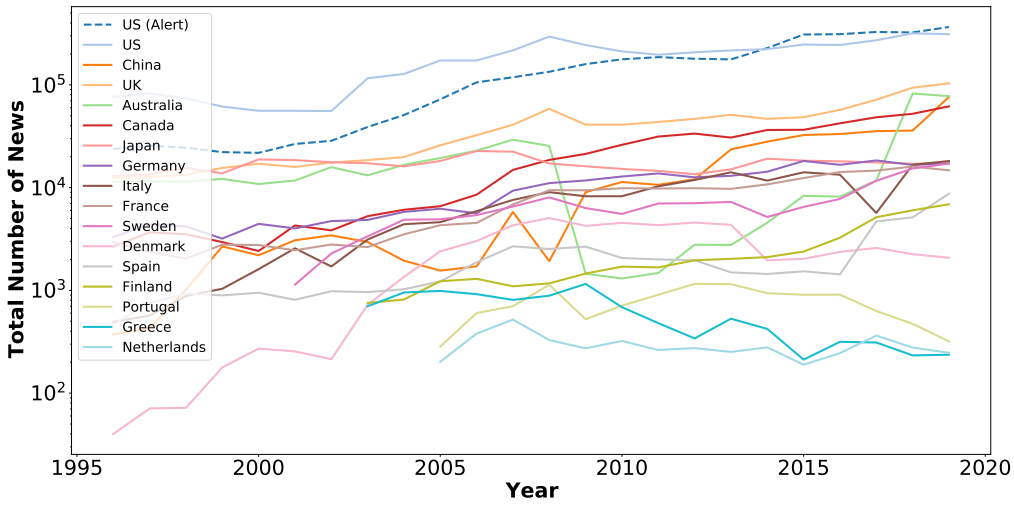
Note: This table summarizes market information and processed datasets for each country. The columns correspond to the language of news articles, local times corresponding to market hours, the overall count of news articles, the initial day of our sample period, the average number of available stocks per month, the total number of days with news articles, and the average monthly count of news articles.

Table 3: Summary Statistics of Characters/Tokens/Words in US News Articles and Alerts

	Article					Alert				
	1%	25%	50%	75%	99%	1%	25%	50%	75%	99%
# of Characters	163	511	1566	3795	29887	32	63	78	103	160
# of LLaMA Tokens	59	175	451	978	10029	19	33	41	51	79
# of RoBERTa Tokens	43	129	348	783	7076	14	25	31	38	59
# of BERT Tokens	44	129	352	802	7234	9	19	23	28	42
# of Words	6	30	89	240	1896	0	3	5	7	14

Note: Row "# of Characters" report the percentiles of the number of characters in the raw article. Rows "# of LLaMA Tokens", "# of RoBERTa Tokens", and "# of BERT Tokens" report the percentiles of the number of tokens converted from news text using model specific tokenizer. Row "# of Words" reports the percentiles of the number of words extracted from an article (after removing pronouns/stop words) that are used by SESTM/Word2vec.

Figure 2: Total Number of News Articles/Alerts



Note: This figure plots the total number of news articles each year for all international markets. The blue dashed line plots the number for news alerts for US equity market.

3.1.2 News Embeddings

We both use word-based models and Large Language Models (LLMs) to generate embeddings for news articles and alerts. LLMs operate at the level of tokens, whereas word-based models take individual words as inputs. Detailed statistics on token and word counts could be found at Table 3, and Table 4 presents the corresponding counts for international news articles.

Word-based models like Word2Vec and SESTM require meticulous data preprocessing to operate effectively at the word level. We follow the procedure outlined in Ke et al. (2019) to derive Bag-of-Words (BOW) representations for news articles. This comprehensive preprocessing includes converting text to lowercase, expanding contractions (e.g., "haven't" to "have n't"), lemmatization (reducing words to their base forms), tokenization, and removal of pronouns, proper nouns, punctuation, special symbols, numbers, non-English words (for English texts), and common stop words like "and," "the," and "is." To operate on text data preprocessing across all languages, we utilize the natural language processing package "spaCy". In this way, each article is then represented using its word count vector, ensuring accurate word-based embeddings.

In contrast, Large Language Models, such as BERT, RoBERTa and LLaMA, possess the advantage of accepting raw, unprocessed text as input. This capability significantly reduces the need for extensive data cleaning. We have selected specific pre-trained LLMs for each country, as detailed in Table IA11 in the Appendix.

Table 4: Summary Statistics of Tokens/Words in International News Articles

	LLaMA					Word2vec/SESTM				
	1%	25%	50%	75%	99%	1%	25%	50%	75%	99%
US	59	175	451	978	10029	6	28	88	239	1882
UK	77	247	590	1184	18280	7	40	108	247	3106
Australia	62	69	219	937	24439	6	19	38	199	4578
Canada	76	356	823	1478	12071	7	59	193	382	2670
China	43	52	65	124	5265	14	16	21	39	1880
Japan	155	240	365	459	1636	71	99	134	159	756
Germany	70	314	508	1026	5492	8	52	90	171	975
Italy	47	277	533	1476	9998	11	64	126	256	2276
France	75	284	528	1093	10283	13	66	126	271	2151
Sweden	85	260	630	1034	5561	13	54	129	218	1214
Denmark	64	248	439	781	4124	8	40	77	144	651
Spain	40	101	260	468	12944	6	18	46	92	1807
Finland	222	561	857	1541	20035	24	83	139	265	2885
Portugal	61	161	313	537	1617	5	18	61	124	394
Greece	136	552	960	1778	4791	16	55	94	171	460
Netherlands	127	450	741	1208	6964	18	84	146	242	1282

Note: Columns under "LLaMA" report the percentiles of the number of tokens converted from text using specific tokenizers for each country. Columns under "Word2vec/SESTM" report the percentiles of the number of words extracted from an article (after removing pronouns/stop words).

3.2 Model Training

On the basis of Word2vec and LLMs, we commence by acquiring P-dimensional pre-trained features, denoted as $x_{i,t}$, for each news article i at time t within our sample dataset. In sentiment analysis, we train model (1) using a cross-entropy loss. In With respect to predicting the cross-section of returns, we employ a penalized squared-error loss for training model (2) and additionally apply ridge penalty as a means of regularization for overall robustness of our models.

Notably, SESTM introduces a distinctive structural assumption that sets it apart from conventional word-based models. Consequently, its training and prediction methodologies diverge from the norm. The SESTM framework imposes structural assumptions on the BOW representation of article i at time t, $d_{i,t}$, and its associated sentiment score $p_{i,t}$:

$$\mathrm{P}(sgn(y_{i,t}) = 1) = g(p_{i,t}), \quad d_{[S],i,t} \sim Multinomial(s_{i,t}, p_{i,t}O_+ + (1 - p_{i,t})O_-),$$

where $g(\cdot)$ is some increasing function, $s_{i,t}$ is the total number of sentiment charged words for article i at time t, O_+ and O_- are $|S| \times 1$ vectors of parameters, the set S is part of the vocabulary with an exclusive list of sentiment charged words, and $d_{[S],i,t}$ is a subvector of $d_{i,t}$ with rows corresponding to words in set S. Ke et al. (2019) proposes this model and suggests that SESTM's training involves the construction of in-sample estimates for various variables. To be more specific, we can construct $\widehat{S}$ by screening based on how frequently each word appears in a positive article and construct $\widehat{O}_\pm$ by running regressions of sentiment word frequencies of each article onto pilot estimates of in-sample

sentiment scores. Based on the estimated $\widehat{O}_{\pm}$ and $\widehat{S}$, we are able to conduct the maximum likelihood estimator of the sentiment score for an article out of sample. Several tuning parameters will be involved in the process, including three in the screening step, and one shrinkage parameter in the (penalized) likelihood estimation step.

We train each model using annually updated rolling windows. Each rolling window consists of a 8-year interval for in-sample training with the first six years for training and the next two for validation. The subsequent one-year data is then set aside for out-of-sample testing. As a result, the out-of-sample data range from 2004 to 2019, totaling 16 years. The tuning parameters are selected in the validation sample.

3.3 Portfolio Performance

3.3.1 Sentiment Analysis

The sentiment analysis aims to predict a binary outcome: one indicating a positive return and zero signifying otherwise. The fitted value of the logistic regression, $\sigma(x'\widehat{\beta})$, is an estimate for the probability of a positive outcome for an article with feature x. A true positive (TP) or true negative (TN) occurs when a predicted "up" probability of greater than 50% coincides with a positive realized return and a probability less than 50% coincides with a negative return.[6] False positives and negatives (FP and FN) are the complementary outcomes. We calculate classification accuracy as follows:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN).$$

Table 5 presents the yearly out-of-sample prediction accuracy, offering several noteworthy observations. Firstly, the first six models consistently outperform a random guess (50%) in terms of average accuracy over these years. Remarkably, the three Language Models (ChatGPT, LLaMA(LLaMA2), ROBERTa and BERT) exhibit higher overall accuracy compared to the word-based models (Word2vec and SESTM), with ChatGPT achieving an average accuracy of 54.28%.

However, it's important to note that even the best-performing model, ChatGPT, does not show a significant increase in accuracy compared to a random guess. These statistical artifacts are primarily due to market efficiency. In a well-functioning market, unpredictable news dominates equity returns, resulting in a small predictable component. This explains why all models achieve accuracy slightly above 50%. Nevertheless, this modest level of predictability can still lead to substantial gains in terms of investment performance.

To evaluate out-of-sample predictive performance in economic terms, we introduce a trading strategy that capitalizes on sentiment estimates for investment decisions. Our trading approach is straightforward: we construct a zero-net-investment portfolio by taking long positions in the top

[6]The threshold of 50% is a natural cutoff for positive sentiment score. Alternatively, we also consider the unconditional up probability as a threshold for the data. As we subsequently show, the empirical results remain consistent as we only trade stocks with extreme sentiment scores.

Table 5: Out-of-Sample Prediction Accuracy

	ChatGPT	LLaMA2	LLaMA	ROBERTa	BERT	Word2vec	SESTM	LMMD
2004	54.34	53.85	53.18	54.08	53.60	53.01	50.14	46.32
2005	53.54	53.02	53.00	53.24	53.25	52.44	51.77	47.93
2006	55.28	54.81	54.52	54.70	54.69	54.35	53.70	46.72
2007	53.91	53.63	53.65	53.65	53.10	52.38	52.55	48.67
2008	51.36	50.89	51.86	51.18	50.35	49.00	51.66	53.45
2009	54.70	54.36	53.07	53.97	54.08	53.93	52.52	46.52
2010	54.76	55.13	53.22	55.31	55.07	55.25	51.58	45.14
2011	53.50	51.45	51.45	51.45	51.45	51.79	52.34	48.55
2012	55.07	53.73	53.27	53.68	53.47	54.15	52.51	46.67
2013	56.55	55.33	54.31	55.10	54.92	55.14	51.70	44.35
2014	53.93	53.35	52.48	53.58	53.27	53.05	51.89	47.09
2015	53.33	52.85	52.77	52.59	52.34	51.35	51.22	49.93
2016	54.73	54.21	53.73	53.83	53.68	53.25	51.90	47.36
2017	55.81	55.48	54.27	55.45	54.91	54.36	51.01	46.30
2018	51.99	51.57	51.42	51.69	51.63	50.81	51.14	50.32
2019	56.83	56.83	55.21	56.39	55.96	55.78	52.85	44.72
Mean	54.35	53.78	53.21	53.74	53.49	53.13	51.91	47.50
Overall	54.28	53.69	53.15	53.66	53.41	53.04	51.88	47.60
Top 20%	57.85	56.75	56.67	56.41	55.66	54.95	55.65	46.54
Bot. 20%	55.17	53.86	53.50	53.42	52.87	52.10	51.72	50.73

Note: The table reports out-of-sample classification accuracy of the sentiment measure for ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, Word2vec, SESTM and LMMD models for each year in the testing sample.

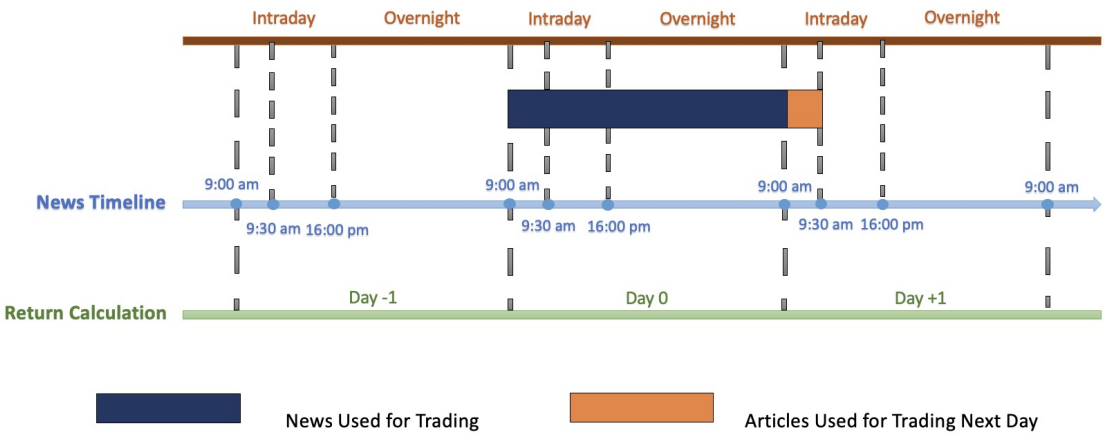
quintile (20%) of stocks with the most positive sentiment scores and short positions in the bottom quintile (20%) of stocks with the most negative sentiment scores.

When forming the long and short sides of the strategy, we consider both equal-weighted and value-weighted schemes. Equal weighting provides a straightforward and robust method to assess the predictive power of sentiment across the spectrum of firm sizes and aligns with common practices in hedge funds' news text-based portfolio construction. On the other hand, value weighting accounts for size effect by placing greater emphasis on large-cap stocks, which can be justified for economic reasons (assigning more weight to more productive firms) and practical trade implementation reasons (such as managing transaction costs).

Additionally, in line with a similar choice made by Ke et al. (2019), we exclude articles published between 9:00 am and 9:30 am Eastern Standard Time (EST) for US markets, and a similar strategy applies to global markets. This decision is driven by our commitment to aligning with realistic considerations. Instead of incorporating the information published during this half-hour window into the same day's prediction input, we opt to shift these articles to the following day. This adjustment allows funds ample time to compute their positions in response to news developments and facilitates trading when market liquidity tends to be at its peak. We illustrate the timeline in Figure 3.

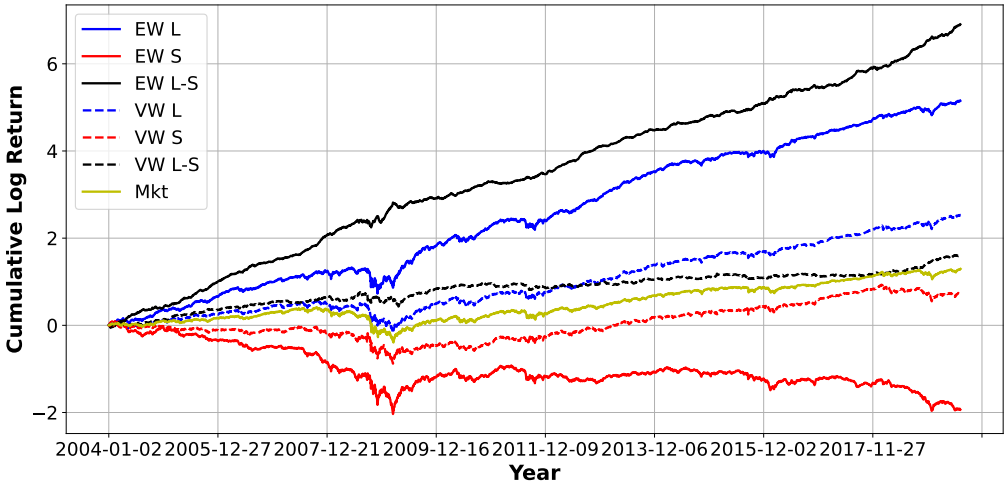
Figure 4 illustrates the cumulative returns of the long-short (L-S), long (L), and short (S) portfolios constructed using LLaMA2 model. The performance of both equal-weighted ("EW") and value-weighted ("VW") versions of the strategy is also presented for comparison. Notably, the long-short strategy successfully avoids significant drawdowns and demonstrates positive returns during the financial crisis, contrasting with the downward movement of the SPY index. Furthermore, Figure

Figure 3: News Timeline



Note: This figure describes the news timeline and our trading activities. We move news from 9:00 am to 9:30 am EST to next trading day for feasibility (in our testing exercise). For news that occur on day 0 , we build positions at the market opening on day 1 , and rebalance at the next market opening, holding the positions of the portfolio within the day. We call this portfolio day +1 portfolio. Similarly, we can define day +2, day +3, ..., day +10 portfolios.

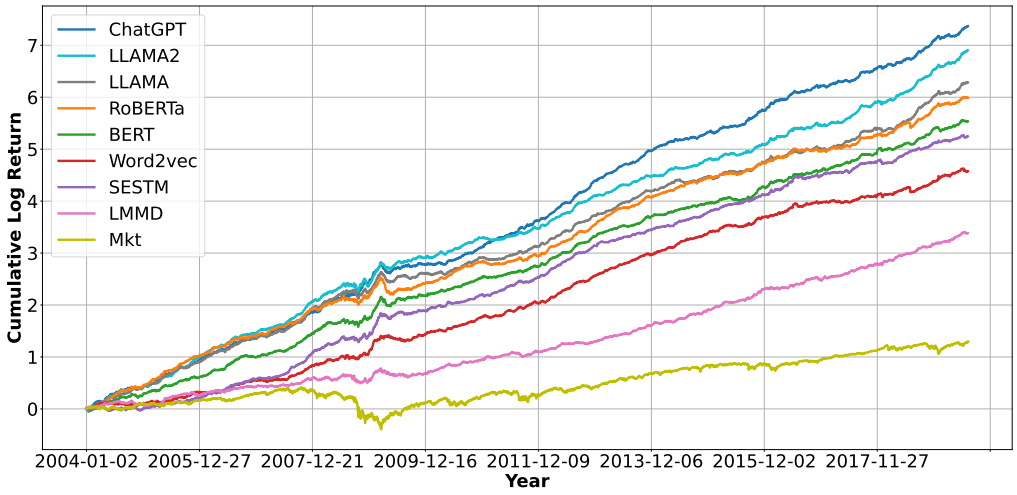
Figure 4: One-day-ahead Portfolio Performance based on LLaMA2



Note: This figure compares the out-of-sample cumulative log returns of portfolios sorted on sentiment scores based on the LLaMA2 model. The black, blue, and red colors represent the long-short (L-S), long (L), and short (S) portfolios, respectively. The solid and dashed lines represent equal-weighted (EW) and value-weighted (VW) portfolios, respectively. The yellow solid line is the S&P 500 return (SPY).

5 showcases the cumulative one-day trading strategy returns, calculated from open-to-open, based on out-of-sample sentiment forecasts. It is evident that all five LLMs (ChatGPT, LLaMA2, LLaMA, ROBERTa and BERT) consistently outperform all word-based models (Word2vec and SESTM).

Figure 5: Performance of US Equal-Weighted Portfolios



Note: This figure presents US equal-weighted long-short portfolio cumulative log returns for portfolios sorted on sentiment scores. The portfolios are built on the basis of ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, Word2vec, SESTM, and LMMD models, respectively. "Mkt" is the cumulative return for S&P 500 return (SPY).

Table 6: Performance of Daily News Sentiment Portfolios in US

	ChatGPT						LLaMA2					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.34	-0.14	0.48	0.19	0.04	0.15	0.35	-0.10	0.45	0.18	0.07	0.11
Std	0.20	0.22	0.10	0.19	0.22	0.11	0.20	0.23	0.11	0.19	0.22	0.11
SR	1.71	-0.62	4.62	1.03	0.18	1.41	1.75	-0.43	4.16	0.97	0.33	0.98
	LLaMA						RoBERTa					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.34	-0.07	0.41	0.19	0.08	0.11	0.33	-0.06	0.39	0.20	0.09	0.11
Std	0.20	0.23	0.11	0.19	0.22	0.11	0.20	0.22	0.10	0.19	0.22	0.11
SR	1.67	-0.33	3.89	1.02	0.36	1.04	1.62	-0.29	3.75	1.08	0.43	0.94
	BERT						Word2vec					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.32	-0.04	0.36	0.16	0.07	0.10	0.29	-0.01	0.30	0.18	0.08	0.09
Std	0.20	0.22	0.10	0.18	0.21	0.10	0.21	0.22	0.10	0.19	0.21	0.10
SR	1.59	-0.19	3.60	0.89	0.31	0.92	1.41	-0.05	3.06	0.93	0.40	0.92
	SESTM						LMMD					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.31	-0.03	0.34	0.18	0.09	0.09	0.24	0.01	0.22	0.14	0.10	0.04
Std	0.20	0.22	0.10	0.19	0.21	0.11	0.20	0.23	0.10	0.18	0.21	0.10
SR	1.53	-0.14	3.43	0.97	0.42	0.86	1.18	0.06	2.29	0.77	0.47	0.39

Note: The table reports the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted on sentiment scores and their long (L) and short (S) legs. The portfolios are built on the basis of ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, Word2vec, SESTM, and LMMD models, respectively.

Table 6 provides an overview of the performance statistics for these portfolios. Analyzing these portfolio results in Table 6 reveals two key findings.

Firstly, equal-weighted portfolios exhibit significantly better performance compared to their value-weighted counterparts. Specifically, the long-short strategy based on LLaMA2 with equal weights achieves an annualized Sharpe ratio of 4.16, whereas the value-weighted case only attains a Sharpe ratio of 0.98. This indicates that news article sentiment is a stronger predictor of future returns for small stocks, assuming all other factors remain constant. Several potential economic explanations exist for this observation. It could be attributed to the facts that i) small stocks receive less attention from investors, resulting in slower responses to news; ii) the underlying fundamentals of small stocks are more uncertain and less transparent, requiring more effort to process news and translate it into actionable price assessments; and iii) small stocks are less liquid, necessitating more time for trading to incorporate information into prices.

Secondly, both the long and short sides of the trade demonstrate significant profitability. The long side marginally outperforms the short side, exhibiting a Sharpe ratio of 1.75 compared to 0.43 (in the equal-weighted case) for LLaMA2. This can be partially attributed to the long side naturally capturing the market equity risk premium, while the short side bears the cost of it. Additionally, it is possible that investors face constraints on short sales, preventing negative news from being fully reflected in prices over short time horizons.

Finally, we provide robustness check in Table 7 by varying some of the implementation choices we make in the portfolio formation step. We focus on the performance of the LLaMA2 model. To begin with, we limit the number of stocks for trading each day from the top/bottom quintile (20%) to decile (10%). This increases the Sharpe ratios for equal weighted portfolio, from 4.16 to 4.29 (EW). Though the Sharpe ratio for the value-weighted portfolio decreases from 0.98 to 0.78 (VW), the average return increases from 0.11 to 0.13 (VW). Next, we exclude all news that occurs around earning announcement days to examine the effect of non-earning news. Concretely, we select again the top/bottom quintile of stocks on each day t, but avoid those whose earning announcements are scheduled on day $t-1$, day t, and day $t+1$. The resulting Sharpe ratios suggest that a large amount of information in our sentiment scores does not directly stem from earnings reports. For example, LLaMA2's Sharpe ratios are 3.62 (EW) and 0.80 (VW), compared with 4.16 (EW) and 0.98 (VW) from the benchmark. Additionally, we experiment with using residual returns rather than raw returns as our supervising labels in the training procedure. Residuals are obtained from time-series regressions of raw returns over a two-year rolling window with respect to either the CAPM model, or the Fama-French three-factor model, or the CAPM model augmented with 17 industry portfolios.[7] The resulting Sharpe ratios are a bit higher than those of the benchmark model: 4.90, 4.90, and 5.31 for equal-weight L-S portfolios, and 1.27, 1.24, 1.44 for value-weight portfolios. This result suggests that the out-of-sample portfolio performance is fairly robust with respect to labels used in the training sample.

[7]These 17 industry portfolios (equal-weighted) are obtained from Kenneth French's website.

Table 7: Performance of Alternative Sentiment Portfolios based on the LLaMA2 Model

Type	BenchMark		Trading Top Decile		w/o. Earning Days	
	Sharpe Ratio	Avg. Ret.	Sharpe Ratio	Avg. Ret.	Sharpe Ratio	Avg. Ret.
EW L-S	4.16	0.45	4.29	0.66	3.62	0.39
EW L	1.75	0.35	2.00	0.42	1.62	0.33
EW S	-0.43	-0.10	-0.96	-0.24	-0.29	-0.07
VW L-S	0.98	0.11	0.78	0.13	0.80	0.09
VW L	0.97	0.18	1.04	0.21	0.95	0.18
VW S	0.33	0.07	0.34	0.09	0.42	0.09
Type	CAPM Resid.		FF3 Resid.		CAPM+ind. Resid.	
	Sharpe Ratio	Avg. Ret.	Sharpe Ratio	Avg. Ret.	Sharpe Ratio	Avg. Ret.
EW L-S	4.90	0.51	4.90	0.50	5.31	0.53
EW L	2.13	0.43	2.09	0.43	2.11	0.44
EW S	-0.35	-0.08	-0.35	-0.08	-0.43	-0.09
VW L-S	1.27	0.15	1.24	0.14	1.44	0.16
VW L	1.16	0.22	1.13	0.21	1.24	0.24
VW S	0.30	0.07	0.33	0.07	0.39	0.08

Note: The table reports the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios and their long (L) and short (S) legs, based on the LLaMA2 model. The performance measures include (annualized) annual Sharpe ratio and annualized expected returns. We have presented results of the benchmark LLaMA2 portfolio in Table 6. Additionally, we report results for alternative portfolios constructed similarly: trading up to 10% stocks per day, excluding firms near their earnings announcement days, using residuals from time series regressions of raw returns (with respect to the CAPM model, the Fama-French three-factor model, or the CAPM model augmented with 17 industry portfolios as factors).

3.3.2 Return Prediction

While sentiment analysis based on LLMs holds promise for investment, the textual features from news may contain information beyond the sentiment (sign) of returns. In this section, we exploit the ability of these features in measuring the cross-section of expected returns within short horizon[8]. Compared to sentiment analysis, there are two key differences. First, the target variables are directly taken as realized returns. Second, the magnitude of returns may carry valuable information for selecting textual features. We exclude SESTM and LMMD in the this section since SESTM is designed to measure contemporary sentiment, and Loughran-McDonald Master Dictionary is specifically designed to capture sentiment in financial text. while the Loughran-McDonald Master Dictionary is specifically engineered to analyze sentiment within financial texts.

Table 8 presents the out-of-sample prediction correlation on an annual basis. All models consistently achieve a positive average correlation over the years. LLaMA2 and LLaMA exhibit out-of-sample correlations exceeding 2%, outperforming other models. ChatGPT and ROBERTa, while having slightly lower correlations, still surpass Word2vec. BERT lags slightly behind with an overall correlation of 1.62, which is lower than Word2vec's correlation of 1.84.

[8]Specifically, we estimate a ridge regression model using the forward one day's open-to-open returns as dependent variable. The penalty coefficient is tuned from the choices of 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 5e1, 1e2.

Table 8: Out-of-Sample Cross-Sectional Correlations

	ChatGPT	LLaMA2	LLaMA	ROBERTa	BERT	Word2vec
2004	0.83	1.21	1.26	1.27	1.17	1.18
2005	1.57	2.07	1.99	1.36	1.08	1.17
2006	0.34	1.15	1.18	0.17	0.86	1.10
2007	1.19	2.51	2.47	2.14	1.57	1.69
2008	2.00	2.19	2.17	1.79	0.89	1.54
2009	1.20	2.35	2.17	1.40	1.12	2.07
2010	2.30	2.82	2.68	2.28	2.02	1.81
2011	2.15	2.01	1.79	1.31	0.98	1.12
2012	2.05	2.65	2.83	2.04	2.08	2.08
2013	1.77	3.28	3.23	2.41	2.35	1.90
2014	2.07	2.45	2.47	2.09	1.77	2.07
2015	2.30	2.90	2.70	2.54	1.53	2.12
2016	2.04	2.22	2.28	1.66	1.29	2.12
2017	2.79	3.89	3.59	2.84	2.34	2.68
2018	3.52	3.94	3.65	2.94	2.65	2.44
2019	2.85	3.97	3.82	3.48	2.85	2.97
Mean	1.94	2.60	2.52	1.98	1.66	1.88
Overall	1.91	2.56	2.48	1.93	1.62	1.84
Top quintile	0.83	1.88	1.82	1.09	0.36	0.22
Bottom quintile	1.43	2.23	2.04	1.27	1.28	1.07

Note: The table reports the time-series average of the cross-sectional rank correlations between the predicted return and the future one day's open-to-open returns for ChatGPT, LLaMA2, LLaMA RoBERTa, BERT, and Word2vec models for each year in the testing sample.

The portfolio construction methodology in this section is similar to that of sentiment portfolios, with one key difference: we use predicted returns as sorting variables instead of sentiment scores. Table 9 presents the performance of portfolios derived from cross-sectional return predictions. The results reveal that LLaMA2, LLaMA and RoBERTa can outperform their sentiment portfolio counterparts in terms of Sharpe ratios, both in equal-weighted and value-weighted scenarios. (e.g. Specifically, LLaMA2 improves from 4.16 (EW) and 0.98 (VW) to 5.31 (EW) and 1.32 (VW)). For less advanced LLM and word-based model, such as BERT and Word2vec, the performance of the return prediction portfolios is slightly worse than that of the sentiment portfolios in value-weighted case, but still can beat sentiment analysis in equal-weighted case. This suggests that directly incorporating next period realized returns assists in selecting textual features that predict returns.

3.4 News Assimilation

Timely information about the market is captured by news, and this information is assimilated into prices rapidly. To assess the speed of information assimilation in economic terms, we utilize a database containing sequences of news alerts and their order in a developing story. We train various models using all the alerts and construct separate portfolios using subsets of news alerts to illustrate the impact of their sequences and evaluate the performance of portfolios formed based on different sequences of news alerts. The comparison of Sharpe ratios between portfolios based on news alerts and news articles is presented in Table 10.

Our analysis yields several significant findings. First, portfolios based on news alerts generate

Table 9: Performance of Portfolios sorted by the Cross-Section of Return Predictions

	ChatGPT						LLaMA2					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.39	-0.04	0.43	0.22	0.11	0.10	0.46	-0.11	0.57	0.23	0.09	0.14
Std	0.21	0.22	0.10	0.20	0.20	0.11	0.21	0.22	0.11	0.20	0.20	0.11
SR	1.87	-0.21	4.23	1.07	0.58	0.91	2.22	-0.50	5.31	1.14	0.44	1.32
	LLaMA						RoBERTa					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.45	-0.11	0.56	0.23	0.08	0.15	0.37	-0.05	0.42	0.21	0.08	0.13
Std	0.21	0.22	0.11	0.20	0.20	0.11	0.20	0.21	0.10	0.20	0.20	0.11
SR	2.17	-0.51	5.17	1.12	0.41	1.35	1.81	-0.23	4.36	1.06	0.42	1.17
	BERT						Word2vec					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	-0.03	0.37	0.16	0.09	0.07	0.33	-0.00	0.33	0.20	0.13	0.08
Std	0.21	0.22	0.09	0.19	0.20	0.10	0.21	0.22	0.10	0.19	0.20	0.10
SR	1.62	-0.15	3.94	0.85	0.45	0.68	1.61	-0.00	3.20	1.06	0.63	0.74

Note: The table reports the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios and their long (L) and short (S) legs. The portfolios are built on the basis of ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, and Word2vec, respectively, using the cross-section of expected returns as sorting variables.

higher Sharpe ratios compared to those based on news articles. This is somewhat unexpected, as news alerts often lack detailed body text and convey information solely through the headline. This suggests that the speed of information dissemination is more crucial than the depth of content provided.

Secondly, there is a noticeable decline in Sharpe ratios from portfolios based on the initial set of news alerts (TS1) to those based on a subsequent set (TS2), and further to later sets. This decline demonstrates that news is rapidly absorbed into market prices, diminishing the value of delayed trading. By the time a second update in a developing story is received, the market has typically already adjusted, resulting in a sharp decrease in Sharpe ratios.

In Figure 6, we present the average returns in basis points per day along with shaded 95% confidence intervals for sentiment portfolio. We observe that sentiment information is effectively assimilated into prices by the beginning of Day 3. Notably, there is a considerable decline in returns from Day 1 to Day 2, particularly for news alerts. This observation aligns with our expectations since our sentiment score predominantly captures the sentiment of fresh news that has not yet been fully incorporated into market prices. It emphasizes the need for timely decision-making and efficient trading strategies to capitalize on news-based signals before they lose their potential profitability due to rapid information assimilation by the market.

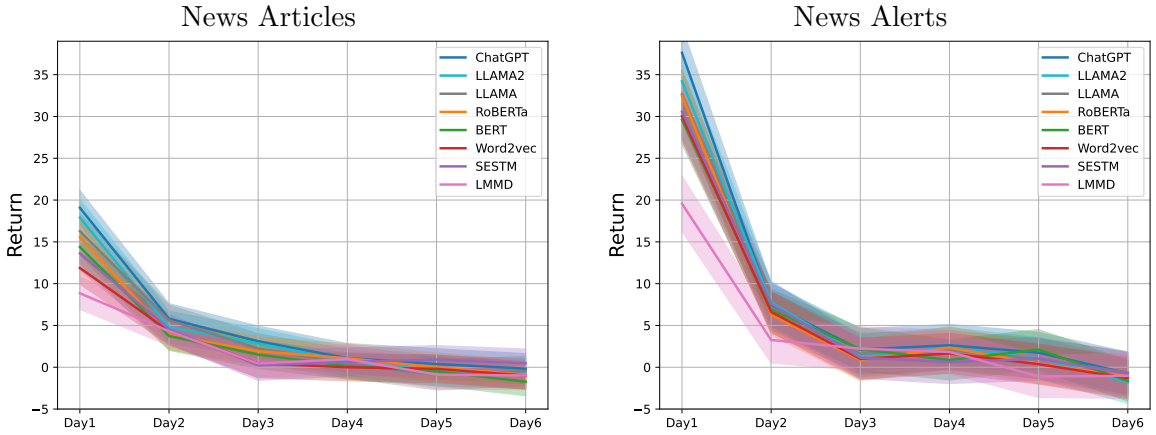
We further investigate the difference in news assimilation with heterogeneity of stocks. We analyze the difference by average return in basis point and the corresponding 95% confidence interval in Table 11. Sole small stocks exhibit an average daily return of 29.06 basis points on the day following a news release—over five times as the 5.66 basis point return of large stock portfolios. And large stocks' the

Table 10: Sharpe Ratios of Portfolios based on News Articles and News Alerts

	EW					VW				
	Article	Alert				Article	Alert			
		All	TS1	TS2	Rest		All	TS1	TS2	Rest
ChatGPT	4.62	6.06	5.81	4.03	3.71	1.41	2.76	2.78	0.80	0.97
LLaMA2	4.16	5.77	5.60	3.92	3.20	0.98	2.54	2.74	0.55	0.91
LLaMA	3.89	5.48	5.32	3.58	2.93	1.04	2.39	2.42	0.60	0.94
RoBERTa	3.75	5.49	5.33	3.14	3.36	0.94	2.28	2.57	0.80	0.91
BERT	3.60	5.05	4.49	3.19	2.52	0.92	1.86	1.80	0.68	1.02
Word2vec	3.06	5.21	4.89	2.92	2.04	0.92	2.10	2.14	0.95	0.54
SESTM	3.43	4.95	4.41	3.02	2.71	0.86	2.20	2.23	0.52	1.01
LMMD	2.29	2.94	2.77	1.03	0.97	0.39	1.13	1.05	0.16	0.16

Note: The table reports Sharpe ratios of the long-short portfolios built on the basis of ChatGPT, LLaMA, LLaMA2, RoBERTa, BERT, Word2vec, SESTM, and LMMD models, respectively. The top panels reports the Sharpe ratios of portfolios based on sentiment scores and the bottom panel reports the Sharpe ratios of portfolios based on predicted returns. Column "TS1" refers to portfolios that only rely on take sequence 1 alerts, "TS2" only take sequence 2 alerts, "Rest" the remaining alerts, and "All" all alerts.

Figure 6: Speed of News Assimilation



Note: This figure compares average one-day holding period returns to the news trading strategy as a function of when the trade is initiated. We consider daily open-to-open returns initiated from one to 6 days following the announcement. We report equal-weighted portfolio average returns (in basis points per day), with 95% confidence intervals given by the shaded regions.

price reponse is complete after one day, while it takes about 3 days to fully incorporate news into the price for small stocks. A similar pattern emerges when contrasting stocks by volatility, which is shown in Table 12. There is no significant difference in the magnitude of initial return response, but in the speed of price adjustment. High volatility stocks return to baseline levels in just one day, whereas low volatility stocks necessitate a three-day period for complete adjustment. These findings are also illustrated in Figures 7.

Table 11: News Assimilation for LLaMA2: Size

	Big Stocks			Small Stocks		
	Avg. Ret	Std	95% CI	Avg. Ret	Std	95% CI
Day 1	5.66	59.26	[3.8, 7.52]	29.06	113.64	[25.49, 32.63]
Day 2	0.08	50.68	[-1.52, 1.67]	9.19	101.00	[6.02, 12.36]
Day 3	0.99	53.37	[-0.68, 2.67]	4.86	90.98	[2.0, 7.71]
Day 4	0.56	51.73	[-1.06, 2.18]	1.26	88.73	[-1.52, 4.05]
Day 5	-0.01	49.36	[-1.56, 1.53]	0.05	94.75	[-2.92, 3.03]
Day 6	-0.14	51.50	[-1.75, 1.48]	0.60	89.17	[-2.2, 3.4]

Note: This table presents heterogeneity in stock size in news assimilation impact over six days. We segment stocks into large stocks and small stocks by cross-sectional market capital size meidan. Average returns (Avg. Ret), standard deviations (Std), and 95% confidence intervals (95% CI) are displayed

Table 12: News Assimilation for LLaMA2: Volatility

	High Vol Stocks			Low Vol Stocks		
	Avg. Ret	Std	95% CI	Avg. Ret	Std	95% CI
Day 1	18.16	109.82	[14.72, 21.61]	17.26	62.54	[15.29, 19.22]
Day 2	3.07	99.87	[-0.06, 6.2]	5.86	51.20	[4.26, 7.47]
Day 3	2.85	91.00	[-0.0, 5.71]	2.90	47.94	[1.39, 4.4]
Day 4	-0.54	90.87	[-3.39, 2.31]	1.10	47.31	[-0.38, 2.59]
Day 5	-0.66	96.29	[-3.68, 2.36]	0.60	47.16	[-0.88, 2.08]
Day 6	-1.10	94.43	[-4.07, 1.86]	0.63	46.90	[-0.84, 2.1]

Note: This table presents heterogeneity in stock volatility in news assimilation impact over six days. We segment stocks into high-volatility stocks and low-volatility stocks by cross-sectional volatility meidan. Average returns (Avg. Ret), standard deviations (Std), and 95% confidence intervals (95% CI) are displayed

3.5 News Momentum

In general, stock returns typically exhibit a pronounced short-term reversal effect. However, as the analysis shows, portfolios constructed using news events demonstrate a positive Sharpe ratio, suggesting a momentum effect. To further explore the impact of news on stock momentum, we construct portfolios based on past 5-day cumulative returns, sorting them into long and short positions. The results, presented in Table 13, indicate that the short-term reversal effect, while still evident, is diminished by the existence of news, shifting from -2.33 to -1.58. Although the gap is small, the presence of news alone can already create some level of momentum, though not strong enough to offset the short-term reversal effect completely. In fact, when portfolios are formed based on sentiment scores derived from news and traded within a universe of news-tagged stocks, a clear decay pattern is observed. This pattern, detailed in Table 14, shows a significant positive Sharpe Ratio on the first day, which gradually fades in the following days, which demonstrates a significant momentum effect.

3.6 Context > Words > Past Returns

We then compare the performance of portfolios based on sentiment scores derived by news and past returns in the entire market and conditioning on stocks universe influenced by news. We first see the results in the entire market. The signals are constructed based on the following rules:

Table 13: News Momentum

	Entire Market						Stocks with news					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.01	0.33	-0.32	0.08	0.30	-0.21	0.06	0.35	-0.29	0.08	0.29	-0.20
Std	0.18	0.23	0.14	0.20	0.27	0.18	0.23	0.27	0.18	0.23	0.30	0.24
SR	0.06	1.46	-2.33	0.41	1.11	-1.19	0.25	1.29	-1.58	0.37	0.95	-0.83

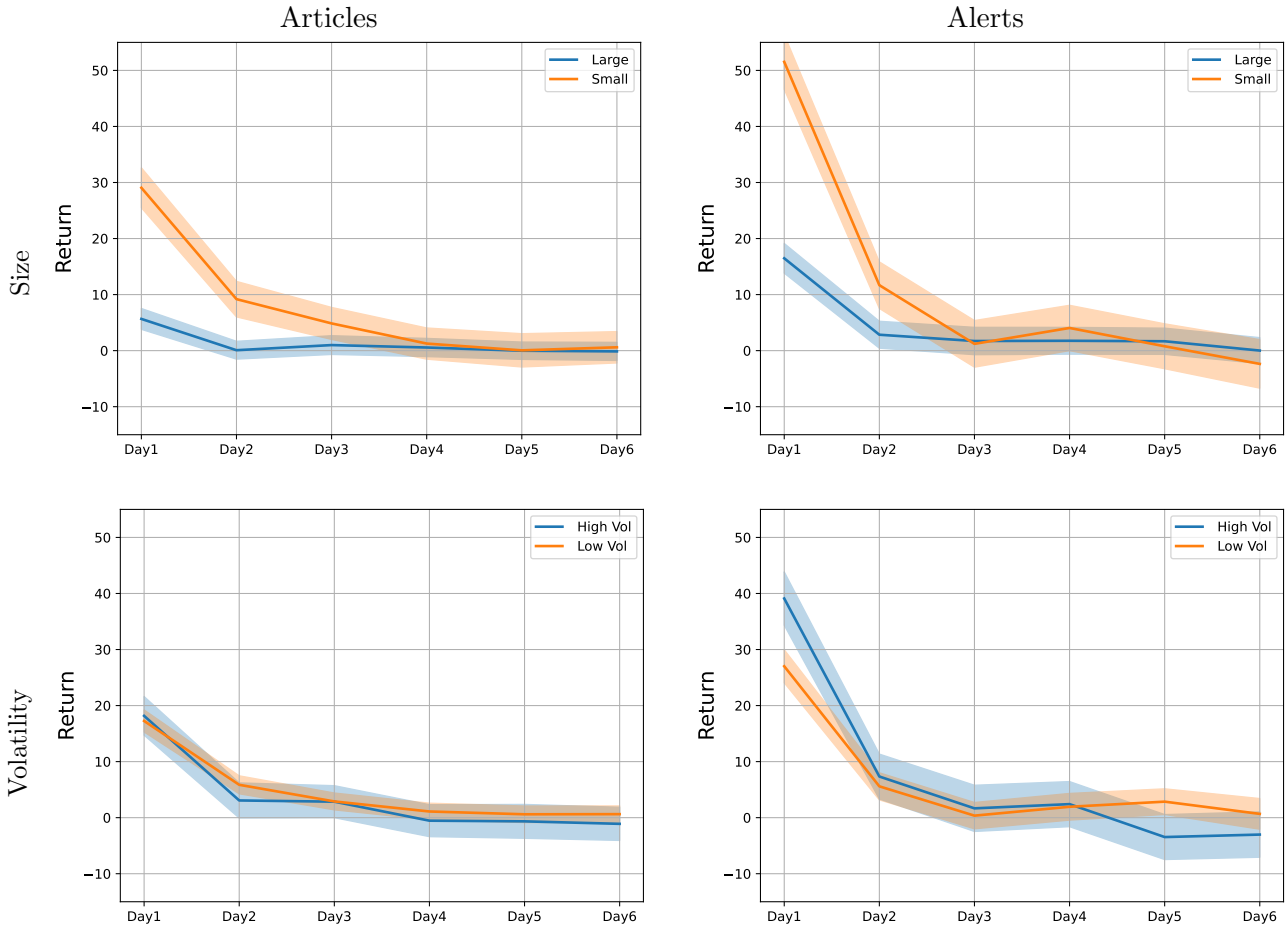
Note: The table presents of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted on past 5-day cumulative returns and their long (L) and short (S) legs. The portfolios are built within entire market and investment universe conditioning on stocks tagged with news.

Table 14: Delayed Portfolio Performance

	Day +1 Portfolios						Day +2 Portfolios					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.35	-0.10	0.45	0.18	0.07	0.11	0.17	0.05	0.12	0.06	0.06	0.00
Std	0.20	0.23	0.11	0.19	0.22	0.11	0.20	0.22	0.10	0.19	0.21	0.11
SR	1.75	-0.43	4.16	0.97	0.33	0.98	0.86	0.22	1.26	0.30	0.26	0.02
	Day +3 Portfolios						Day +4 Portfolios					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.16	0.09	0.07	0.06	0.08	-0.01	0.12	0.10	0.02	0.09	0.08	0.01
Std	0.20	0.22	0.09	0.19	0.22	0.11	0.20	0.22	0.09	0.20	0.21	0.11
SR	0.80	0.39	0.81	0.34	0.36	-0.12	0.62	0.48	0.21	0.46	0.40	0.06
	Day +5 Portfolios						Day +6 Portfolios					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.15	0.16	-0.01	0.11	0.05	0.06	0.11	0.13	-0.02	0.08	0.05	0.02
Std	0.20	0.22	0.09	0.19	0.21	0.10	0.20	0.22	0.09	0.19	0.20	0.10
SR	0.75	0.72	-0.12	0.56	0.25	0.54	0.54	0.58	-0.22	0.39	0.25	0.24

Note: This table reports the LLaMA2's performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted on sentiment scores and their long (L) and short (S) legs in four consecutive days.

Figure 7: Heterogeneous News Assimilation: Size and Volatility



Note: This figure compares average one-day holding period returns to the news trading strategy as a function of when the trade is initiated. We consider daily open-to-open returns initiated from one to 6 days following the announcement. We report equal-weighted portfolio average returns (in basis points per day), with 95% confidence intervals given by the shaded regions.

- For stocks associated with recent news, portfolio signals are generated using a combination of sentiment scores and probability-adjusted past 5-day cumulative returns (calculated via logistic probability).
- For stocks without news tagged to them, the signal is simply the probability-adjusted returns.

The results are summarized in Table 15, where the past returns portfolio in the first row's left panel has been adjusted to ensure the long leg consistently represents the higher returns. Although only a limited number of stocks are tagged with news, it is still apparent that all LLMs outperform the past returns portfolio across the entire market. In contrast, word-based models only achieve a Sharpe Ratio comparable to the past returns portfolio, suggesting that the impact of words alone is limited.

Table 15: Performance of Portfolios based on with Past Returns in Entire Market

	Entire Market						ChatGPT					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	0.01	0.32	0.30	0.08	0.21	0.36	-0.00	0.36	0.22	0.08	0.13
Std	0.23	0.18	0.14	0.27	0.20	0.18	0.22	0.18	0.12	0.20	0.20	0.08
SR	1.46	0.06	2.33	1.11	0.41	1.19	1.63	-0.01	2.96	1.09	0.42	1.60
	LLaMA2						LLaMA					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.36	-0.00	0.36	0.21	0.08	0.13	0.35	0.01	0.35	0.22	0.09	0.13
Std	0.22	0.18	0.12	0.20	0.20	0.09	0.22	0.18	0.12	0.20	0.19	0.08
SR	1.62	-0.00	2.92	1.06	0.43	1.48	1.60	0.04	2.86	1.09	0.46	1.63
	RoBERTa						BERT					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.36	0.00	0.36	0.21	0.09	0.12	0.36	0.00	0.35	0.22	0.09	0.13
Std	0.22	0.18	0.12	0.20	0.20	0.08	0.22	0.18	0.12	0.20	0.20	0.09
SR	1.62	0.01	2.91	1.09	0.48	1.40	1.61	0.02	2.87	1.09	0.43	1.52
	SESTM						LMMD					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.32	0.03	0.30	0.29	0.11	0.18	0.32	0.06	0.26	0.28	0.14	0.14
Std	0.22	0.19	0.12	0.25	0.19	0.14	0.22	0.19	0.12	0.25	0.18	0.12
SR	1.45	0.15	2.37	1.13	0.55	1.31	1.43	0.30	2.28	1.10	0.76	1.15

Note: This table presents the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted based on sentiment scores and their respective long (L) and short (S) positions. The first row's left panel shows the performance of entire market portfolios sorted using past 5-day cumulative returns. The rest of the panels present the performance of portfolios constructed from sentiment scores derived from LLMs and word-based models. For stocks associated with recent news, portfolio signals are generated using a combination of sentiment scores and probability-adjusted past 5-day cumulative returns (calculated via logistic probability). In cases where no news is tagged to a stock, the signal is simply the probability-adjusted returns. The models employed include ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, SESTM, and LMMD.

The effects are more marked when considering stocks associated with news, as shown in Table 16. Here, using past returns combined with sentiment scores derived from news articles significantly outperform the past returns portfolio with notable improvements in performance. Specifically, LLaMA2 achieves a Sharpe ratio of 4.43, and ChatGPT reaches 5.03. Additionally, when augmented by past returns, all models display an improved Sharpe Ratio compared to those based solely on sentiment scores, as detailed in Table 6). We also present the results for using 1-day close-to-close return in Table IA12 and Table IA13 in Appendix.

3.7 Disagreement among Strategies

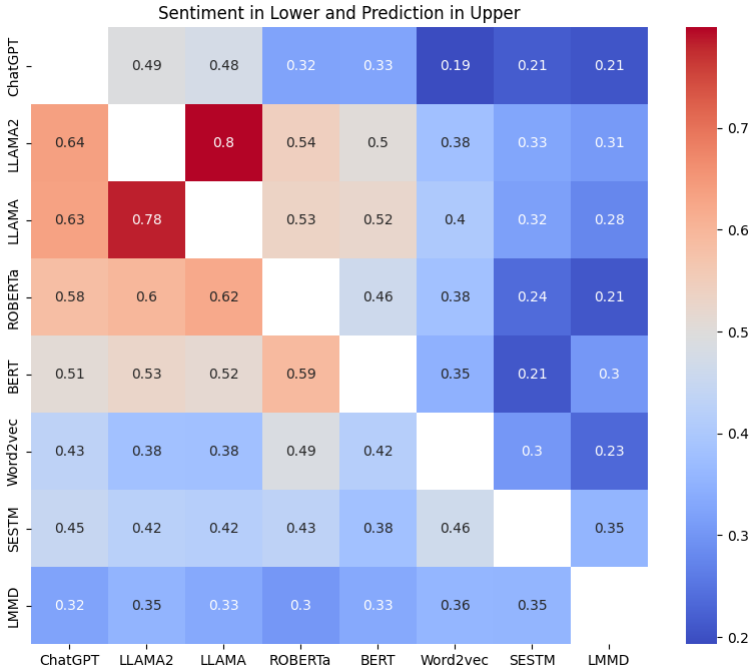
We then explored disagreements among different investment strategies by analyzing the pairwise correlations of daily portfolio returns from different models, with these correlations visualized in a heatmap in Figure 8. The results reveal that word-based models display low correlations with each other, suggesting a lack of consensus on market movements even within this model category. In

Table 16: Performance of Portfolios based on with Past Returns Conditional on Stocks with News

	Stocks with news						ChatGPT					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.35	0.06	0.29	0.29	0.08	0.20	0.38	-0.16	0.54	0.22	0.04	0.18
Std	0.27	0.23	0.18	0.30	0.23	0.24	0.21	0.22	0.11	0.19	0.22	0.11
SR	1.29	0.25	1.58	0.95	0.37	0.83	1.86	-0.71	5.03	1.13	0.20	1.58
	LLaMA2						LLaMA					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.40	-0.12	0.52	0.21	0.07	0.14	0.37	-0.08	0.45	0.21	0.09	0.12
Std	0.21	0.23	0.12	0.20	0.21	0.12	0.21	0.22	0.11	0.20	0.22	0.12
SR	1.88	-0.53	4.43	1.06	0.34	1.19	1.79	-0.34	4.00	1.06	0.42	1.00
	RoBERTa						BERT					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.39	-0.11	0.49	0.24	0.08	0.16	0.38	-0.07	0.44	0.20	0.06	0.14
Std	0.21	0.22	0.11	0.20	0.22	0.12	0.21	0.22	0.11	0.19	0.21	0.11
SR	1.84	-0.48	4.46	1.22	0.36	1.38	1.80	-0.31	4.13	1.01	0.28	1.22
	SESTM						LMMD					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.42	-0.03	0.45	0.30	0.05	0.26	0.26	-0.01	0.28	0.17	0.09	0.07
Std	0.26	0.22	0.17	0.27	0.23	0.21	0.20	0.22	0.10	0.18	0.21	0.10
SR	1.63	-0.12	2.61	1.10	0.21	1.21	1.30	-0.06	2.79	0.90	0.44	0.71

Note: This table presents the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted based on sentiment scores and their respective long (L) and short (S) positions. The first row's left panel shows the performance of entire market portfolios sorted using past 5-day cumulative returns. The rest of the panels present the performance of portfolios constructed from sentiment scores derived from LLMs and word-based models. Portfolio signals are generated using a combination of sentiment scores and probability-adjusted past 5-day cumulative returns (calculated via logistic probability). The models employed include ChatGPT, LLaMA2, LLaMA, RoBERTa, Word2vec, SESTM, and LMMD.

Figure 8: Correlation of Returns among Different Strategies



Note: The figure presents the correlation of daily portfolio returns among different strategies. The portfolios are equal-weighted portfolios built on the basis of ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, Word2vec, SESTM, and LMMD models, respectively. The correlation of sentiment portfolios is shown in the lower triangle, while the correlation of portfolios based on predicted returns is shown in the upper triangle.

contrast, correlations among LLMs are generally strong. Every pairwise correlation between LLMs used in sentiment analysis exceeds 0.5. This indicates that LLMs share common characteristics, such as a better ability to understand context.

It is also important to note that the correlation between LLMs and word-based models is notably lower. To further address this discrepancy, we use the SHAP (SHapley Additive exPlanations) method developed by Lundberg and Lee (2017). SHAP is a game-theoretic approach that provides interpretability for machine learning models. SHAP values represent the expected change in predictions based on specific features. For LLaMA2 interpretation, we highlight segments contributing to positive SHAP values in red, while segments with negative SHAP values are highlighted in blue. The intensity of the color indicates the magnitude of the SHAP value. For Word2vec and SESTM, we utilize a waterfall plot that highlights the SHAP values for all features, with blue features pushing predictions towards the negative side and red features towards the positive side.

Figure 9 presents an example that illustrates the disagreement among different methods. The accompanying text reads as follows:

Brussels has warned British Airways owner IAG ICAG.L that its favored strategy to allow it to continue flying freely in and around Europe in the event of a no deal Brexit will not work, the Financial Times reported on Tuesday. After Brexit, European carriers will have to show they are

more than 50 per cent EU owned and controlled to retain flying rights in the bloc, the FT said. IAG, which also owns the Spanish flag carrier Iberia, is registered in Spain but headquartered in the United Kingdom and has diverse global shareholders. The FT said part of IAG's strategy to retain both EU and UK operating rights is to stress that its important individual airlines are domestically owned through a series of trusts rather than being part of the bigger a high proportion of nonEU investors. The FT quoted an unnamed senior EU official as saying, "For IAG, I can't see how it can be a solution." Concerns have been raised with IAG over its post-Brexit ownership structure, the FT quoted a second Brussels official familiar with the conversations as saying. IAG was not immediately available.

Upon analyzing this news article, it becomes apparent that the message conveyed is negative for British Airways. However, both Word2vec and SESTM label this news as positive, primarily due to the presence of the sentiment word "raise," which is often associated with positive returns. By considering the contextual information provided, such as the phrase "concerns have been raised" and the overall content of the paragraph, LLaMA2 accurately identifies the sentiment as negative.

This example highlights the importance of analyzing the context surrounding sentiment words and demonstrates LLaMA2's ability to capture nuanced sentiment by considering the entire article.

3.8 Interpreting Textual Narratives

In comparison to technical signals, news provides a more interpretable and transparent source of information. LLMs, with their advanced ability to understand context, are particularly effective tools for uncovering narratives within news data. Acknowledging the importance of this capability, we have dedicated this section to discussing and deepening our understanding of the empirical results derived from using news data in financial economics.

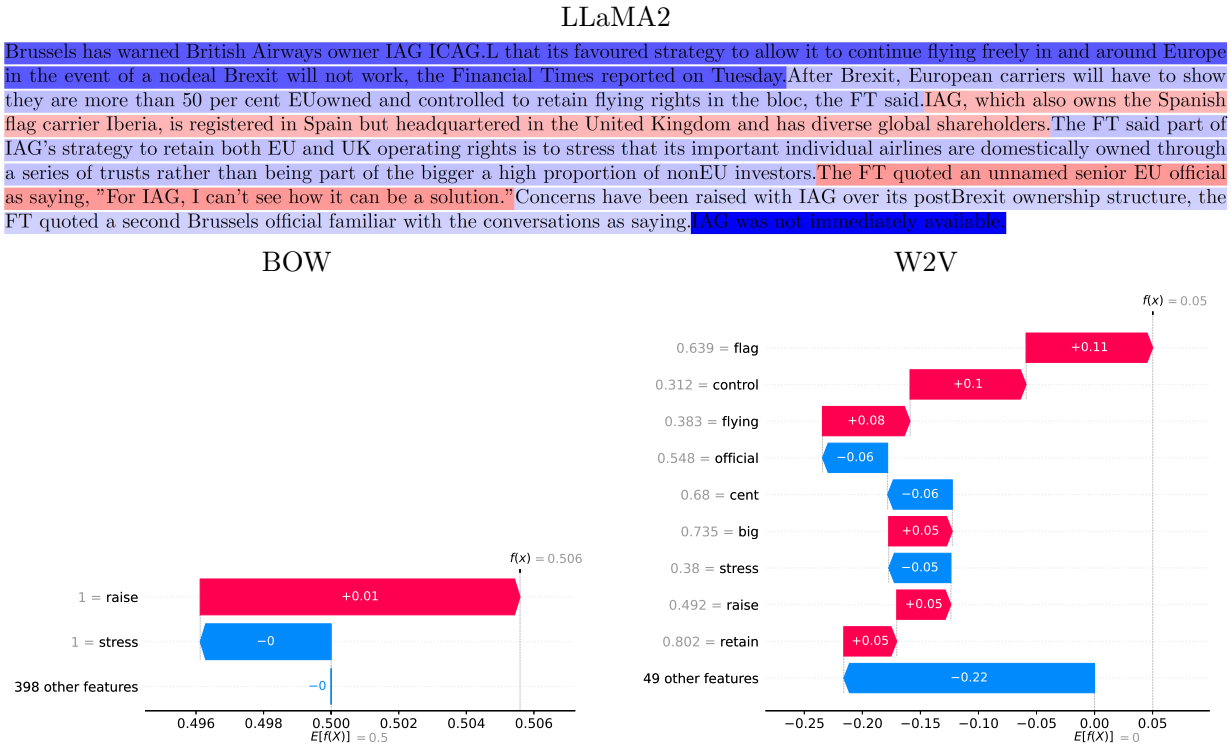
3.8.1 Impact of Negation Words

The example above demonstrates the problem with word-based approaches – the ignorance of context can lead to sentiment classification errors. We now consider a specific context – negation and examine errors caused by the presence of negation words systematically.

In order to highlight the disagreement between LLMs and word-based methods, we specifically focus on news articles that employ negation words. To do so, we construct double-sorted long-short portfolios based on news containing negation words and compare them to portfolios constructed from news without negation words. The performance of these relative equal-weighted portfolios, based on sentiment analysis, is presented in Table 17.

Upon analyzing the results, we observe that almost all LLMs exhibit higher Sharpe ratios for the partition of news articles containing negation words compared to the partition without negation words. Specifically, LLaMA2 increases from 3.29 to 4.18, LLaMA from 3.34 to 4.23, RoBERTa from 3.14 to 4.35, and BERT from 2.94 to 3.37. Though ChatGPT Sharpe Ratio slightly decreases,

Figure 9: Disagreement between LLaMA2 and Word-based Approaches



Note: This figure includes a piece of news for which LLaMA labels as negative, whereas Word2vec and SESTM both recognize as positive. Segments/words that contribute to positive SHAP values are highlighted in red and segments/words with negative SHAP values are highlighted in blue.

we can still get higher expected return for news with negation words (from 0.56 to 0.66). This suggests that there is greater profitability to be found within the partition of news articles that utilize negation words. This can also be confirmed from the Sharpe ratios earned from shorting side of these portfolios.

On the contrary, the word-based models demonstrate even worse portfolio performance when negation words are present. Word2vec's performance decreases from 2.71 to 2.21, while SESTM's performance decreases from 3.00 to 2.58. It becomes apparent that word-based models tend to make more errors or misinterpretations when handling news articles that contain negation words. In contrast, LLMs are able to effectively leverage the presence of negation words to identify hidden predictive signals or seize additional opportunities in the data.

To quantitatively assess the influence of negation words on the performance of word-based models, we further undertake a regression as follows:

$$r_{h,i,t+1}(s_{i,t}^{LLM} - s_{i,t}^{\text{word-based model}}) = \alpha + \beta \text{Negation}_{i,t} + \text{Fixed Effect} + \text{Control Variables}_{i,t} + \epsilon_{i,t}$$

For the left hand side of the regression, our primary focus is on the return-weighted difference between LLM signals and word-based signals. This choice of dependent variable is essential as it enables us to account for the differential signals and their correlation with actual returns, which could also be interpreted as the difference in bets weighted by the success of the bets. We still employ Word2Vec and SESTM as benchmark comparators.

For the right hand side of the regression, we employ negation word count as the measure of negation. Additionally, we introduce several stock-level characteristics as control variables to better determine the impact of negation words. Moreover, we integrate text feature-related variables as controls to enhance the precision of our analysis.

The results of the regression is displayed in Table 18, where negation word count exhibits significant, positive coefficient. Specifically, this effect is accentuated in models that typically exhibit superior performance. This signifies that, when negation words are present, LLMs tend to outperform word-based models in their predictions. We also report results in the appendix by using negation word ratio as the measure (see Table IA14).

These findings underscore the advantage of LLMs over word-based models in the context of negation. While word-based models struggle with the presence of negation words, LLMs demonstrate their ability to capture the underlying information and exploit the predictive signals that may be concealed within such articles. This distinction further highlights the superior performance and interpretability of LLMs in narratives.

3.8.2 Impact of Context Complexity

We then investigate the impact of context complexity on the performance of sentiment analysis models. The properties of headlines and bodies in the same news article can serve as a perfect proxy for context complexity. Because The headline typically serves as a concise summary, whereas the body offers a detailed description, both presumably conveying similar information. To analyze this, we constructed double-sorted long-short portfolios based on sentiment scores derived from either the headline or the body. The performance of these equal-weighted portfolios is detailed in Table 19.

Our analysis reveals a consistent trend where, across all Large Language Models (LLMs), portfolios based on the body of articles consistently outperform those based on headlines. For instance, with LLaMA2, the Sharpe ratio improves from 3.51 for the headline to 4.16 for the body. Conversely, word-based models like SESTM and Word2Vec show better performance when utilizing headlines rather than bodies, suggesting that while the body contains richer information, LLMs are particularly adept at leveraging this to enhance portfolio outcomes. An exception is observed with LMMD, which performs better when analyzing the body, possibly due to the minimal content in headlines leading to zero sentiment scores in most article headlines and poorer performance metrics.

Table 17: Portfolio Performance Comparison on News With and Without Negation Words

	ChatGPT						LLaMA2					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.40	-0.15	0.56	0.43	-0.23	0.66	0.35	-0.07	0.42	0.48	-0.22	0.70
Std	0.21	0.24	0.13	0.21	0.25	0.17	0.21	0.24	0.13	0.22	0.25	0.17
SR	1.96	-0.64	4.34	2.05	-0.90	3.98	1.70	-0.28	3.29	2.21	-0.87	4.18
	LLaMA						RoBERTa					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.36	-0.06	0.43	0.50	-0.21	0.71	0.34	-0.07	0.41	0.51	-0.19	0.70
Std	0.21	0.24	0.13	0.22	0.25	0.17	0.21	0.24	0.13	0.22	0.24	0.16
SR	1.74	-0.27	3.34	2.32	-0.82	4.23	1.64	-0.30	3.14	2.37	-0.76	4.35
	BERT						SESTM					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	-0.03	0.36	0.45	-0.11	0.56	0.33	-0.05	0.38	0.38	-0.01	0.40
Std	0.21	0.23	0.12	0.22	0.25	0.17	0.21	0.24	0.13	0.22	0.25	0.15
SR	1.56	-0.14	2.94	2.06	-0.45	3.37	1.57	-0.22	3.00	1.78	-0.05	2.58
	Word2vec						LMMD					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.28	-0.04	0.32	0.32	-0.01	0.33	0.26	-0.03	0.28	0.29	0.04	0.25
Std	0.21	0.23	0.12	0.22	0.24	0.15	0.21	0.24	0.12	0.21	0.24	0.15
SR	1.35	-0.18	2.71	1.49	-0.02	2.21	1.25	-0.11	2.31	1.35	0.17	1.66

Note: This table presents the differential performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios, which are organized based on sentiment scores and the presence of negation words in context. It includes their respective long (L) and short (S) positions as well.

3.9 The Virtue of Complexity

In the realm of LLMs, there's evidence that more complex models generally outperform simpler ones in NLP tasks. However, this trend does not clearly extend to trading strategies derived from news data, which involve low signal-to-noises and the challenge of extracting relevant information. For trading strategies, there exists the possibility that simpler models might suffice to capture the essential predictive signals for return prediction. The bottleneck might lie in the return prediction issue itself rather than the textual analysis. This will shift the focus to the economic impact of LLMs' size and complexity in such tasks rather than LLM itself.

We evaluated this by testing different-sized LLaMA models, including both LLaMA and LLaMA2 models ranging from 7 billion to 70 billion parameters in Table 20. Performance trends in both sentiment-based and cross-sectional return prediction portfolios showed improvement up to 13 billion parameters (i.e. LLaMA13B and LLaMA2_13B respectively) but not beyond, suggesting a peak in effectiveness at this complexity level. Meanwhile, performance across models remained stable for portfolios based on concise news alerts, which means simpler LLMs are sufficient for shorter texts. In this way, while more complex LLMs may benefit deeper text analysis, they offer limited additional value for tasks involving simpler or shorter texts, like news alerts.

Table 18: Impact of negation word count

	SESTM				Word2Vec			
	LLaMA2	LLaMA	ROBERTa	BERT	LLaMA2	LLaMA	ROBERTa	BERT
neg words count	0.0137***	0.0134***	0.0077*	0.0076*	0.0189***	0.0186***	0.0129***	0.0128***
	(0.0046)	(0.0047)	(0.0046)	(0.0045)	(0.0045)	(0.0046)	(0.0042)	(0.0041)
size	-0.0890***	-0.0793***	-0.0667***	-0.0888***	-0.0620***	-0.0523***	-0.0397***	-0.0617***
	(0.0169)	(0.0171)	(0.0168)	(0.0166)	(0.0166)	(0.0168)	(0.0152)	(0.0150)
BM	0.0046	0.0061	-0.0030	-0.0012	0.0020	0.0035	-0.0055	-0.0038
	(0.0075)	(0.0076)	(0.0075)	(0.0074)	(0.0074)	(0.0075)	(0.0068)	(0.0067)
liquidity	0.0420***	0.0423***	0.0355***	0.0272***	0.0630***	0.0633***	0.0565***	0.0482***
	(0.0105)	(0.0106)	(0.0104)	(0.0103)	(0.0104)	(0.0105)	(0.0095)	(0.0094)
IdioRisk	0.0173***	0.0114*	0.0204***	0.0086	0.0381***	0.0322***	0.0412***	0.0294***
	(0.0065)	(0.0065)	(0.0064)	(0.0064)	(0.0064)	(0.0065)	(0.0058)	(0.0058)
sic2D	-0.0393***	-0.0268*	-0.0328**	-0.0114	-0.0376**	-0.0250*	-0.0311**	-0.0097
	(0.0151)	(0.0152)	(0.0149)	(0.0147)	(0.0148)	(0.0150)	(0.0135)	(0.0133)
Constant	0.0259***	0.0207***	0.0191***	0.0221***	0.0186***	0.0134**	0.0118**	0.0148***
	(0.0053)	(0.0054)	(0.0053)	(0.0052)	(0.0052)	(0.0053)	(0.0048)	(0.0047)
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of obs	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769
Adj R-squared	0.0029	0.0035	0.0046	0.0047	0.0048	0.0036	0.0032	0.0030

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents regression results examining the impact of negation word count on the difference between LLM and word-based model performance with firm-level characteristics as controls. We use LLM's signals minus word-based signals multiplied by next period return as the dependent variable. The first 4 columns show the results with SESTM as a word-based model benchmark, and the rest 4 columns show the results with Word2Vec as a word-based model benchmark

Table 19: Article Healine vs Article Body

	ChatGPT						LLaMA2					
	Headline			Body			Headline			Body		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	-0.09	0.42	0.34	-0.14	0.48	0.32	-0.04	0.36	0.35	-0.10	0.45
Std	0.20	0.22	0.10	0.20	0.22	0.10	0.20	0.22	0.10	0.20	0.23	0.11
SR	1.65	-0.41	4.12	1.71	-0.62	4.62	1.58	-0.19	3.51	1.75	-0.43	4.16
	LLaMA						RoBERTa					
	Headline			Body			Headline			Body		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	-0.02	0.35	0.34	-0.07	0.41	0.35	-0.00	0.35	0.33	-0.06	0.39
Std	0.20	0.22	0.10	0.20	0.23	0.11	0.21	0.22	0.10	0.20	0.22	0.10
SR	1.62	-0.11	3.51	1.67	-0.33	3.89	1.69	-0.02	3.47	1.62	-0.29	3.75
	BERT						Word2vec					
	Headline			Body			Headline			Body		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.31	-0.02	0.33	0.32	-0.04	0.36	0.35	-0.01	0.37	0.29	-0.01	0.30
Std	0.20	0.21	0.10	0.20	0.22	0.10	0.21	0.23	0.11	0.21	0.22	0.10
SR	1.55	-0.09	3.48	1.59	-0.19	3.60	1.70	-0.06	3.32	1.41	-0.05	3.06
	SESTM						LMMD					
	Headline			Body			Headline			Body		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.38	-0.02	0.40	0.31	-0.03	0.34	0.17	-0.09	0.25	0.24	0.01	0.22
Std	0.21	0.22	0.10	0.20	0.22	0.10	0.23	0.25	0.15	0.20	0.23	0.10
SR	1.84	-0.07	4.05	1.53	-0.14	3.43	0.74	-0.35	1.66	1.18	0.06	2.29

Note: This table presents the comparative performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios, stratified based on sentiment scores derived from either the headline or the body of articles. It includes respective long (L) and short (S) positions across ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, Word2vec, SESTM, and LMMD.

Table 20: Sharpe Ratios Comparison between Portfolios Built on Variants of LLaMA Model

	EW					VW				
	Article	Alert				Article	Alert			
		All	TS1	TS2	Rest		All	TS1	TS2	Rest
Portfolios based on Sentiment Analysis										
LLAMA7B	3.93	5.40	5.27	3.45	2.68	1.12	2.39	2.46	0.52	0.79
LLAMA13B	3.89	5.48	5.32	3.58	2.93	1.04	2.39	2.42	0.60	0.94
LLAMA33B	3.54	5.51	4.77	3.60	3.02	1.03	2.29	2.39	0.64	0.56
LLAMA65B	2.73	4.63	4.03	2.80	2.07	0.43	1.99	1.95	0.87	0.31
LLAMA2_7B	4.07	5.29	5.42	3.60	3.16	1.06	2.41	2.87	0.48	0.61
LLAMA2_13B	4.16	5.77	5.60	3.92	3.20	0.98	2.54	2.74	0.55	0.91
LLAMA2_70B	4.24	5.95	5.63	3.78	3.60	1.14	2.37	2.75	0.57	0.73
Portfolios based on Return Prediction										
LLAMA7B	4.90	4.13	4.08	2.51	1.96	1.00	1.47	1.37	0.93	0.95
LLAMA13B	5.17	4.52	4.53	3.33	2.23	1.35	1.87	1.95	1.38	1.05
LLAMA33B	4.37	3.78	3.14	2.50	2.22	0.86	1.37	1.01	0.66	0.83
LLAMA65B	3.05	2.18	1.96	1.34	1.17	0.82	0.92	0.59	0.59	0.43
LLAMA2_7B	5.15	4.03	4.18	3.06	2.36	1.05	1.72	1.61	1.19	1.16
LLAMA2_13B	5.31	3.99	3.75	2.63	2.63	1.32	1.17	1.38	1.02	1.04
LLAMA2_70B	4.61	3.91	3.75	3.00	2.54	1.07	0.96	1.10	0.66	0.67

Note: The table reports Sharpe ratios of the quintile long-short portfolios built on LLaMA models, with the number of parameters 7 billion, 13 billion (baseline), 33 billion, 65 billion , and LLaMA2 model with the number of parameters 7 billion, 13 billion and 70 billion respectively. The top panels reports the Sharpe ratios of portfolio based on sentiment scores and the bottom panel reports the Sharpe ratios of portfolio based on predicted returns. Column "TS1" refers to portfolios that only rely on take sequence 1 alerts, "TS2" only take sequence 2 alerts, "Rest" the remaining alerts, and "All" all alerts.

3.10 Polyglot Evidence from International Equity Markets

To further assess the effectiveness of our strategy in international markets, we have implemented it in the US market and constructed zero-net-investment portfolios for each country. These portfolios long the top quintile of stocks and short the bottom quintile based on predicted returns, with the corresponding Sharpe ratios detailed in Table 21. Notably, for models catering to non-English languages, LLMs generally outperform word-based models. Specifically, RoBERTa leads with an average equal-weighted Sharpe ratio of 1.07, followed closely by LLaMA at 1.06, LLaMA2 at 0.95, and BERT at 0.88. In comparison, word-based models like Word2vec and SESTM achieve lower ratios of 0.52 and 0.85, respectively.

This trend holds even when excluding the United States from the analysis, indicating that LLMs consistently outperform word-based models across various international markets by capturing return predictability more effectively. These findings highlight the potential advantages of utilizing LLMs, particularly RoBERTa and LLaMA, for building investment portfolios internationally, as these models provide superior risk-adjusted returns.

However, an unexpected observation is that LLaMA2 does not outperform word-based models when data from the US is excluded. This underperformance is particularly noticeable in countries like Portugal, Greece, and the Netherlands, where the limited number of stocks with news coverage renders the portfolio performance in these countries less reliable. To further investigate, we ana-

lyzed the cross-sectional differences in the Sharpe ratios, illustrated in Figure 10. This figure plots the relationship between the logarithmic average number of stocks with available news per day and the equal-weighted Sharpe ratios across all 16 countries. It reveals a consistent positive correlation between the profitability of investment strategies and the availability of news-covered stocks, irrespective of the model used. The United States, with the highest number of stocks available for trade, shows the highest Sharpe ratios, underscoring the beneficial impact of extensive news coverage on return prediction strategies. Following the US, the United Kingdom also shows high Sharpe ratios due to its substantial stock and news volume. Conversely, smaller markets like the Netherlands and Greece, which have fewer than ten stocks covered by news, generally exhibit negative Sharpe ratios, highlighting the challenges of executing profitable return prediction strategies in markets with limited stock availability.

In conclusion, these findings emphasize the crucial role of stock availability and extensive news coverage in enhancing the profitability of return prediction strategies across different markets. Larger markets provide a favorable environment for generating positive risk-adjusted returns, while smaller markets may present significant challenges due to limited stock and news coverage.

Table 21: Sharpe Ratios of Portfolios based on the Cross-Section of Return Predictions

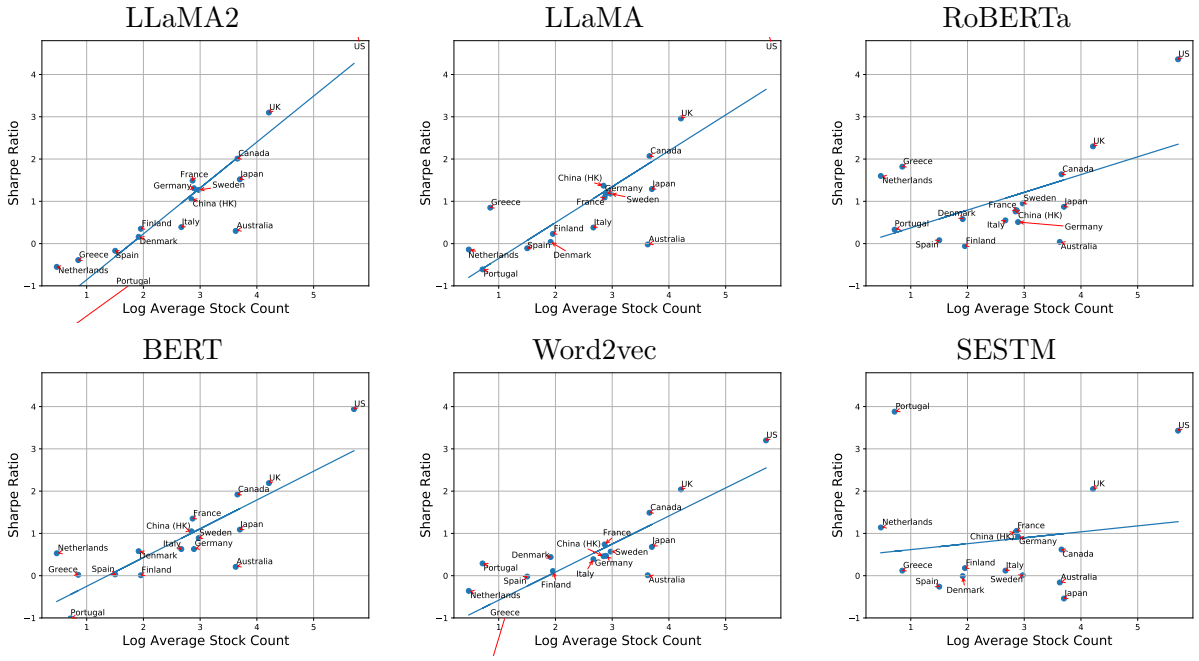
	LLaMA2		LLaMA		RoBERTa		BERT		Word2vec		SESTM		LMMD	
	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW
US	5.31	1.32	5.17	1.35	4.36	1.17	3.94	0.68	3.20	0.74	3.43	0.86	2.29	0.41
UK	3.10	1.64	2.96	1.34	2.30	0.60	2.19	1.22	2.04	0.81	2.05	0.73	0.82	0.34
Australia	0.30	0.25	-0.02	0.02	0.04	-0.01	0.21	0.07	0.01	-0.02	-0.16	-0.11	0.37	0.02
Canada	2.01	1.27	2.07	0.76	1.64	0.60	1.92	0.89	1.49	0.79	0.62	0.33	0.76	0.37
China (HK)	1.05	0.54	1.37	1.07	0.76	0.55	1.05	0.87	0.46	0.33	1.03	0.76		
Japan	1.52	0.56	1.29	0.65	0.87	0.39	1.09	0.54	0.68	0.45	-0.54	-0.29		
Germany	1.31	0.63	1.18	0.40	0.51	0.23	0.63	0.34	0.47	0.20	0.92	0.70		
Italy	0.39	0.08	0.38	0.14	0.55	0.13	0.63	0.06	0.39	-0.04	0.12	0.21		
France	1.49	0.63	1.09	0.67	0.79	0.40	1.35	0.72	0.74	0.19	1.06	0.14		
Sweden	1.27	0.76	1.18	0.67	0.95	0.58	0.89	0.21	0.57	0.59	0.01	0.53		
Denmark	0.16	0.02	0.04	-0.06	0.58	0.49	0.58	0.53	0.44	0.31	-0.01	-0.16		
Spain	-0.17	-0.15	-0.11	0.05	0.08	0.02	0.03	0.07	-0.02	0.14	-0.26	-0.43		
Finland	0.35	0.09	0.23	0.01	-0.06	-0.21	0.01	0.01	0.11	0.11	0.18	-0.06		
Portugal	-1.99	-2.00	-0.61	-0.62	0.33	0.34	-1.01	-1.04	0.29	0.29	3.88	3.86		
Greece	-0.39	-0.39	0.85	0.85	1.82	1.82	0.02	0.02	-2.14	-2.14	0.12	0.12		
Netherlands	-0.55	-0.55	-0.14	-0.14	1.60	1.60	0.53	0.53	-0.36	-0.36	1.14	1.14		
Mean	0.95	0.29	1.06	0.45	1.07	0.54	0.88	0.36	0.52	0.15	0.85	0.52	1.06	0.29
Mean (Excluding US)	0.66	0.23	0.79	0.39	0.85	0.50	0.67	0.34	0.34	0.11	0.68	0.50	0.65	0.24
Median (Excluding US)	0.39	0.25	0.85	0.40	0.76	0.40	0.63	0.34	0.44	0.20	0.18	0.21	0.76	0.34

Note: The table reports Sharpe ratios of long-short portfolios for international market portfolios. The portfolios are built on the basis of LLaMA, LLaMA2, RoBERTa, BERT and Word2vec model, respectively, using cross-sectional predicted return as sorting variables.

3.11 Advanced Machine Learning Models

In this section, we evaluate the performance of advanced machine learning models for news-based trading strategies. We use the LLaMA2 model as a benchmark and test various models, including RIDGE, LASSO, Random Forest, and Neural Networks, to predict stock returns from news data. For Neural Networks, a simple feedforward architecture with three hidden layers is employed. We construct quintile portfolios based on the predicted cross-sectional returns and calculate the Sharpe

Figure 10: Sharpe Ratios vs the Number of Stocks for International Markets



Note: The figures show annualized out-of-sample Sharpe ratios of equal-weight portfolios versus the logarithm of the average number of stocks available at each rebalance for each country.

ratios for each model, as presented in Table 22. The results indicate that Neural Networks, the most complex model tested, deliver superior performance with an equal-weighted Sharpe ratio of 5.83 and a value-weighted Sharpe ratio of 1.44. However, the performance does not always correlate directly with model complexity across all weighting schemes, as the Random Forest Model cannot outperform the LASSO model despite its higher complexity. These results suggest that the performance of advanced machine learning models in news-based trading strategies is not solely determined by model complexity, but also by the specific characteristics of the data and the model's ability to capture the underlying signals effectively.

3.12 Transaction Cost Analysis

In previous sections, our evaluations focused primarily on providing economic context and magnitude to the predictive content of each model without accounting for transaction costs. However, to truly gauge the practical viability of our trading strategies, especially given their high turnover, it is essential to incorporate trading costs into our performance assessment.

To accurately reflect the net performance of our strategies, we have constructed a transaction cost model. This model accounts for the differential costs associated with trading large and small stocks. We assume daily transaction costs of 10 basis points (bps) for large stocks and 20 bps for smaller stocks, below the NYSE 20% breakpoints, aligning with the average costs experienced by

Table 22: Advanced Machine Learning Models Performance

	RF						LASSO					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.32	-0.03	0.34	0.19	0.09	0.10	0.38	-0.04	0.42	0.15	0.07	0.08
Std	0.20	0.22	0.11	0.19	0.20	0.11	0.19	0.21	0.10	0.18	0.19	0.10
SR	1.55	-0.12	3.25	0.97	0.45	0.88	2.01	-0.21	4.14	0.84	0.37	0.78
	RIDGE						NN					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.46	-0.11	0.57	0.23	0.09	0.14	0.53	-0.15	0.68	0.24	0.07	0.17
Std	0.21	0.22	0.11	0.20	0.20	0.11	0.21	0.22	0.12	0.21	0.20	0.12
SR	2.22	-0.50	5.31	1.14	0.44	1.32	2.49	-0.66	5.83	1.15	0.36	1.44

This table presents the comparative performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios overall four machine learning models: Random Forest (RF), LASSO, Ridge Regression (RIDGE), and Neural Networks (NN).

large asset managers as reported by Frazzini et al. (2018). In addition, our model introduces a turnover reduction strategy. This strategy, which we refer to as Exponentially-Weighted Calendar Time (EWCT) and derive from Ke et al. (2021), limits portfolio turnover to a fixed proportion each period. It also assigns exponentially decaying weights to stocks based on the recency of their appearance in news, effectively extending their holding period.

Tables 23 present EWCT portfolio performances under varying turnover limits ($\gamma = 0.9$ to $\gamma = 0.1$) based on alert sentiment analysis. Net returns increase from 1.03 bps to 7.79 bps as γ increases, and the net Sharpe ratio peaks at 1.54 for a γ value of 0.4 before slightly declining as turnover further increases. Our analysis reveals that increased turnover restrictions tend to lower the gross Sharpe ratio due to a loss in the immediacy of predictive signals. However, this negative impact is mitigated by a corresponding reduction in trading costs. While gross Sharpe ratios improve with higher γ values, the influence of transaction costs is significant. At lower γ levels, the turnover is sufficiently minimal to sustain a positive net Sharpe ratio, but at higher levels, even substantial gross Sharpe ratios are negated by transaction costs.

4 Conclusion

In this study, we have harnessed the power of state-of-the-art LLMs in NLP to obtain contextualized embeddings of news text. Our comprehensive analysis has encompassed a wide range of news articles from 16 countries, written in 13 different languages, providing compelling evidence of the predictability of returns based on news data.

Our analysis has demonstrated that news significantly influences immediate price reactions, suggesting that markets respond contemporaneously to new information. However, this integration is not always immediate or efficient, leading to a momentum effect where prices adjust over a period,

Table 23: Performance Analysis of Trading Strategies with Transaction Cost

	Turnover	Gross Return	Gross Sharpe Ratio	Net Return	Net Sharpe Ratio
0.10	10.05	3.54	4.83	1.03	1.40
0.20	20.17	7.08	5.19	2.05	1.50
0.30	30.37	10.59	5.37	3.02	1.53
0.40	40.62	14.07	5.49	3.94	1.54
0.50	50.94	17.51	5.57	4.81	1.53
0.60	61.32	20.91	5.64	5.64	1.52
0.70	71.78	24.29	5.69	6.42	1.50
0.80	82.34	27.64	5.72	7.14	1.48
0.90	93.12	30.96	5.75	7.79	1.45

Note: This table presents the performance of the EWCT portfolio with varying turnover limits (γ), using alert sentiment analysis based on the LLaMA2 model. Turnover is represented as the average daily turnover percentage. Returns are expressed in daily basis points, and the Sharpe ratio is annualized. Gross returns and Sharpe ratios are calculated before accounting for transaction costs, whereas net returns and Sharpe ratios incorporate these costs.

highlighting potential short-run predictability. This delay in price adjustment underscores inefficiencies in how markets process and react to new information.

Large Language Models (LLMs) have proven to be more effective tools for textual analysis compared to traditional word-based method and number-based technical analysis. By transforming unstructured text into structured vectorized embeddings, LLMs manage to retain more context and semantic content. This capability is a significant advancement over older techniques that primarily depend on word frequency counts, offering a more nuanced and comprehensive analysis.

Overall, the use of LLMs in extracting and processing information from textual data represents a significant step forward for empirical finance. It not only enhances our understanding of how news affects market dynamics but also opens up new avenues for research into the predictive power of textual analysis within financial markets. The findings from this study lay the groundwork for further exploration into modern textual analysis techniques and their application in understanding and predicting financial market behaviors.

References

Araci, Dogu, 2019, Finbert: Financial sentiment analysis with pre-trained language models, *arXiv preprint arXiv:1908.10063* .

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, 2014, Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473* .

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* 5, 135–146.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, 2020, Language models are few-shot learners, in H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901 (Curran Associates, Inc.).

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu, 2020, The structure of economic news, Technical report, National Bureau of Economic Research.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, 2020, Unsupervised cross-lingual representation learning at scale, in Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 8440–8451 (Association for Computational Linguistics).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Firth, J. R., 1957, A synopsis of linguistic theory 1930-1955, in *Studies in Linguistic Analysis*, 1–32 (Oxford: Philological Society).

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov, 2018, Learning word vectors for 157 languages, in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Harris, Zellig S, 1954, Distributional structure, *Word* 10, 146–162.

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

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu, 2019, Predicting returns with text data, Technical report, National Bureau of Economic Research.

Kelly, Bryan, Asaf Manela, and Alan Moreira, 2018, Text selection, Working paper.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 2019, RoBERTa: A robustly optimized BERT pretraining approach, *arXiv preprint arXiv:1907.11692* .

LOUGHRAN, TIM, and BILL MCDONALD, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance* 66, 35–65.

Lundberg, Scott M, and Su-In Lee, 2017, A unified approach to interpreting model predictions, in *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning, 2015, Effective approaches to attention-based neural machine translation, *arXiv preprint arXiv:1508.04025* .

Malo, Pekka, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala, 2014, Good debt or bad debt: Detecting semantic orientations in economic texts, *Journal of the Association for Information Science and Technology* 65, 782–796.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* .

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin, 2018, Advances in pre-training distributed word representations, in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 2018, Deep contextualized word representations, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237 (Association for Computational Linguistics, New Orleans, Louisiana).

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., 2018, Improving language understanding by generative pre-training .

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., 2019, Language models are unsupervised multitask learners, *OpenAI blog* 1, 9.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams, 1986, Learning representations by back-propagating errors, *Nature* 323, 533–536.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Ar-

mand Joulin, Edouard Grave, and Guillaume Lample, 2023, Llama: Open and efficient foundation language models.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 2017, Attention is all you need, in *Advances in neural information processing systems*, 5998–6008.

Yang, Yi, Mark Christopher Siy Uy, and Allen Huang, 2020, Finbert: A pretrained language model for financial communications, *arXiv preprint arXiv:2006.08097* .

Appendix A Additional Tables and Figures

A.1 Truncation

Our Large Language Model (LLM) utilizes a partial token integration strategy, leading to truncation for longer text sequences. This truncation might result in the loss of important information, as our model processes only a portion of the complete token sequence, leaving out potentially crucial data found in the truncated sections.

A key feature of our LLaMA2 model is its ability to process up to 4096 tokens, significantly exceeding the typical 512-token limit seen in other LLMs. Analysis using the LLaMA2 tokenizer shows that 47.9% of articles contain fewer than 512 tokens, and an additional 49.8% have between 512 and 4096 tokens. This increased capacity allows for a more detailed examination of the effects of token truncation and whether utilizing more tokens can enhance prediction accuracy.

To investigate this, we conducted a comparative analysis of the LLaMA2 models using different token lengths—512 versus 4096 tokens. The results, shown in Table IA1, compare sentiment classification accuracy and cross-sectional prediction correlation. We also report the performance of portfolios based on these models. The results indicate that the LLaMA2 model with 512 tokens is the most suitable for model comparisons, as the differences in performance is not significant, even LLaMA2 will perform better than LLaMA2TK4096.

Table IA1: Performance Comparison with Different Token Truncation

	LLaMA2		LLaMA2TK4096	
	Acc.	Corr.	Acc.	Corr.
token length ≤ 512	54.15	2.04	54.06	1.39
$512 <$ token length < 4096	54.03	2.67	53.93	2.25
token length ≥ 4096	53.28	1.33	53.42	3.59
Overall	54.07	2.48	53.99	2.00
	Return	Sharpe Ratio	Return	Sharpe Ratio
Sentiment	45.22	4.16	41.96	3.81
Prediction	57.14	5.31	49.93	4.80

Note: This table reports the performance of LLaMA2 and LLaMA2TK4096 models with different token number. The upper part of the table reports the sentiment classification accuracy and cross-sectional prediction correlation of the models with different token number. The lower part of the table reports the portfolio performance of the models.

To further explore the impact of token truncation, we categorized articles into three groups based on token count—fewer than 512, between 512 and 4096, and more than 4096—and analyzed the correlation between model signal differences and token type. We also adjusted for individual stock characteristics and included fixed effects and control variables in our regression model:

$$\text{Model Signal Difference}_{i,t} = \alpha + \beta \text{Token Type}_{i,t} + \text{Fixed Effect} + \text{Control Variables}_{i,t} + \epsilon_{i,t}$$

The regression results, as detailed in Table IA2, indicate that the coefficient of token type is not

Table IA2: Difference in LLaMA2 and LLaMA2TK4096

token type	0.0018	0.0015	0.0016	0.0041	0.0042
	(0.0122)	(0.0122)	(0.0122)	(0.0122)	(0.0122)
size	-0.0651*	-0.0885**	-0.0184	0.0562	0.0559
	(0.0392)	(0.0404)	(0.0446)	(0.0479)	(0.0479)
IdioRisk		-0.0409**	-0.0409**	-0.0155	-0.0149
		(0.0174)	(0.0174)	(0.0184)	(0.0184)
BM			0.0790***	0.0834***	0.0827***
			(0.0213)	(0.0214)	(0.0214)
liquidity				-0.1268***	-0.1280***
				(0.0295)	(0.0295)
sic2D					-0.0830*
					(0.0427)
Constant	0.0208	0.0224	0.0072	0.0238	0.0222
	(0.0244)	(0.0244)	(0.0247)	(0.0250)	(0.0250)
Stock FE	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Number of obs	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769
Adj R-squared	0.0066	0.0066	0.0066	0.0066	0.0066

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents regression results examining the impact of truncation in token length on the difference between LLaMA2 model truncated at 512 token and full LLaMA2TK4096 model performance. We use LLaMA2TK4096 signals minus LLaMA2 signals multiplied by next period return as the dependent variable. We include size, idiosycratic risk, book-to-market ratio, liquidity, industry and other controls as independent variables. The sample period is from 2004 to 2019.

statistically significant. This finding suggests that the LLaMA2 model truncated at 512 tokens is sufficient for model comparisons, confirming that the full token length of 4096 does not necessarily provide a significant advantage.

A.2 Alpha Regression

To further evaluate the performance of our trading strategies, we conducted an alpha regression analysis, regressing portfolio returns against established financial factors. These included the components of the Fama-French three-factor model, alongside Momentum, and both short-run and long-run reversal factors. It's important to note that, since our portfolio strategy involves open to open price transactions, and most benchmarks use close to close metrics, we adapted our approach by formulating our portfolio at market open but calculating daily portfolio values at market close for comparison with standard benchmarks. The results are displayed in Table IA3.

The results of this regression revealed several key insights. Generally, both model-based and LLM-

Table IA3: Sentiment Performance of Alpha Test on Equal-weighted Portfolio

	ChatGPT	LLaMA2	LLaMA	ROBERTa	BERT	SESTM	W2V	LMMD
Mkt	-0.0517***	-0.0345***	-0.0262**	-0.0383***	-0.0302***	-0.0286***	-0.0579***	0.0050
	(0.0107)	(0.0111)	(0.0110)	(0.0106)	(0.0104)	(0.0102)	(0.0100)	(0.0101)
SMB	-0.0189	-0.0299	-0.0337	-0.0379*	-0.0303	-0.0090	0.0056	-0.0145
	(0.0200)	(0.0206)	(0.0205)	(0.0198)	(0.0194)	(0.0191)	(0.0187)	(0.0188)
HML	-0.0011	-0.0131	-0.0151	-0.0695***	-0.0401*	0.0052	-0.0441**	-0.0458**
	(0.0227)	(0.0234)	(0.0232)	(0.0225)	(0.0220)	(0.0216)	(0.0212)	(0.0213)
LT Rev	-0.0240	-0.0687***	-0.0854***	-0.0175	-0.0271	-0.0044	0.0170	-0.0258
	(0.0242)	(0.0249)	(0.0247)	(0.0239)	(0.0234)	(0.0230)	(0.0226)	(0.0227)
ST Rev	-0.0340**	-0.0338**	-0.0340**	-0.0241*	-0.0211	-0.0102	0.0031	-0.0453***
	(0.0140)	(0.0144)	(0.0143)	(0.0138)	(0.0135)	(0.0133)	(0.0130)	(0.0131)
Mom	0.0166	-0.0100	-0.0102	0.0012	-0.0001	0.0085	-0.0052	0.0005
	(0.0138)	(0.0142)	(0.0141)	(0.0137)	(0.0134)	(0.0132)	(0.0129)	(0.0130)
Constant	0.4870***	0.4557***	0.4130***	0.3972***	0.3665***	0.3478***	0.3072***	0.2250***
	(0.0266)	(0.0274)	(0.0272)	(0.0264)	(0.0258)	(0.0254)	(0.0249)	(0.0250)
Number of obs	3,901	3,902	3,902	3,902	3,902	3,902	3,902	3,902
Average Return	0.4811	0.4524	0.4111	0.3930	0.3634	0.3448	0.3017	0.2235
Adj R-squared	0.0146	0.0101	0.0103	0.0160	0.0089	0.0025	0.0134	0.0055

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table reports the results of regressing portfolio returns on Fama-French 3 factors, long/short-term reversal and momumtum. The portfolios are built on the basis of ChatGPT, LLaMA2, LLaMA, RoBERTa, BERT, SESTM and Word2vec respectively. All portfolio returns and factors are annulized.

based portfolios showed a negative correlation with the market factor.This trend was particularly pronounced in our more sophisticated strategies, such as LLaMA and LLaMA2, which also exhibited a more pronounced negative correlation with the SMB (Small Minus Big) factor. Notably, all models displayed positive alpha, indicating that our portfolios could generate excess returns after controlling for common market risk factors. Furthermore, the portfolio alphas slightly exceeded the raw average returns. This finding suggests that, beyond their raw performance metrics, these portfolios offer an added layer of value, successfully navigating market risks to deliver superior returns. However, despite the statistical significance, we observe a relatively modest magnitude of the factor coefficients. This implies a limited economic significance and indicates a constrained risk exposure of our models to the market.

A.3 Trade

A.3.1 Summary Statistics

In this section of our paper, we provide additional information about details in trading. We begin our analysis with the summary statistics of our constructed portfolios, which includes both the long and short sides, as well as the combined long-short portfolios. Tables IA4 and Table IA5 present the statistical characteristics of our portfolios, which are constructed based on sentiment analysis and return prediction respectively. These tables cover the number of stocks traded, their liquidity—measured by daily trading volume shares (in unit) sourced from WRDS—and their market capitalization. These statistics are articulated in both average and median values across each model.

In addition, Tables IA6 and IA7 display the average monthly returns from sentiment analysis and return prediction strategies, respectively. This analysis reveals cyclical patterns, particularly elevated

Table IA4: Summary Statistics of Sentiment Analysis Portfolios

	ChatGPT						LLaMA2					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Avg Traded	85	85	170	85	85	170	85	85	170	85	85	170
Median Traded	82	82	164	82	82	164	82	82	164	82	82	164
Avg Liquidity	3.11	4.29	3.71	3.12	4.28	3.70	3.03	4.15	3.59	3.10	3.88	3.49
Median Liquidity	0.82	0.75	0.76	0.82	0.75	0.77	0.79	0.73	0.74	0.75	0.72	0.71
Avg Cap	16.59	13.98	15.28	16.56	13.93	15.24	17.05	13.78	15.41	16.25	13.76	15
Median Cap	2.90	1.43	1.98	2.88	1.44	1.98	3.01	1.42	1.98	2.79	1.46	1.95
	LLaMA						RoBERTa					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Avg Traded	85	85	170	85	85	170	85	85	170	85	85	170
Median Traded	82	82	164	82	82	164	82	82	164	82	82	164
Avg Liquidity	3.04	3.94	3.49	3.35	3.25	3.30	2.84	4.16	3.50	2.86	4.14	3.50
Median Liquidity	0.77	0.73	0.72	0.99	0.56	0.73	0.80	0.72	0.75	0.80	0.73	0.74
Avg Cap	17.36	13.15	15.25	18.94	11.48	15.21	15.42	14.76	15.09	15.46	14.65	15.05
Median Cap	2.81	1.48	1.93	3.60	1.26	2.05	2.75	1.59	2.01	2.71	1.60	2
	BERT						Word2vec					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Avg Traded	85	85	170	85	85	170	84	85	169	85	85	170
Median Traded	82	82	164	82	82	164	82	82	164	82	82	164
Avg Liquidity	2.89	4.06	3.47	2.87	4.07	3.46	3.07	4.36	3.72	3.07	4.35	3.72
Median Liquidity	0.74	0.73	0.71	0.73	0.74	0.71	0.75	0.77	0.74	0.74	0.78	0.73
Avg Cap	15.75	15.29	15.52	15.72	15.11	15.42	16.52	15.79	16.15	16.37	15.64	16
Median Cap	2.56	1.69	1.98	2.52	1.71	1.98	2.61	1.78	2.07	2.52	1.79	2.04
	SESTM						LMMD					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Avg Traded	85	84	169	85	84	169	85	84	169	85	84	169
Median Traded	82	82	164	82	82	164	82	82	164	82	82	164
Avg Liquidity	2.89	4.58	3.73	2.89	4.58	3.73	2.74	4.84	3.78	2.74	4.84	3.78
Median Liquidity	0.66	0.82	0.71	0.66	0.82	0.71	0.63	0.99	0.77	0.63	0.99	0.77
Avg Cap	14.22	16.18	15.20	14.22	16.18	15.20	14.10	19.45	16.77	14.10	19.45	16.77
Median Cap	2.25	1.86	1.97	2.25	1.86	1.97	2.24	2.49	2.22	2.24	2.49	2.22

Note: This table provides summary statistics for sentiment analysis portfolios, detailing the number of stocks traded, their liquidity, and market capitalization. Market capitalization is expressed in billions of USD, while liquidity is measured by the total number of shares sold per day and is presented in millions of shares.

Table IA5: Summary Statistics of Return Prediction Portfolios

	ChatGPT						LLaMA2					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Avg Traded	85	85	170	85	85	170	85	85	170	85	85	170
Median Traded	82	82	164	82	82	164	82	82	164	82	82	164
Avg Liquidity	2.46	3.89	3.18	2.35	4.89	3.62	2.37	4.49	3.43	2.14	5.97	4.04
Median Liquidity	0.56	0.79	0.65	0.57	0.87	0.67	0.52	0.93	0.66	0.53	1.19	0.71
Avg Cap	12.03	16.22	14.13	11.67	18.58	15.13	10.34	18.90	14.63	10.98	23.43	17.20
Median Cap	1.73	1.95	1.74	1.74	2.40	1.93	1.67	2.24	1.76	1.67	3.61	2.11
	LLaMA						RoBERTa					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Avg Traded	85	85	170	85	85	170	85	85	170	85	85	170
Median Traded	82	82	164	82	82	164	82	82	164	82	82	164
Avg Liquidity	2.37	4.50	3.44	1.57	6.75	4.16	2.72	3.68	3.19	2.20	5.07	3.63
Median Liquidity	0.50	0.97	0.66	0.40	1.47	0.68	0.59	0.78	0.65	0.53	0.99	0.67
Avg Cap	10.86	18.95	14.91	6.52	27	16.76	12.61	16.92	14.77	10.75	21.47	16.09
Median Cap	1.58	2.51	1.78	1.19	5.10	1.99	1.85	2.03	1.83	1.64	3.04	1.99
	BERT						Word2vec					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Avg Traded	85	85	170	85	85	170	85	85	170	84	85	169
Median Traded	82	82	164	82	82	164	82	82	164	82	82	164
Avg Liquidity	2.56	4.05	3.31	2.04	4.91	3.47	3	3.54	3.27	2.43	4.69	3.57
Median Liquidity	0.59	0.79	0.66	0.52	0.97	0.66	0.64	0.74	0.67	0.58	0.94	0.70
Avg Cap	11.91	17.73	14.82	9.98	21.17	15.57	14.04	15.90	14.98	11.17	20.27	15.78
Median Cap	1.80	2.05	1.82	1.60	2.82	1.93	2.13	1.79	1.85	1.67	2.85	1.97

Note: This table provides summary statistics for portfolios built on cross-sectional predicted returns, detailing the number of stocks traded, their liquidity, and market capitalization. Market capitalization is expressed in billions of USD, while liquidity is measured by the total number of shares sold per day and is presented in millions of shares.

Table IA6: Average Monthly Return based on Sentiment Analysis

	ChatGPT	LLaMA2	LLaMA	ROBERTa	BERT	Word2vec	SESTM	LMMD
Jan	19.63	9.84	9.44	10.63	12.61	12.76	16.18	12.53
Feb	25.23	24.75	22.13	24.09	21.65	13.12	19.91	8.64
Mar	14.20	17.96	13.11	13.53	14.72	14.26	11.81	10.03
Apr	10.58	12.18	12.59	7.51	7.26	5.15	5.47	4.67
May	24.94	24.24	22.95	19.09	14.04	14.18	18.97	12.02
Jun	9.41	14.76	11.70	12.82	11.12	5.42	14.93	10.57
Jul	23.47	23.94	20.51	20.46	14.20	13.42	14.36	7.36
Aug	24.93	18.71	19.26	15.21	17.53	10.52	10.39	7.39
Sept	12.85	12.70	12.38	7.80	9.80	9.61	8.86	8.01
Oct	23.61	24.42	21.17	22.79	21.36	18.46	13.73	6.67
Nov	25.81	18.80	21.06	18.55	15.75	12.21	18.30	6.85
Dec	15.64	12.97	10.11	14.81	13.01	14.27	11.26	11.36
Overall	19.12	17.94	16.34	15.57	14.40	11.93	13.65	8.86

Note: This table displays the average daily returns of sentiment analysis-based portfolios for each month, quantified in basis points.

Table IA7: Average Monthly Return based on Return Prediction

	ChatGPT	LLaMA2	LLaMA	ROBERTA	BERTLARGE	Word2vec
Jan	15.95	16.95	15.46	14.04	8.01	6.07
Feb	20.42	30.44	27.55	21.24	17.77	20.94
Mar	19.91	23.23	22.27	18.87	12.74	19.15
Apr	9.05	16.68	18.10	10.78	12.98	8.67
May	18.77	28.75	27.36	16.17	19.95	13.49
Jun	19.27	21.38	19.50	12.73	14.42	6.74
Jul	13.19	13.31	15.32	11.76	9.03	3.18
Aug	24.72	29.66	27.56	19.96	17.63	15.56
Sept	15.06	22.58	21.80	21.03	15.02	11.42
Oct	12.59	20.77	20.27	19.25	20.46	15.37
Nov	20.92	20.50	20.25	13.62	8.38	18.20
Dec	15.72	27.76	30.34	20.13	17.27	18.80
Overall	17.15	22.67	22.14	16.60	14.51	13.10

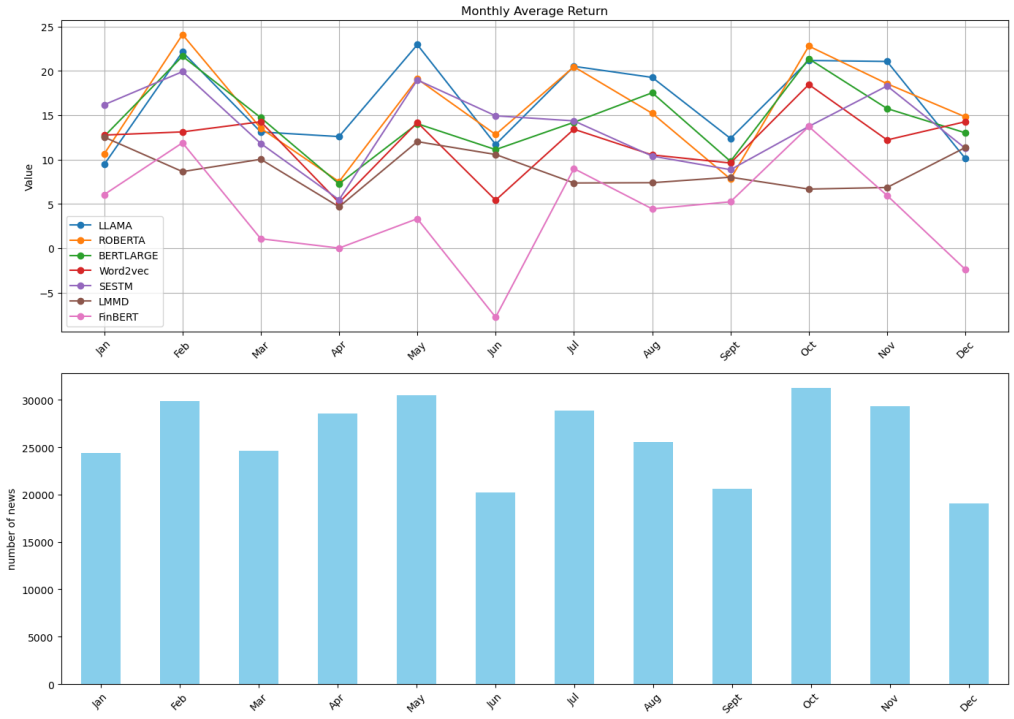
Note: This table displays the average daily returns of cross-sectional predicted return-based portfolios for each month, quantified in basis points.

returns during months like February, May, August, and October, which align closely with the financial reporting calendar. This timing suggests that earnings reports and corporate disclosures significantly influence market dynamics. The pronounced sensitivity of our portfolios to these periodic financial disclosures underscores their responsiveness to market information fluctuations. We further illustrate the findings in Figure IA1 and Figure IA2.

A.3.2 Trading in Russell 1000 Stocks

Initially, our trading universe encompassed all stocks, but we have now narrowed our focus to include only those stocks listed in the Russell 1000 Index. This index, maintained by the FTSE Russell, a subsidiary of the London Stock Exchange Group, represents the largest 1000 publicly traded companies in the United States. It is widely regarded as a benchmark for the performance of large-cap U.S. equities, covering approximately 90% of the U.S. stock market capitalization and

Figure IA1: Sentiment Analysis: Monthly Returns



Note: This figure plots the average monthly returns of portfolios based on sentiment analysis. The top panel shows the average daily returns in basis point for each month of portfolios based on sentiment analysis, while the bottom panel shows the average number of news each month.

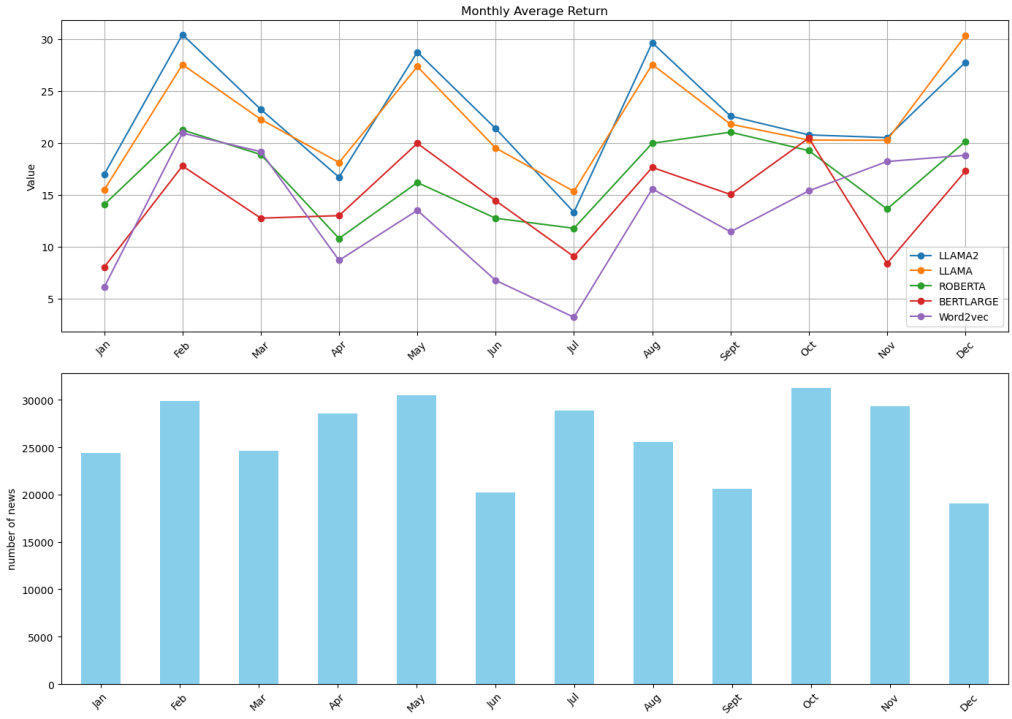
encompassing a diverse range of industries and sectors.

In constructing our portfolio, we initially apply our original methodology, which involves long positions in the top 20% and short positions in the bottom 20% of stocks based on alpha sorting. Subsequently, we further refine our portfolio to include only those stocks that are part of the Russell 1000 Index.

Taking the LLaMA2 model as an example, we report in Table IA8 the performance and characteristics of portfolios that exclusively trade Russell-indexed stocks, in comparison to those trading in the broader stock universe. An analysis of the data reveals notable differences. Portfolios limited to Russell 1000 stocks exhibit significantly larger capital sizes and higher liquidity.

Furthermore, in terms of the number of stocks traded, the disparity between the long and short sides of the Russell-only portfolios is minimal. This observation suggests that our original strategy provided a relatively balanced selection of the most influential and largest-cap stocks for both long and short positions. By focusing on Russell 1000 stocks, we not only adhere to a more targeted approach in our trading strategy but also gain valuable insights into the impact of market capitalization and liquidity on portfolio performance and characteristics.

Figure IA2: Prediction Analysis: Monthly Returns



Note: This figure plots the average monthly returns of portfolios based on sentiment analysis. The top panel shows the average daily returns in basis point for each month of portfolios based on sentiment analysis, while the bottom panel shows the average number of news each month.

A.3.3 Trading Timeliness and Portfolio Performance

In this section, we examine the impact of intraday trade timing on the performance of our portfolios. While constructing our portfolios, we utilized opening prices for training, formulating strategies based on the prediction of open-to-open (O2O) returns, and ideally executing trades at market open. However, practical constraints often preclude immediate execution at the opening, introducing a potential deviation from anticipated performance. To address this deviation, our methodology still utilizes predictions based on O2O returns but incorporates settlement using both the volume-weighted average price (VWAP) and close-to-close (C2C) returns. VWAP accounts for the reality of executing trades at different times during the trading day, while the C2C price represents the latest possible execution within a day.

Table IA9 provides the results of the comparative analysis using three distinct trading timings. The data from both panels consistently show that the O2O Portfolio stands out in terms of efficacy, yielding returns of 0.45 (0.57) and Sharpe Ratios of 4.16 (5.31) for sentiment analysis and return prediction strategies, respectively. This notable performance underscores the critical role of timely trade execution, which appears to be in perfect harmony with our predictive models. Conversely, the portfolios based on C2C and VWAP trading, while still profitable, exhibit a decline in both

Table IA8: Comparison of Trading in Russell 1000 Stocks

	All Stocks						Russell Only					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.37	-0.09	0.45	0.19	0.08	0.11	0.23	0.09	0.14	0.19	0.08	0.11
Std	0.20	0.23	0.11	0.19	0.22	0.11	0.21	0.24	0.10	0.19	0.21	0.11
SR	1.81	-0.38	4.16	1.03	0.38	0.98	1.12	0.38	1.39	1.02	0.37	1
Avg Traded	85	85	170	85	85	170	35	35	70	35	35	70
Median Traded	82	82	164	82	82	164	36	36	72	36	36	72
Avg Liquidity	3.03	4.15	3.59	3.03	4.15	3.59	5.44	8.93	7.18	5.44	8.93	7.18
Median Liquidity	0.79	0.73	0.74	0.79	0.73	0.74	2.29	3.26	2.62	2.29	3.26	2.62
Avg Cap	17.05	13.78	15.41	17.05	13.78	15.41	33.04	35.50	34.27	33.04	35.50	34.27
Median Cap	3.01	1.42	1.98	3.01	1.42	1.98	12.44	11.54	11.52	12.44	11.54	11.52

Note: This table provides a comparison of the performance and characteristics of portfolios trading in all stocks versus those trading exclusively in Russell 1000 stocks. The data is presented for both equal-weighted (EW) and value-weighted (VW) portfolios. Market capitalization is expressed in billions of USD, while liquidity is measured by the total number of shares sold per day and is presented in millions of shares.

returns and Sharpe Ratios. This trend can be attributed to the potential costs arising from delayed trading, indicating a deviation from the ideal timing prescribed by our models. However, the fact that these strategies remain profitable despite the less-than-optimal timing showcases the resilience and adaptability of our trading methodologies.

Table IA9: Performance Comparison of Portfolios Based on Trade Execution Timings

Panel A: Performance of Sentiment Analysis-Based Portfolios

Ret Type	O2O Portfolio			C2C Portfolio			VWAP Portfolio		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.35	-0.10	0.45	0.21	0.03	0.18	0.17	-0.02	0.18
Std	0.20	0.23	0.11	0.20	0.22	0.09	0.17	0.21	0.11
SR	1.75	-0.43	4.16	1.04	0.12	2.06	0.98	-0.08	1.66

Panel B: Performance of Return Prediction-Based Portfolios

Ret Type	O2O Portfolio			C2C Portfolio			VWAP Portfolio		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.46	-0.11	0.57	0.21	0.03	0.18	0.19	-0.06	0.25
Std	0.21	0.22	0.11	0.20	0.22	0.09	0.18	0.19	0.08
SR	2.22	-0.50	5.31	1.03	0.16	1.91	1.07	-0.32	2.95

Note: This table compares the performance of LLaMA2 portfolios based on different trade execution timings and strategies: open-to-open (O2O), close-to-close (C2C), and volume-weighted average price (VWAP). Panel A evaluates portfolios built on sentiment analysis, while Panel B focuses on return prediction strategies.

A.3.4 Other Complementary Tables

Table IA10: International News Summary Statistics

	Raw Articles			Articles Tagged with Single Stock			Articles With Returns	After Filtering Short Articles	After Filtering Similarity > 0.8
	RTRS	3PTY	Total	RTRS	3PTY	Total			
UK	707,288	1,050,467	1,757,755	196,573	773,266	969,839	906,705	901,838	571,285
Australia	261,020	1,203,784	1,464,804	100,444	1,113,347	1,213,791	388,585	382,114	249,190
Canada	255,933	473,686	729,619	126,281	431,401	557,682	481,891	478,205	350,549
China (HK)	3,537,487	7,287,688	10,825,175	1,140,542	5,558,763	6,699,305	2,086,045	305,335	182,363
Japan	3,259,103	38,860	3,297,963	1,210,077	16,850	1,226,927	405,341	399,185	310,244
Germany	2,423,671	1,751,231	4,174,902	480,264	880,650	1,360,914	238,577	229,265	178,039
Italy	1,022,204	337,322	1,359,526	194,650	227,599	422,249	173,250	168,410	130,168
France	2,422,338	1,587,490	4,009,828	298,886	670,469	969,355	174,917	174,784	153,779
Sweden	288,395	189,424	477,819	96,039	124,862	220,901	126,211	126,168	115,195
Denmark	261,146	124,209	385,355	93,596	57,768	151,364	53,056	52,381	43,584
Spain	2,748,601	165,468	2,914,069	257,739	46,829	304,568	47,541	45,597	34,159
Finland	81	110,123	110,204	38	87,226	87,264	38,159	38,119	28,633
Portugal	747,069	39,086	786,155	124,017	13,638	137,655	11,265	11,212	6,158
Greece	85,915	14	85,929	19,156	6	19,162	10,093	10,082	7,710
Netherlands	194	183,668	183,862	53	66,669	66,722	4,313	4,312	3,751

Note: In this table, we report the remaining sample size after each filter applied on the news articles for international equity markets. Column "Raw Articles" presents the numbers of available articles separately from Thomson Reuters Real-time News Feed (RTRS) and Archive (3PTY). Columns under "Articles Tagged with Single Stock" presents the number of articles tagged with a single stock. Column "Articles with Available Returns" presents the number of remaining articles after matching returns data. Column "After Filtering Short Articles" reports the number of articles with at least 100 characters and at most 100,000 characters.

Table IA11: Specification of Tokenizers

	BOW/Word Embeddings	BERT	RoBERTa
US	en_core_web_sm	bert-large-uncased	roberta-large
UK	en_core_web_sm	bert-large-uncased	roberta-large
Australia	en_core_web_sm	bert-large-uncased	roberta-large
Canada	en_core_web_sm	bert-large-uncased	roberta-large
China (HK)	jieba	ckiplab/bert-base-chinese	xlm-roberta-large
Japan	nagisa	cl-tohoku/bert-base-japanese	xlm-roberta-large
Germany	de_core_news_sm	bert-base-german-cased	xlm-roberta-large
Italy	it_core_news_sm	dbmdz/bert-base-italian-cased	xlm-roberta-large
France	fr_core_news_sm	dbmdz/bert-base-french-europeana-cased	xlm-roberta-large
Sweden	sv_core_news_sm	KB/bert-base-swedish-cased	xlm-roberta-large
Denmark	da_core_news_sm	Maltehb/danish-bert-botxo	xlm-roberta-large
Spain	es_core_news_sm	dccuchile/bert-base-spanish-wwm-uncased	xlm-roberta-large
Finland	sv_core_news_sm	TurkuNLP/bert-base-finnish-cased-v1	xlm-roberta-large
Portugal	pt_core_news_sm	neuralmind/bert-base-portuguese-cased	xlm-roberta-large
Greece	el_core_news_sm	nlpaueb/bert-base-greek-uncased-v1	xlm-roberta-large
Netherlands	nl_core_news_sm	wietsedv/bert-base-dutch-cased	xlm-roberta-large

Note: This table reports the specification of the pre-trained model for each country. Column "BOW/Word Embeddings" reports the specific tokenizer in spaCy that we use. Column "BERT" and "RoBERTa" reports pre-trained models for BERT and RoBERTa from Hugging Face.

Table IA12: Context, Words, Past Returns Comparison Conditional on News

	Stocks with news						ChatGPT					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.05	0.32	-0.26	0.11	0.25	-0.14	0.39	-0.14	0.53	0.21	0.05	0.16
Std	0.23	0.26	0.17	0.23	0.27	0.22	0.20	0.22	0.11	0.19	0.22	0.11
SR	0.24	1.22	-1.53	0.48	0.92	-0.63	1.88	-0.65	4.96	1.11	0.24	1.43
	LLaMA2						LLaMA					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.38	-0.10	0.48	0.19	0.07	0.12	0.36	-0.06	0.42	0.20	0.08	0.11
Std	0.21	0.23	0.12	0.20	0.22	0.12	0.21	0.22	0.11	0.20	0.22	0.12
SR	1.82	-0.46	4.17	0.96	0.33	0.98	1.74	-0.28	3.74	1.00	0.39	0.96
	RoBERTa						SESTM					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.36	-0.08	0.45	0.21	0.09	0.12	0.37	-0.04	0.41	0.28	0.06	0.22
Std	0.21	0.22	0.11	0.19	0.22	0.12	0.25	0.22	0.15	0.24	0.22	0.17
SR	1.75	-0.38	4.00	1.10	0.41	1.03	1.51	-0.16	2.67	1.17	0.27	1.31
	Word2vec						LMMD					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.32	-0.03	0.35	0.21	0.09	0.11	0.25	-0.01	0.26	0.16	0.11	0.05
Std	0.21	0.22	0.10	0.19	0.21	0.10	0.20	0.22	0.10	0.18	0.21	0.11
SR	1.55	-0.13	3.46	1.07	0.46	1.09	1.26	-0.03	2.64	0.87	0.51	0.49

Note: This table presents the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted based on sentiment scores and their respective long (L) and short (S) positions. The first row's left panel shows the performance of entire market portfolios sorted using past 1-day close-to-close returns. The rest of the panels present the performance of portfolios constructed from sentiment scores derived from LLMs and word-based models. For stocks associated with recent news, portfolio signals are generated using a combination of sentiment scores and probability-adjusted past 1-day close-to-close returns (calculated via logistic probability). In cases where no news is tagged to a stock, the signal is simply the probability-adjusted returns. The models employed include ChatGPT, LLaMA2, LLaMA, RoBERTa, Word2vec, SESTM, and LMMD.

Table IA13: Context, Words, Past Returns Comparison in Entire Market

	Entire Market						ChatGPT					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.28	0.04	0.24	0.22	0.11	0.11	0.33	0.04	0.28	0.20	0.10	0.09
Std	0.22	0.18	0.12	0.25	0.21	0.16	0.21	0.18	0.11	0.19	0.20	0.08
SR	1.29	0.24	1.91	0.88	0.51	0.72	1.55	0.24	2.62	1.01	0.52	1.17
	LLaMA2						LLaMA					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	0.05	0.28	0.19	0.11	0.08	0.32	0.05	0.27	0.20	0.11	0.09
Std	0.21	0.18	0.11	0.20	0.20	0.08	0.21	0.18	0.11	0.20	0.20	0.08
SR	1.55	0.25	2.58	0.98	0.54	1.01	1.52	0.27	2.52	1.02	0.55	1.20
	RoBERTa						BERT					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	0.05	0.28	0.19	0.12	0.07	0.32	0.05	0.27	0.19	0.10	0.09
Std	0.21	0.18	0.11	0.19	0.20	0.08	0.21	0.18	0.11	0.20	0.20	0.08
SR	1.54	0.26	2.55	0.99	0.60	0.90	1.53	0.27	2.50	0.99	0.52	1.10
	SESTM						LMMD					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.28	0.07	0.21	0.23	0.13	0.09	0.28	0.09	0.19	0.22	0.15	0.08
Std	0.21	0.19	0.11	0.24	0.19	0.12	0.21	0.19	0.10	0.24	0.19	0.11
SR	1.31	0.38	1.89	0.94	0.70	0.77	1.30	0.48	1.81	0.93	0.79	0.68

Note: This table presents the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted based on sentiment scores and their respective long (L) and short (S) positions. The first row's left panel shows the performance of entire market portfolios sorted using past 1-day close-to-close returns. The rest of the panels present the performance of portfolios constructed from sentiment scores derived from LLMs and word-based models. Portfolio signals are generated using a combination of sentiment scores and probability-adjusted past 1-day close-to-close returns (calculated via logistic probability). The models employed include ChatGPT, LLaMA2, LLaMA, RoBERTa, Word2vec, SESTM, and LMMD.

Table IA14: Impact of negation word ratio

	SESTM				Word2Vec			
	LLaMA2	LLaMA	ROBERTa	BERT	LLaMA2	LLaMA	ROBERTa	BERT
neg ratio	0.0134***	0.0123***	0.0074*	0.0071	0.0178***	0.0168***	0.0119***	0.0116***
	(0.0044)	(0.0045)	(0.0044)	(0.0043)	(0.0044)	(0.0044)	(0.0040)	(0.0039)
size	-0.0891***	-0.0795***	-0.0668***	-0.0888***	-0.0622***	-0.0526***	-0.0398***	-0.0619***
	(0.0169)	(0.0171)	(0.0168)	(0.0166)	(0.0166)	(0.0168)	(0.0152)	(0.0150)
BM	0.0045	0.0060	-0.0030	-0.0012	0.0019	0.0035	-0.0056	-0.0038
	(0.0075)	(0.0076)	(0.0075)	(0.0074)	(0.0074)	(0.0075)	(0.0068)	(0.0067)
liquidity	0.0421***	0.0424***	0.0355***	0.0272***	0.0631***	0.0635***	0.0566***	0.0483***
	(0.0105)	(0.0106)	(0.0104)	(0.0103)	(0.0104)	(0.0105)	(0.0095)	(0.0094)
IdioRisk	0.0173***	0.0114*	0.0204***	0.0086	0.0381***	0.0323***	0.0412***	0.0294***
	(0.0065)	(0.0065)	(0.0064)	(0.0064)	(0.0064)	(0.0065)	(0.0058)	(0.0058)
sic2D	-0.0393***	-0.0267*	-0.0328**	-0.0114	-0.0376**	-0.0250*	-0.0311**	-0.0097
	(0.0151)	(0.0152)	(0.0149)	(0.0147)	(0.0148)	(0.0150)	(0.0135)	(0.0133)
Constant	0.0259***	0.0207***	0.0191***	0.0222***	0.0186***	0.0134**	0.0118**	0.0149***
	(0.0053)	(0.0054)	(0.0053)	(0.0052)	(0.0052)	(0.0053)	(0.0048)	(0.0047)
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of obs	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769	1,552,769
Adj R-squared	0.0029	0.0035	0.0046	0.0047	0.0048	0.0036	0.0032	0.0030

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents regression results examining the impact of negation word ratio on the difference between LLM and word-based model performance with FF3 factors where we use LLM signals minus word-based signals multiplied by next period return as the dependent variable. The first 4 columns show the results with SESTM as a word-based model benchmark, and the rest 4 columns show the results with Word2Vec as a word-based model benchmark

Table IA15: Sharpe Ratios of Portfolios based on Sentiment Analysis

	LLaMA2		LLaMA		RoBERTa		BERT		Word2vec		SESTM		LMMD	
	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW
US	4.16	0.98	3.89	1.04	3.75	0.94	3.60	0.92	3.06	0.92	3.43	0.86	2.29	0.39
UK	2.79	1.22	2.74	1.29	1.44	0.71	1.42	0.59	1.38	0.72	2.05	0.73	0.80	0.32
Australia	-0.15	-0.22	-0.04	0.14	-0.24	0.15	-0.07	-0.04	-0.23	-0.29	-0.16	-0.11	0.38	0.03
Canada	1.96	1.16	2.30	1.12	1.74	0.99	2.14	0.84	1.26	0.39	0.62	0.33	0.69	0.33
China (HK)	0.77	0.54	0.52	0.46	0.93	0.75	1.00	0.69	0.71	0.46	1.03	0.76		
Japan	-0.63	-0.47	-0.32	-0.19	-0.32	-0.07	-0.42	-0.36	0.54	0.79	-0.54	-0.29		
Germany	2.08	0.95	1.68	0.65	1.45	0.78	1.21	0.70	0.70	0.67	0.92	0.70		
Italy	0.41	0.44	0.56	0.40	0.42	0.29	0.38	0.18	0.09	0.23	0.12	0.21		
France	0.97	0.65	0.70	0.04	0.96	0.13	0.91	0.42	0.65	0.49	1.06	0.14		
Sweden	0.48	0.51	0.49	0.21	0.69	0.88	0.68	0.31	0.76	0.17	0.01	0.53		
Denmark	0.34	0.30	-0.13	-0.16	0.30	0.31	0.25	0.12	0.37	0.25	-0.01	-0.16		
Spain	0.04	-0.02	0.05	-0.13	-0.13	-0.33	0.11	-0.02	0.11	-0.12	-0.26	-0.43		
Finland	0.99	0.85	0.53	0.30	0.92	0.52	0.88	0.50	0.13	-0.26	0.18	-0.06		
Portugal	1.16	1.12	1.55	1.54	1.50	1.49	0.21	0.19	1.23	1.21	3.88	3.86		
Greece	-0.48	-0.48	-0.57	-0.57	1.00	1.00	-0.05	-0.05	1.04	1.04	0.12	0.12		
Netherlands	-0.45	-0.45	-0.62	-0.62	-0.54	-0.54	0.48	0.48	0.75	0.75	1.14	1.14		
Mean	0.90	0.44	0.83	0.34	0.87	0.50	0.80	0.34	0.78	0.46	0.85	0.52	1.04	0.27
Mean (Excluding US)	0.68	0.41	0.63	0.30	0.67	0.47	0.61	0.30	0.63	0.43	0.68	0.50	0.62	0.23
Median (Excluding US)	0.48	0.51	0.52	0.21	0.92	0.52	0.48	0.31	0.70	0.46	0.18	0.21	0.69	0.32

Note: The table reports Sharpe ratios of long-short portfolios for international market portfolios. The portfolios are built on the basis of LLaMA, LLaMA2, RoBERTa, BERT, SESTM, Word2vec and LMMD model, respectively, using sentiment scores as sorting variables.

Table IA16: Performance Analysis of Trading Strategies with Transaction Cost

Panel A: Article-Based Portfolio Performance

	Turnover	Gross Return	Gross Sharpe Ratio	Net Return	Net Sharpe Ratio
0.10	10.23	2.77	3.02	0.08	0.08
0.20	20.80	5.52	3.39	0.06	0.04
0.30	31.65	8.23	3.62	-0.06	-0.03
0.40	42.73	10.90	3.79	-0.28	-0.10
0.50	54.03	13.53	3.92	-0.58	-0.17
0.60	65.51	16.13	4.02	-0.97	-0.24
0.70	77.17	18.69	4.11	-1.43	-0.31
0.80	89.03	21.22	4.18	-1.96	-0.39
0.90	101.14	23.74	4.24	-2.57	-0.46

Panel B: CAPM Residual Return Portfolio Performance

	Turnover	Gross Return	Gross Sharpe Ratio	Net Return	Net Sharpe Ratio
0.10	10.24	3.00	3.63	0.33	0.39
0.20	20.85	6.04	4.05	0.60	0.40
0.30	31.77	9.05	4.31	0.78	0.37
0.40	42.94	12.03	4.49	0.87	0.32
0.50	54.33	14.97	4.63	0.86	0.27
0.60	65.93	17.89	4.74	0.79	0.21
0.70	77.72	20.79	4.82	0.65	0.15
0.80	89.71	23.67	4.89	0.44	0.09
0.90	101.97	26.54	4.94	0.17	0.03

Panel C: Alert-Based Portfolio Performance

	Turnover	Gross Return	Gross Sharpe Ratio	Net Return	Net Sharpe Ratio
0.10	10.10	3.74	4.17	1.26	1.41
0.20	20.35	7.48	4.37	2.48	1.45
0.30	30.75	11.18	4.46	3.63	1.45
0.40	41.27	14.85	4.53	4.73	1.44
0.50	51.91	18.49	4.58	5.77	1.43
0.60	62.65	22.10	4.63	6.75	1.41
0.70	73.49	25.69	4.67	7.69	1.40
0.80	84.47	29.24	4.69	8.58	1.38
0.90	95.64	32.78	4.72	9.39	1.35

Note: This table presents the performance of trading strategies that incorporate transaction costs, trading the top and bottom 10% of stocks ranked by sentiment score across three distinct scenarios. Panel A evaluates the performance of article-based portfolios, Panel B delves into portfolios using CAPM residual returns, and Panel C assesses the performance of portfolios based on alerts.