Separating Sentiment from Fundamentals Using Large Language Model: How Does Sentiment Speak to the Crowd

Zhaohui Chen, Athanasios A. Pantelous, Ruohua Tang, Yuxin Xie*

September 4, 2025

Abstract

We fine-tune a large language model to extract article-level sentiment that is explicitly separated from fundamentals, and we show that it outperforms GPT variants and FinBERT. In firm-level analyses of financial media commentary, we find that positive (negative) sentiment predicts higher (lower) cumulative abnormal returns (CARs) and higher abnormal trading volume. These return effects subsequently reverse, consistent with noise trader theory (De Long et al., 1990). In contrast, sentiment measured using word-based methods is less informative: it does not consistently predict CARs, abnormal trading volume, or price reversals. A simple trading strategy exploiting our sentiment measure consistently outperforms strategies based on word-based proxies. Taken together, the results demonstrate that fine-tuned LLMs can isolate economically meaningful sentiment at the article level and provide sharper predictions for asset prices.

Keywords: Textual Sentiment, Large Language Models, Noise Traders, Narrative Economics, ChatGPT

JEL: C45, G10, G41

*McIntire School of Commerce, University of Virginia, USA (zc8j@virginia.edu), Department of Econometrics and Business Statistics, Monash Business School, Monash University, Australia (athanasios.pantelous@monash.edu, ruohua.tang@monash.edu), School of Finance, Southwestern University of Finance and Economics, China (yuxinxie@swufe.edu.cn). We are grateful to Chunrong Ai, Tony Cookson, Diego García, Liyuan Liu, Gordon Phillips, Rui Shen, Stavros Stavroglou, and George Baihan Wang for their valuable comments and suggestions. We also thank seminar and conference participants at Southwestern University of Finance and Economics, the Boyue Academic Forum at CUHK-Shenzhen, the 2025 Shenzhen Postdoctoral Academic Annual Conference, and the Quantitative Finance and Risk Analysis Symposium (QFRA 2025) for their helpful feedback. All remaining errors are our own. The authors declare that they have no relevant financial interests or other potential conflicts of interest.

1 Introduction

Economists (e.g., De Long et al., 1990) have long posited that investor sentiment, i.e., non-fundamental beliefs, can move market prices, which Keynes (1936) called "animal spirits". However, sentiment is hard to measure. For example, Baker & Wurgler (2006)'s sentiment index combines size, profitability, volatility, NYSE turnover, IPO average first-day returns, and other measures that may capture fundamentals as much as sentiment. Textual analysis literature, pioneered by Tetlock (2007), has documented that textual data can move stock prices.[1] A typical study in this literature counts the positive and negative words or sentences of a publication to quantify sentiment. Because positive or negative words may reflect either the author's sentiment or the underlying economic fundamentals, it is unclear whether the documented price movements are driven by investor sentiment or fundamentals. One way to differentiate sentiment from fundamentals is to look at reversals after initial price movements when sentiment fades. If price reversals cancel the initial movements, it supports the hypothesis that sentiment causes price movements. However, at the firm level, there is little evidence of reversals (e.g., Chen et al., 2014), but we do not know whether the lack of evidence is due to noisy sentiment measurement or to the failure of the sentiment hypothesis.

In this paper, we seek to cleanly separate sentiment from fundamentals and then examine both the immediate price response and any subsequent reversal. First, we define sentiment as the authors' attitudes, that is, their opinions and tones that are independent of the facts. Secondly, we define sentiment at the article level, rather than at the word or phrase level in the literature, to capture the subtle sentiment expressed "between the lines". The following example illustrates the pitfalls of word-count approach: "Apple achieved breakthrough innovations, successful market expansion, and improved operational efficiency, though I remain bearish on the stock; meanwhile, Microsoft faces significant challenges, declining margins, and mounting regulatory pressures, yet I find it attractively undervalued." A word-based method would classify Apple as positive (three

[1] A large body of research has documented that data such as company filings (Loughran & McDonald, 2011; Jegadeesh & Wu, 2013; Cohen et al., 2020), media coverage (Huberman & Regev, 2001; Chan, 2003; Dougal et al., 2012; Baker et al., 2016; Goldman et al., 2024), and social media (Antweiler & Frank, 2004; Jegadeesh & Wu, 2013; Chen et al., 2014; Cookson & Niessner, 2020; Chen & Hwang, 2022; Cookson et al., 2024), are informative. See Tetlock (2014) and Gentzkow et al. (2019) for comprehensive reviews.

favorable descriptors versus one negative opinion) and Microsoft as negative (three unfavorable descriptors versus one positive view), thus inverting the actual attitudes of the author. Finally, to analyze abnormal returns of assets associated with sentiment, we focus exclusively on sentiment directed at specific assets, excluding sentiment related to industries or the broader market.

Implementing our sentiment definition systematically and consistently at scale is nontrivial. While general-purpose LLMs (e.g., ChatGPT) offer a baseline, they exhibit inconsistent prompt adherence and variable task performance (e.g., OpenAI, 2023). To address this, we develop Sentiment-GPT, a domain-specific model fine-tuned for financial sentiment classification. We begin with a rigorously annotated training corpus that encodes our definition of sentiment. We use the Eastmoney.com dataset because it offers two properties relevant to sentiment classification: (i) linguistic diversity across millions of retail-investor discussions, yielding rich lexical and stylistic variation that improves generalization; and (ii) scale sufficient to train high-capacity models and support low-variance, stratified evaluation. These properties enable precise annotation and reliable downstream classification. We then fine-tune the Qwen2.5-72B base model on this corpus to obtain our final sentiment-classification system.

Our fine-tuned model outperforms the GPT variants based on four common performance measures in machine learning, *Precision*, *Recall*, F_1 *Score*, and *Accuracy*. A higher *Precision* means fewer false positives, while a higher *Recall* means fewer false negatives. A higher F_1 *Score* indicates that a model is performing well, balancing both *Precision* and *Recall*. Finally, a higher *Accuracy* indicates that the machine learning classification model is, on average, making more correct classifications.

The overall *Accuracy* of Sentiment-GPT in classifying sentiment is 90.67%, compared to 72.83%–80.17% for the GPT variants. FinBERT performs substantially worse, with an *Accuracy* of only 39.67%. The F_1 *Scores* of Sentiment-GPT for neutral, positive, and negative sentiments are 85.11%, 91.08%, and 95.35%, respectively. By comparison, the highest F_1 *Scores* achieved by the GPT variants are 65.32% (GPT-4.1), 86.94% (GPT-5), and 84.93% (GPT-4.1), respectively. FinBERT's F_1 *Scores* are 49.12%, 36.93%, and 14.85%, respectively.

A natural concern in LLM applications is look-ahead bias: if the training corpus contains information about firms we subsequently evaluate, our sentiment measure could

appear overly predictive due to prior model exposure. In principle, however, this concern does not apply to our approach because sentiment classification relies on the textual content and expressed opinions within each message, independent of any background knowledge about specific firms. Moreover, our sentiment definition is newly introduced and absent from pre-training sources, ensuring that performance gains stem from learning our classification framework rather than firm-specific information. Finally, our empirical tests confirm that look-ahead bias does not affect our sentiment classification performance.

Sentiment-GPT's superior performance extends beyond its Chinese training domain, as demonstrated by English-language evaluations. In particular, the model maintains comparable accuracy when evaluated on English translations of our benchmark, performs consistently across diverse text genres, from informal social media posts to formal financial journalism, and remains robust to variation in document length.

Equipped with Sentiment-GPT, we study whether firm-level sentiment from media moves stock prices as predicted by the theory: sentiment initially moves the prices as investors overreact, and then the prices revert to fundamentals after the sentiment wears off. To maximize the presence of sentiment, we focus on commentary articles about the 100 largest publicly traded U.S. firms published in the Wall Street Journal (WSJ), the New York Times (NYT), and the Financial Times (FT). All commentary articles are downloaded from *Factiva*. These platforms are widely viewed as shaping investor perceptions of blue-chip stocks (e.g., see Tetlock, 2007; Garcia, 2013). We include firms as long as they have entered the top 100 list at least once during the sample period. In total, we collected 55,275 articles covering 238 firms from January 1992 to December 2024. We then match the sentiment of these articles with the corresponding firm characteristics from CRSP.

Our empirical findings support the noise trader theory of De Long et al. (1990). We show that both positive and negative sentiment in media coverage induce abnormal returns, consistent with noise traders creating excess buying or selling pressure. Specifically, stocks discussed with positive (negative) sentiment exhibit significantly higher (lower) cumulative abnormal returns (CARs) in the short term, as well as larger abnormal trading volume. However, these CARs disappear about a month after publication because the prices revert to fundamentals. The reversals continue for at least another month. We

benchmark these results against conventional word-based sentiment measures derived from the Loughran & McDonald (2011) (LM) dictionary. Although LM measures partially predict CARs, their explanatory power becomes insignificant when our sentiment measures are included. In addition, LM measures do not consistently predict either price reversals or abnormal trading volume.

To further assess the economic significance of the relationship between textual sentiment and stock prices, we evaluate the performance of sentiment-based trading strategies using our sentiment measures. Covering the same period as our regression analysis, the results show that a simple long-short strategy based on our sentiment measures generates an excess return of 2.46% per month, an average annual return of 24.40% with a Sharpe ratio of 0.56. In comparison, an identical strategy using the word-based approach yields a lower average annual return of 16.10% and a Sharpe ratio of 0.38. However, when a 10 basis point round-trip transaction cost is applied, the performance of the word-based strategy drops sharply, producing a negative average return of –6.00% and a Sharpe ratio of –0.141. In contrast, the strategy based on our sentiment measures remains robust, maintaining an average annual return of 13.90% with a Sharpe ratio of 0.32.

1.1 Related Literature

As mentioned above, many have attempted to extract sentiment from text. This literature began with word-based approaches (e.g., Tetlock, 2007; Tetlock et al., 2008; Loughran & McDonald, 2011; Garcia, 2013; Chen et al., 2014; Garcia et al., 2023; Bin Hasan et al., 2024; Cookson et al., 2024, 2025). Machine-learning classification (e.g., Das & Chen, 2007; Huang et al., 2014; Azimi & Agrawal, 2021) and LLMs (e.g., Huang et al., 2023; Lopez-Lira & Tang, 2023) have subsequently emerged as stronger alternatives. We believe studies that rely on counting positive and negative words or phrases are limited in their ability to separate sentiment from fundamentals. Approaches that use LLMs without fine-tuning risk relying on the LLMs' definitions of sentiment and fundamentals, which, as we show, may not align with our needs. As a result, even though many documented sentiment measures can move firms' stock prices, there is no evidence that prices revert to fundamentals. Tetlock (2007) and Garcia (2013) document market index reversals, but because these reversals are not abnormal returns by definition, they do not confirm

mispricing.[2]

We thus contribute to the literature in two respects. First, we show sentiment's full price impact, initial mispricing and subsequent reversal, at the firm level, for the first time in the literature. Second, the new results hinge on our innovative methodology: annotating a large training corpus that clearly incorporates our definition of sentiment, communicating that definition to an LLM via fine-tuning, and then using the fine-tuned model to extract media sentiment.

In principle, our methodology for fine-tuning LLMs can be applied to improve the quantification of variables of interest from texts in various economic settings.[3] The importance of this contribution is highlighted by the increasing popularity of powerful general-purpose LLMs such as ChatGPT in textual analysis.[4] Although general-purpose LLMs offer powerful tools to scholars, a growing concern is their inaccuracy. For example, many LLMs generate wrong classifications for no reason, i.e. the "hallucination". Our methodology demonstrates that fine-tuning a large language model (LLM) for a specific task can outperform general-purpose models such as ChatGPT, significantly enhancing domain-specific capabilities while limiting hallucinated outputs.

The remainder of the paper is organized as follows. Section 2 introduces our sentiment classification framework. Section 3 validates the sentiment classification. In Section 4, we extract sentiment from financial commentary and study its market impacts. Section 5 concludes.

2 Methodology

This section presents our novel framework for textual sentiment extraction. We first define sentiment as the author's attitude toward a specific asset, explaining our focus

[2] Goetzmann et al. (2024) uses a large language model to recover market-wide investor sentiment and shows that subjective crash beliefs are associated with negative sentiment. Because the measure is aggregate, any associated market responses do not constitute abnormal returns.

[3] For example, Soo (2018) measures the sentiment of housing news and shows a positive association between a sentiment index of the media and future house prices. Cohen et al. (2020) quantifies changes in the language and construction of financial reports and shows how these changes impact the firms' future returns and operations. Engle et al. (2020) quantifies the intensity of climate news and then constructs portfolios based on the intensity to hedge innovations in climate news.

[4] Several recent studies demonstrate this growing trend. For example, Lopez-Lira & Tang (2023) uses ChatGPT to evaluate the impact of news headlines on stock prices. Bybee (2023) explores the potential of LLMs to generate economic expectations. Kim et al. (2024) uses ChatGPT to summarize complex corporate disclosures such as Management Discussion and Analysis (MD&A) and earnings conference calls.

on personal opinions and emotional tone. We then detail the human annotation process employed to create our training data, outlining the annotation guidelines and discussing the challenges inherent in classifying a large number of posts. We also address the issue of textual unreadability, a factor often neglected in prior research but crucial for accurate sentiment extraction. Finally, we describe the construction of our annotated corpus, Andromeda, encompassing a large volume of diverse textual data, and the development of our fine-tuned LLM, Financial-Sentiment-GPT, trained using Andromeda.

2.1 Definition

We define textual sentiment as the explicit expression of personal attitudes and emotional tone conveyed by the author regarding a specific asset. Our definition explicitly excludes purely factual information, objective financial metrics, and neutral descriptions of news or events, focusing solely on content that reflects the author's subjective assessment or emotional engagement with financial objects.

We prioritize personal attitudes and emotional tone for two key reasons. First, in the finance literature, sentiment is typically referred to as a subjective assessment or emotional disposition that can cause asset prices to deviate from their fundamental values (Zhou, 2018). Textual content that primarily conveys fundamentals, such as news reports, corporate disclosures, or descriptive accounts, does not align with this conceptualization of sentiment. Rather, "sentiment" should closely reflect investors' "animal spirits", characterized by spontaneous optimism or pessimism. In this regard, personal attitudes and emotions closely align with this concept—they are of particular interest to individuals because they offer readily actionable narratives and evoke immediate emotional responses. This cognitive shortcut makes personal attitudes and emotions highly attention-grabbing and contagious, enabling their rapid dissemination and their potential to override rational decision-making (Shiller, 2017; Graeber et al., 2024).[5]

Second, including content beyond personal attitudes can introduce significant measurement errors. For example, qualitative fundamentals are inherently difficult to inter-

[5]Prioritizing personal attitudes does not diminish the influence of other content (e.g., fundamentals). Rather, personal attitudes—direct reflections of sentiment based on the author's information set—bypass the interpretive ambiguities associated with fundamentals, thereby implicitly incorporating their informational impact. Consequently, interpreting an author's personal attitudes through their writing becomes more direct: by focusing on personal attitudes, the impact of fundamentals is implicitly and more accurately incorporated into the sentiment extraction process.

pret objectively and often lack a clear ground truth (see Hong & Stein, 2007; Zhou, 2018). As a result, heterogeneous interpretations are much more common.[6] Conversely, tangible fundamentals, such as earnings or profit margins, can endogenously influence stock prices or trading volume, potentially overstating the effect of textual sentiment. Consequently, assigning a specific sentiment label to fundamentals is inherently flawed. This limitation becomes particularly problematic when conducting sentiment analysis on large volumes of text, where even small inaccuracies can have a substantial cumulative effect.

Our emphasis on asset-specific sentiment addresses the practical challenges of multi-asset discussions commonly found in online forum posts. Authors frequently discuss several assets within a single post, and their sentiment toward one asset may differ significantly from their sentiment toward another. Therefore, if the target of the sentiment is not explicitly identified, any attempt to assign a single sentiment to the entire post becomes inherently ambiguous and potentially misleading. By requiring explicit identification of the specific asset being evaluated, our approach ensures precision in sentiment attribution and enables more accurate measurement of investor attitudes toward individual securities.

2.2 Human Annotation

2.2.1 The Annotation Rules

We discuss our procedures of sentiment extraction step by step. These procedures are carefully designed to achieve a systematic and comprehensive approach to human annotation. Human annotation is the process of manually labeling data by human annotators based on one or more features. In our case, the main challenge is ensuring that each annotator can not only understand the definition but also consistently apply it to each post classification with high accuracy. This can be particularly difficult given the frequent occurrence of unpredictable exceptions and textual unreadability.

We focus on categorizing posts into three distinct sentiment categories: "positive", "neutral", and "negative". There are several compelling reasons for not differentiating between multiple grades of the same sentiment tag.

[6] As evidence, despite decades of academic study, much economic reasoning related to news or fundamentals remains controversial. It is therefore unlikely that individuals, with only a cursory reading of texts, can derive insightful and uniform economic judgments.

First, increasing the number of grades introduces more subjective judgments, making it exceptionally challenging to ensure that all annotators can maintain consistent grading standards across all posts. For example, to what extent should a post be regarded as "strong negative" compared to "negative"? Different annotators tend to have their own cognitive thresholds, which are difficult to explain or detect. This hidden inconsistency in annotation could severely impair the quality of the resulted corpus.

Second, even if a corpus offers consistent grades, its application with LLMs is still problematic as pre-training or fine-tuning LLMs with such a corpus requires detailed grading criteria to differentiate between neighboring grades. This often relies on implementing complex reasoning structures like chain-of-thought (CoT). CoT involves explicitly outlining the reasoning steps an annotator would take to arrive at a specific grade, thereby transforming additional subjective judgments into a series of logical steps. Although CoT can provide a framework for consistency and transparency, its implementation significantly complicates the annotation process. Annotators not only need to classify the sentiment but also need to offer detailed reasoning, a process that generates additional complexity and inconsistencies.

Third, in sentiment analysis, the primary objective is often to explore the qualitative effects of various sentiments on different entities, such as individuals, organizations, or broader market trends. This qualitative focus seeks to determine how positive, negative, or neutral sentiments influence decisions, behaviors, and ultimately, economic outcomes. Therefore, although more detailed sentiment grades can provide more nuanced information, given their potential measurement errors, increasing the number of grades seems to add little value in enhancing theoretical understanding or improving statistical analysis.

Given that each annotator has to evaluate a large number of posts in a relatively short time, minimizing their cognitive burden is our top priority. To best simplify the annotators' decision-making process, the annotation starts by decomposing the entire set of posts into five mutually exclusive subsets:

- Subset 1: Texts not reflecting their own attitudes, such as factual descriptions or inquiries.
- Subset 2: Texts reflecting their own attitudes, but not related to a specific asset.
- Subset 3: Texts exclusively containing the author's opinions towards a specific asset.

- Subset 4: Texts expressing both the author's opinions and emotional tones towards a specific asset.
- Subset 5: Texts exclusively reflecting the author's emotional tones towards a specific asset.

For posts in Subsets 1 and 2, which fall outside the scope of our definition, we assign a neutral sentiment. Sentiment extraction for posts in Subsets 3 and 5 is grounded in either opinions or emotional tones expressed by the author. In the case of Subset 4, where posts blend opinions with emotional tones, we prioritize the author's opinions during sentiment evaluation. We further stipulate that if the sentiment based on the author's opinions conflicts with that based on the author's emotional tones, we focus solely on the opinions. Therefore, when a post contains both personal attitudes and emotional tones, we extract the sentiment based on the author's emotional tones only if they hold neutral opinions.

To further assist annotators, we use a flowchart, as shown in Figure 1, to direct each subset into a specific workflow. Within this pipeline, each post will be clearly identified and assessed based on the rules specific to its subset. To further clarify how a post is assigned a sentiment tag, we walk through the process from the annotator's perspective below.

1. The process begins by evaluating whether the post is understandable. If a post is hard to understand, the annotator should mark it as "complicated" and continue to evaluate.
2. For those posts that are understandable, the next step is to determine if the post reflects the author's own attitudes, which may include opinions, emotional tones, or both. If the post does not reflect the author's attitudes, it is categorized as "neutral" and classified into Subset 1. The evaluation ends.
3. If the post does reflect the author's attitudes, the annotator should continue to check whether it relates to any specific asset. Posts that do not relate to a specific asset are classified into Subset 2 and assigned "neutral". The evaluation ends.
4. For posts that are related to a specific asset and contain only opinions, the post is classified into Subset 3 and be assessed accordingly. The evaluation ends.

5. For posts that are related to a specific asset and contain both opinions and emotional tones, the post is classified into Subset 4 and assessed solely based on opinions. The evaluation ends.

6. Posts that contain only emotional tones are classified into Subset 5 and assessed based on the emotional tones. The evaluation ends.

[Figure 1 about here.]

After categorizing all posts into their corresponding subsets, annotators assign a sentiment to each post based on their individual interpretations. Naturally, this process entails a significant degree of subjective judgment. Our human annotation team consists of five members—one professor and four Ph.D. students—each with an extensive background in finance. While we could maintain a much larger group of annotators, early practices have shown that the expertise and relatively small size of the group are crucial for strict adherence to the established annotation guidelines.

To achieve the best accuracy, all posts are simultaneously assessed by all annotators. We attach the company name to each post so that all annotators know which company the writer is referring to, simulating the reality where readers of the post would also know the company being discussed. Before commencing the formal annotation process, we conducted ten training rounds, each containing 500 posts, to familiarize the annotators with the definition. We discuss all posts for which at least one annotator assigns a different sentiment than the others. After these ten rounds of training sessions, the average disagreement rate among the group decreased from 55% to 20%. No significant further drop in the average disagreement rate was observed after additional training rounds beyond the initial ten. For posts where agreement cannot be reached even after group discussion, we categorize these posts as "neutral" and mark them as "complicated". The key motivation behind this treatment is that if a post evokes persistent controversy among the five experienced annotators, it is likely to do the same for others. Thus, our "complicated" posts come from two sources: either they are subjectively marked by the annotators at first sight or they are objectively sorted based on the annotators' disagreements.

2.2.2 Textual Unreadability

A clear definition of textual sentiment provides a theoretical guideline for sentiment extraction. The flowchart and training sessions are essential layers of quality control, aimed at ensuring each annotator can consistently apply the definition. Are all these efforts enough to help annotators assign consistent sentiments to all posts? Of course not. Divergences may still arise. These remaining divergences are primarily due to the inherent unreadability of the text, which hinders the precise interpretation of the writings provided.

Given the vast volume of texts, unreadability is not negligible, as it poses challenges for both human annotators and LLMs during the sentiment extraction process. Current pre-training and fine-tuning datasets lack specific inclusion to these unreadable texts, especially in the Chinese community. Our annotation process, however, ensures that a large number of unreadable posts are promptly identified and labeled as "Complicated". This additional dimension grants us the opportunity to incorporate a solid coverage of textual unreadability by deliberately including 50% of complicated posts in the corpus. As a result, our fine-tuned LLM is expected to gain an extra boost in understanding posts with textual unreadability. Below, we summarize two typical causes of textual unreadability. For interested readers, we also provide a detailed collection of unreadable posts in Appendix D.2.

Unclear Expressions and Typos. Unclear expressions refer to phrases or sentences that are either confusingly written or clearly written but present a vague meaning. These might be due to complex jargon, grammatical errors, distorted sentence structures, implicit irony, or expressions that are not universally understood across different individuals. Additionally, there are also typos resulting from a careless and hasty typing, as well as typos resulting from auto-correct and text-predictive features provided by input software. The prevalence of unclear expressions and typos can significantly weaken the clarity of communication, creating barriers for annotators attempting to understand the intended meaning. When faced with posts containing unclear expressions or typos, we rely on annotators' own judgment and adaptability to extract potential sentiments. We posit that if five diligent annotators cannot converge on an identical classification for a given post, such controversy would likely extend to the investor community. As a result, posts that are debated among annotators are classified as "neutral", as their associated

unreadability is likely to limit their real-world impact on others.

Inconsistent multi-horizon sentiments. Another common form of unreadability arises when a post expresses both short-term and long-term viewpoints that are contradictory. This situation often occurs when an investor is tempted by short-term speculation that goes against their long-term expectations. For instance, an investor may have a bearish outlook on a particular stock but still decide to go long within a time window during which they believe the price will go up. Consequently, debates among annotators emerge due to the lack of decisive solution for distinguishing between short-term and long-term sentiment signals and weighing their relative importance in determining the overall sentiment of the post.

To address this issue, we recommend prioritizing short-term sentiments over long-term sentiments when they are in conflict. This compromise is based on two rationales. First, short-term sentiments are much more relevant for evoking immediate trading, as the majority of investors may find short-term speculative opportunities more attractive than long-term investments. Second, as it is counter-intuitive to think that today's sentiment will impact the market one month later, sentiment analysis is often conducted at a relatively higher frequency. Therefore, long-term perspectives on a stock tend to be much less predictive than short-term ones, making the latter a more suitable lens for capturing the dynamic relationship between sentiment and market variations. Finally, there is one exception: if the short-term sentiment is neutral while the long-term sentiment is not, we propose that the long-term sentiment should be used to represent the overall sentiment of the post.

2.3 The Corpus: Andromeda

It is essential to computationally perform sentiment extraction based on the definition. Fine-tuning LLMs for sentiment analysis has emerged as a highly effective method for achieving accurate and consistent classifications. The foundation lies in creating a large and high-quality corpus through human annotation, which is critical for fine-tuning LLMs to perform on par with the most experienced human annotators (e.g., see Devlin et al., 2018; Kaplan et al., 2020). In this section, we describe how we construct the applied corpus based on the human annotation.

An abundant and relevant collection of raw texts is fundamental for building a spe-

cialized and representative corpus. A large volume of raw texts not only offers a comprehensive source of linguistic and contextual patterns, but also enables the targeted corpus to be tailored, balanced and expanded over time to meet the evolving needs of model fine-tuning. The relevance of the raw texts is also essential to ensure that the developed corpus is domain-specific, effectively capturing the wide array of jargon, terminology, topics, and narrative styles encountered in financial communications.

We identify Eastmoney.com as the principal source for our textual content collection.[7] Eastmoney.com is a prominent financial services and information portal in China, acclaimed for its exhaustive coverage of financial news, stock market data, and investment insights. The website features specialized sub-forums, known as "GuBa", where investors engage in discussions about all listed Chinese stocks. These forums are meticulously structured, displaying historical posts across 80 posts per page, ensuring comprehensive archival access. As of the end of December 2022, Eastmoney.com hosted discussions on 5,612 stocks, accumulating a total of 252 million posts. The site's significant reach within the Chinese investment community, coupled with its consistent post formatting and extensive historical data, makes it an ideal source for our textual analysis.

The textual content for our study is sourced from Belief Analytics, a specialized lab focusing on the extraction and analysis of large volumes of textual data and the development of finely tuned large language models for specific sectors.[8]

Next, we introduce Andromeda, a high-quality corpus meticulously crafted for fine-tuning LLMs in extracting sentiment within the financial domain. This corpus integrates the essence of extensive and relevant textual data with rigorous human annotation standards. We maintain both a large size and linguistic diversity within the corpus to ensure it provides fruitful features for later fine-tuning. Andromeda includes 9,000 pairs of post-annotated sentiment.

In addition to size, linguistic diversity is another cornerstone that affects the corpus quality. To formally construct Andromeda, we proceed with an ex ante data grouping as follows. First, we categorize the entire set of posts into ten groups based on word count. This step is crucial because Transformer-based models often struggle to efficiently process longer contexts due to the inherent quadratic computational demands in both time and

[7] See Appendix D for detailed discussions of the textual content selection.
[8] For more information, please visit Belief Analytics at https://www.belief-analytics.com.

memory.[9] Second, within each word count group, we further divide the texts into those with and without special meanings, including buzzwords, financial terminology, emojis, and slang. These specific-domain texts often contain meanings vastly different from their everyday use. Finally, we evenly and randomly collected raw posts for human annotation from these processed subgroups to ensure that the resulting corpus best represents diversity in both post length and words used. In particular, we sample across each word count decile, posts with special and non-special meaning, common and complicated posts, as well as each sentiment category. As a result, we group the entire posts in Andromeda into 10×2×2×3=120 groups and keep the number of posts within each group as 75. In practice, since the labeled sentiment is unknown before annotation, we consistently collect posts from each group until the number of posts within each sentiment group meets the target.

2.4 A Fine-tuned Model: Financial-Sentiment-GPT

Based on Andromeda that contains 9,000 carefully classified posts, we fine-tune Qwen2.5-72B to incorporate the definition. This process led to the creation of our fine-tuned LLM—Financial-Sentiment-GPT (Sentiment-GPT). [10]

As a general strategy, we divide the posts in Andromeda into training and validation sets, with the training set comprising 80% of the posts and the validation set comprising 20%.[11] Furthermore, within each dataset, we maintain a balanced distribution across the 120 linguistic characteristics discussed in Section 2.2.2.

[9]For instance, Liu et al. (2023) note that the performance of LLMs substantially declines as input context lengthens, even for models designed to handle long contexts.

[10]The detailed implementation and environment setup are standard procedures, which can be found in OpenAI's official tutorial at: https://learn.microsoft.com/en-us/azure/ai-services/openai/tutorials/fine-tune?tabs=python-new%2Ccommand-line.

[11]The Pareto (80/20) principle is commonly used and proves effective for dataset splitting in machine learning applications.

3 Validating the Sentiment Classification

3.1 Performance Evaluation

3.1.1 The Benchmark: Libra

LLMs are increasingly utilized in financial studies for textual analysis (e.g., Kim et al., 2023; Lopez-Lira & Tang, 2023; Ke, 2024), yet the actual accuracy of these models remains largely unknown due to the absence of specialized benchmarks for these specific downstream tasks. Even GPT-4, for instance, may exhibit excellent performance in certain tasks while struggling in others, as proficiency in one area does not necessarily guarantee competence in another.[12] Consequently, a rigorous external benchmark is essential to ensure that LLM-based outputs are genuinely precise before applying them in real-world analysis.

In this section, we introduce Libra, a specially designed benchmark to verify the accuracy of sentiment extraction based on the definition. For any LLM-based application, an objective benchmark is vital to quantify the exact performance of LLMs, especially on domain-specific tasks. Libra serves two important functions. First, it provides an objective benchmark to examine whether the fine-tuned LLM has successfully gained domain-specific capabilities. For example, hallucinated outputs can be persistent and hard to detect since they are the model's false interpretations that seem real. Just as it is challenging to distinguish psychological hallucinations from reality, in the context of LLM applications, it is similarly difficult to detect hallucinated texts without a proper benchmark. Second, Libra enables a crucial feedback loop within the fine-tuning process, as it sets clear and consistent standards for model iteration and prompt optimization. This guides developers in purposefully enhancing performance and accuracy.

Following the same sampling rules used in constructing Andromeda, we select 600 posts to construct Libra.[13] As a domain-specific benchmark, the size of Libra is comparable to those benchmarks for general NLP tasks (e.g., Wang et al., 2018), this ensures our benchmark are diversified based on various linguistic characteristics and be reflective

[12]For example, GPT-4 excels in English language tests (score: 85.5%) but underperforms in advanced coding tasks, such as hard-level Leetcode problems (score: 6.7%). See OpenAI (2023) for a reference.

[13]Libra serves as the testing set to ensure that performance evaluation is independent of the fine-tuning process.

to model's true performance. The posts included in Libra are entirely independent of those in Andromeda. This is an important measure to avoid overstating the model's ability, as the benchmark data is not used in the model's fine-tuning session.[14] Specifically, we evenly sample the posts into 120 groups, with each group containing 5 posts. Additionally, Libra contains 50% complicated posts marked from the annotation process, which amounts to 300 out of the entire 600. The diversity of Libra benchmark enables it to evaluate the model's capability to extract positive, negative, and neutral sentiments from both ordinary and complicated posts.

3.1.2 Setup

In this section, we present the setup of performance evaluations for Sentiment-GPT, while also comparing it against the current leading LLMs: GPT-5, GPT-4.1, GPT-4o and GPT-4-turbo.[15] We set the "temperature" of generative LLMs to 0 and evaluate LLMs in a zero-shot setting using Libra. "Temperature" is a critical hyperparameter in generative large language models, governing the randomness of the outputs. Setting the temperature to zero means the model consistently selects the highest probability word, a condition maintained throughout this paper to ensure output consistency.

Considering cost-effectiveness, we use zero-shot settings to evaluate LLMs on their general capability for textual sentiment extraction, despite the potential marginal performance gains from few-shot evaluations.[16] We incorporate the definition directly into the prompt through natural language and include the detailed rules to be as clear as possible. The sentiment-defined prompt is constructed as follows:

Sentiment-defined Prompt

Definition

We define sentiment in financial text as the author's own clearly expressed opinion

[14]Several LLMs include part of the benchmark data in the pre-training or fine-tuning dataset to enhance their performance based on related benchmarks. See Zhou et al. (2023) for the definition and potential risks of data (benchmark) leakage.

[15]We use the *gpt-5-chat-latest* version for GPT-5, the *gpt-4.1-2025-04-14* version for GPT-4.1, the *gpt-4o-2024-11-20* version for GPT-4o, and the *gpt-4-turbo-2024-04-09* version for GPT-4-turbo.

[16]Given that computational costs are tied to text length, a few-shot evaluation dramatically increases the input length for each round of model inference. In large-scale sentiment analysis, where millions of posts are processed, the few-shot evaluation significantly amplifies inference costs. Therefore, due to the substantial cost increase and minimal performance benefits, we prefer zero-shot prompts over few-shot prompts during fine-tuning.

or emotional language directed at a single financial asset. Any text that falls outside this definition—such as vague opinions, ambiguous emotional language, or objective financial news—should be labeled as **Neutral**. When both personal opinion and emotional language are present, sentiment should be extracted based on the **personal opinion**. If the sentiment expressed by the author is ambiguous or likely to be interpreted differently by the general reader, it should also be labeled as **Neutral**.

Instruction
Below is a piece of financial text. Please choose **only one** of the following three options to represent the sentiment specifically toward **(Firm name here...)**: [Negative, Neutral, Positive]. Do **not** output any other content.

Financial Text
(Text here...)

Sentiment Answer

We evaluate model performance using Precision, Recall, F_1 Score, and Accuracy, which are standard evaluation metrics commonly used in computer science. In our sentiment classification task with three possible categories, namely -1 (negative), 0 (neutral), and 1 (positive), the model's performance is summarized using a 3×3 confusion matrix as shown in Table 1. In this matrix, rows represent the true sentiment labels, and columns represent the predicted sentiment labels. N_{ij} denote the number of instances where the true label is i and the predicted label is j.

[Table 1 about here.]

For each sentiment class $c \in \{-1, 0, 1\}$, we track three key quantities derived from the confusion matrix in Table 1. The number of *true positives* (TP_c) corresponds to the diagonal entry where the model correctly predicts class c, that is,

$$TP_c = N_{c,c}. \tag{1}$$

The number of *false negatives* (FN_c) captures how often true instances of class c are incorrectly classified as something else, given by the sum of the off-diagonal entries in

row c:

$$FN_c = \sum_{j \neq c} N_{c,j}. \tag{2}$$

The number of *false positives* (FP_c) measures how often the model incorrectly assigns class c to instances that belong to a different class, which corresponds to the sum of the off-diagonal entries in column c:

$$FP_c = \sum_{i \neq c} N_{i,c}. \tag{3}$$

Once these quantities are obtained, *precision* for class c, the proportion of correct predictions among all instances predicted as c is defined as

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{4}$$

while *recall*, refereed as the proportion of actual class c examples that were correctly identified, is computed as

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}. \tag{5}$$

The F_1 $Score$, the harmonic mean that balances the two, is

$$F_1\ Score_c = \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{6}$$

Finally, we calculate the overall accuracy, which is the proportion of all cases the model classifies correctly: add together the diagonal entries of the confusion matrix (the true-positive counts for the three sentiments) and divide this sum by the total number of evaluated instances.

$$\text{Accuracy} = \frac{N_{-1,-1} + N_{0,0} + N_{1,1}}{\sum_{i\in\{-1,0,1\}} \sum_{j\in\{-1,0,1\}} N_{i,j}} = \frac{TP_{-1} + TP_0 + TP_1}{\sum_{c\in\{-1,0,1\}} (TP_c + FN_c)}. \tag{7}$$

3.1.3 Evaluation

To test the capability of sentiment extraction using existing LLMs, we consider state-of-the-art generalized LLMs, including GPT-5, GPT-4.1, GPT-4o and GPT-4-turbo (e.g., Rathje et al., 2024), and FinBERT (e.g., Huang et al., 2023). Table 2 presents the zero-

shot model evaluation based on Libra.[17] In Panel A, we observe that Sentiment-GPT dominates in detecting neutral texts, with an F_1 *Score* of 85.11%, compared to 63.82% for GPT-5, 65.32% for GPT-4.1, 65.08% for GPT-4o, and 60.38% for GPT-4-turbo, indicating that Sentiment-GPT not only shows a higher performance in both *Precision* and *Recall*, but also achieves a better balance between *Precision* and *Recall* than other models. This indicates that GPT-5, GPT-4.1, GPT-4o, and GPT-4-turbo miss a significant portion of actual neutral texts, reflecting either a conservative decision boundary or an inability to detect neutral texts comprehensively. This suggests that GPT-5, GPT-4.1, GPT-4o, and GPT-4-turbo are not capable of understanding the definition, even when it is incorporated within the prompt. In contrast, by fine-tuning based on Andromeda, Sentiment-GPT is able to understand the definition. Our finding also provides insights for economic studies using LLMs, suggesting that LLMs may not fully understand natural language and that validation based on a corpus tailored to specialized research questions is necessary.

Besides, Sentiment-GPT outperforms other models in detecting both positive and negative texts, with an F_1 *Score* of 91.08% and 95.35%, respectively. Overall, Sentiment-GPT dominates in all tasks, achieving an *Accuracy* of 90.67%. Given that, in practice, neutral texts are more common, the actual *Accuracy* of models is likely more influenced by their performance in understanding the definition and detecting neutral texts.

FinBERT shows underperformance compared to larger LLMs. Specifically, its F_1 *Scores* for identifying positive and negative texts are 36.93% and 14.85%, respectively, which are roughly equivalent to the F_1 *Score* for random selection (33.33%). While FinBERT achieves a higher *Recall* (89.36%) than Sentiment-GPT, its precision drops significantly to 33.87%. It seems that FinBERT struggles with semantic comprehension across entire articles and frequently categorizes texts as "neutral" due to input-length constraints, limiting its analysis to sentence-level processing.

Next, we further test the model performance within the ordinary post and compli-

[17]Since FinBERT operates at the sentence level, we follow Huang et al. (2023) to classify individual sentences and determine the overall sentiment score of an article through aggregation. Specifically, the sentiment score is calculated as the proportion of positive sentences minus that of negative sentences. We define an article's FinBERT sentiment as positive when the score exceeds τ, negative when it falls below $-\tau$, and neutral otherwise. Results are presented with τ set to 0.25. In untabulated tests where τ ranges from 0.1 to 0.9, we observe consistent performance across thresholds, with FinBERT's underperformance remains evident. For Chinese and English content, we employ the respective versions of FinBERT to ensure language-specific validity.

cated post subgroups to evaluate the upper and lower bounds of model performance. In Panel B, we show that Sentiment-GPT consistently outperforms other models within ordinary posts. Notably, the F_1 *Score* of Sentiment-GPT in detecting neutral texts achieves 95.38%, compared to 75.00% for GPT-5, 76.19% for GPT-4.1, 78.39% for GPT-4o, 72.45% for GPT-4-turbo, and 50.00% for FinBERT. The overall *Accuracy* of Sentiment-GPT reaches 97.00%.

Within complicated posts, as shown in Panel C, the F_1 *Score* of Sentiment-GPT in detecting neutral texts is 74.03%, while that of GPT-5, GPT-4.1, GPT-4o, GPT-4-turbo, and FinBERT is 48.44%, 52.97%, 50.28%, 46.86%, and 48.26%, respectively. The overall *Accuracy* of Sentiment-GPT is 84.33%. These results indicate that model performance declines due to the unreadability of texts, highlighting the necessity of fine-tuning based on specialized corpora, especially given the prevalence of complicated posts in practice.

In summary, according to the evaluation of Sentiment-GPT and other existing LLMs, we show that the definition can be effectively incorporated through fine-tuning LLMs on a specialized corpus. Sentiment-GPT significantly outperforms GPT-5, GPT-4.1, GPT-4o, and GPT-4-turbo with the sentiment-defined prompt. However, using prompt engineering alone on the state-of-the-art LLMs fails to achieve this improvement. Prompt engineering seems to hardly help GPT-5, GPT-4.1, GPT-4o, and GPT-4-turbo truly "understand" the definition. Limited by original constraints of BERT, FinBERT is unlikely to perform in-context understanding and follow the definition without additional fine-tuning. Therefore, the limitations of prompt engineering and small-size LLMs make it evident that model fine-tuning via a comprehensive corpus, although requiring much more data and effort, is a more promising avenue that leads to competent sentiment extraction.

[Table 2 about here.]

3.2 Look-ahead Bias

While LLMs demonstrate impressive capabilities with textual data, their performance may partly reflect a look-ahead bias. This bias arises when information relevant to a model's predictions is already embedded in its training set. Recent financial studies have examined the issue and proposed mitigation strategies.[18] We argue that Sentiment-GPT's

[18]For example, on the content side, Engelberg et al. (2025) show that removing all identifying information from texts can effectively reduce look-ahead bias, since LLMs cannot recognize entities in "neutered"

performance is unlikely to be affected by this concern. First, sentiment classification relies exclusively on the textual content provided, not on any prior knowledge about specific firms or their historical performance. Our model evaluates sentiment based solely on the linguistic features and expressed opinions within each text, making it independent of any background information about particular companies that may have been encountered during pre-training. Second, our sentiment definition and classification framework represent new conceptual knowledge that was not available in any pre-training data, ensuring that performance gains stem from learning our unique approach rather than from prior exposure to firm-specific information.

Therefore, the superior performance of Sentiment-GPT is unlikely to be driven by look-ahead bias related to its base model, Qwen2.5-72B. To further validate this point, we replicate the tests in Table 2 using the base Qwen2.5-72B model to assess whether it had already internalized forward-looking information relevant to our definition and corpus. The results, reported in Appendix Table A1, show that without fine-tuning Qwen2.5-72B performs substantially worse, with an overall *Accuracy* of only 74.00% and an F_1 *Score* less than 60.59%. This evidence suggests that Sentiment-GPT's strong performance arises from learning our unique corpus during fine-tuning rather than from forward-looking information acquired during pre-training.

3.3 Generalized Performance

A central question is whether Sentiment-GPT generalizes beyond its training domain: after fine-tuning on a Chinese corpus, can it still apply our definition across different languages, text genres, and input lengths? To address this, we conduct four external validations: (i) challenging scenarios requiring consistent adherence to the definition, such as descriptive news without attitudes or non-targeted commentaries, (ii) cross-language transfer, (iii) stylistic variation in social media, and (iv) long-form financial journalism.

We start by testing whether our model follows the two core instructions in our definition when applied to out-of-distribution texts. Panel A of Table 3 shows the results. First, using 150 descriptive news articles that should be classified as neutral, Sentiment-GPT achieves an accuracy of 92.00%, outperforming GPT-5 (86.00%), GPT-4.1 (84.67%),

texts. On the model training side, He et al. (2025) propose training chronologically consistent LLMs exclusively on historical data available at the time.

GPT-4o (80.67%), GPT-4-turbo (82.67%), and FinBERT-tone (79.33%). This confirms that Sentiment-GPT can correctly classify factual reporting as neutral despite the presence of positive or negative words. Second, to test whether the model is misled by non-targeted opinions, we analyze 150 articles expressing positive or negative views about one asset but replace the firm name with another unrelated one in the prompt. Under our definition, these commentaries should still be classified as neutral toward the non-targeted asset.[19] On this task, Sentiment-GPT again leads with an accuracy of 89.33%, compared to GPT-5 (84.67%), GPT-4.1 (82.67%), GPT-4o (81.33%), GPT-4-turbo (77.33%), and FinBERT-tone (80.67%).

Next, we replicate the zero-shot evaluation in Table 2 using an English translation of Libra. As shown in Panel A of Table 3, Sentiment-GPT achieves an overall accuracy of 89.33%, nearly identical to the 90.67% obtained on the original Chinese dataset. This consistency suggests that the model's reliance on the definition, rather than language-specific features, underpins its cross-language robustness. In addition, competing LLMs exhibit the same relative ranking as in Chinese, further reinforcing that the observed stability is not model-specific but highlights Sentiment-GPT's comparative advantage.

Additionally, we conducted an additional evaluation to ensure our results were not simply due to stylistic similarities with the training corpus (Andromeda). We used financial social media posts from Stocktwits, which provided a strict test of generalization. This dataset contains both native and non-native English with informal and varied writing styles that differ significantly from Andromeda. We annotated the posts using our definition and created a balanced benchmark of 210 samples (70 positive, 70 negative, and 70 neutral) that is independent of Andromeda. Panel B of Table 3 shows that Sentiment-GPT again outperformed all competing models in per-class metrics and overall accuracy, achieving 87.62%. These results demonstrate that the model's superior performance is not limited to the training data style but extends to financial texts from broader, less structured domains.

Finally, large language models often struggle with extended contexts, which poses challenges for analyzing long-form financial journalism. Therefore, we test Sentiment-GPT's ability to handle lengthy texts by evaluating its performance on 135 long articles evenly split across positive, neutral, and negative sentiment. We annotated each article

[19] For example, when asked to extract sentiment toward NVDA from the commentary "I'm bearish on AAPL, and I love it", the correct classification is "neutral" toward NVDA.

using our definition and compared the results with model outputs. Panel C of Table 3 shows that Sentiment-GPT again outperformed all baselines, achieving 88.89% overall accuracy. While all models performed worse on longer inputs, Sentiment-GPT maintained a clear advantage. This indicates that the benefits of fine-tuning extend beyond short-form posts. Since these articles are in native English, the results provide additional evidence of the model's cross-language robustness across different text lengths and domains.

[Table 3 about here.]

Taken together, these results demonstrate that Sentiment-GPT's definition-driven framework remains robust across languages, domains, text lengths, and challenging scenarios involving factual content and non-targeted sentiment. This supports the view that language serves primarily as a transmission channel, while sentiment evaluation operates in a language-agnostic conceptual space (Fedorenko et al., 2024; Rathje et al., 2024; Wendler et al., 2024). For financial applications, such robustness suggests that Sentiment-GPT can be reliably applied in multilingual markets and heterogeneous information environments where investor sentiment shapes asset prices.

4 Sentiment Analysis of Financial Commentary

4.1 Sourcing Article

In this section, we use professional commentary articles from the "opinion columns" of the Wall Street Journal (WSJ), the New York Times (NYT) and the Financial Times (FT) as our primary text source for sentiment analysis. This choice is motivated by three key reasons. First, our sentiment measures are designed to extract genuine personal opinions and emotions, and the selected articles are explicitly labeled as opinions or commentary, aligning perfectly with our objective. Second, these articles are authored by professionals, ensuring high-quality, well-structured content. Compared to the unstructured, often ambiguous, and scattered nature of social media texts, these structured articles provide a clear thematic focus, making them an ideal data source for our analysis. Finally, these media platforms are among the most popular sources not only for finance literature studying the effect of financial news on markets (e.g., Huberman & Regev, 2001; Tetlock, 2007;

Dougal et al., 2012; Garcia, 2013; Goldman et al., 2024) but also for widely circulated national media disseminating market insights to both U.S. domestic and international investors.

We examine all articles about the largest 100 U.S. firms from January 1992 to December 2024.[20] We focus on these large-cap firms because only large firms are likely to receive substantial media attention, especially given the highly uneven distribution of coverage between big and small firms. Since the composition of the largest 100 firms varies over time, we include firms as long as they have entered the top 100 at least once during the sampling period. This results in a sample of 55,275 articles covering 238 firms. We then obtain and match firms' corresponding return data and firm characteristics from the CRSP.

4.2 Categorizing Sentiment Based on the word-based Method

The sentiment classifications suggested by our fine-tuned LLM, referred to as Senti-GPT henceforth, naturally take the form of negative (-1), neutral (0), and positive (1). However, constructing the word-based sentiment requires an additional step to convert word counts into categorical sentiments, as the word-based method only returns counts of positive and negative words. We first employ the conventional dictionary outlined in Loughran & McDonald (2011) to calculate the number of positive and negative words. The sentiment derived from the word-based method, referred to as LM_τ henceforth, is then computed using the following equations:

$$\text{LM}_\tau = \begin{cases} 1, & \text{if } \text{LM}_{\text{ratio}} > \tau, \\ -1, & \text{if } \text{LM}_{\text{ratio}} < -\tau, \\ 0, & \text{if } -\tau \leq \text{LM}_{\text{ratio}} \leq \tau, \\ 0, & \text{if } P_{\text{count}} = N_{\text{count}} = 0, \end{cases} \tag{8}$$

where τ is the threshold used for sentiment classification, and LM_{ratio} is defined as:

$$\text{LM}_{\text{ratio}} = \frac{P_{\text{count}} - N_{\text{count}}}{P_{\text{count}} + N_{\text{count}}}, \tag{9}$$

[20] The selection of 1992 as the inception year is primarily motivated by the fact that the annual volume of media articles surpassed 1,000 starting this year, ensuring adequate data coverage.

where P_{count} and N_{count} represent the number of positive and negative words, respectively, in each text. In plain terms, LM_τ is classified as positive (1) if LM_{ratio} exceeds the threshold τ, classified as negative (-1) if it falls below $-\tau$, and classified as neutral (0) otherwise. Instances where both P_{count} and N_{count} are zero, rendering LM_τ inapplicable, are labeled as neutral (0).

4.3 How Does Our Sentiment Differ: A Controlled Experiment

To better understand the distinctions between our model and those word-based methods, we created a controlled experiment. We generated eight articles using GPT-4o, each strategically crafted in the style of the Wall Street Journal and intentionally designed to manipulate the use of positive and negative lexicon as defined by Loughran & McDonald (2011). For interested readers, all prompts used and full-length articles can be found in Appendix E.

[Table 4 about here.]

Table 4 summarizes the characteristics we use to generate each article and their corresponding sentiment assessments. The "Narrative" column indicates whether negative (-1) or positive (1) words are used to generate content related to company fundamentals, news, and financial disclosures. A value of 1 & -1 signifies that both negative and positive words are applied equally in content generation. The "Attitude" column represents whether the article contains explicit negative (-1) or positive (1) opinions or emotions. P_{count} and N_{count} denote the word counts of positive and negative words, respectively, while $P - N$ is the difference between the positive and negative word counts. $\text{LM}_{0.25}$ is the LM indicator using $\tau = 0.25$. Finally, the last column, "Senti-GPT", displays the sentiment classification generated by our Sentiment GPT.

The first two articles used positive (Article 1) and negative (Article 2) terms from the lexicon to describe fundamentals, company disclosures, and financial data, without expressing any personal opinions or emotions. These examples show that a word-based method, which relies on the frequency of positive or negative terms, would likely assign a corresponding positive or negative sentiment to the texts. However, according to our approach, our method remains unaffected by the presence of these sentimental words when no personal opinion is conveyed (Senti-GPT = 0).

The next pair, Articles 3 and 4, presented a deliberate mismatch between the sentiment of the lexicon words used and the personal opinions expressed. Article 3 used negative lexicon words to describe company narrative while expressing positive personal opinions, whereas Article 4 used positive lexicon words in stating company narrative but conveyed negative personal sentiments. These cases highlight a critical weakness in the word-based approach: it systematically misclassifies the overall sentiment by focusing solely on word frequency, thereby ignoring or distorting the actual personal opinions and emotions expressed. In contrast, our method accurately captures the true sentiment based on the personal expressions within the text, regardless of the overall positive or negative view from the word count.

The third pair, Articles 5 and 6, presented company narratives containing an equal mix of positive and negative terms but differed in their expression of personal opinions and emotions—explicitly positive in Article 5 and explicitly negative in Article 6. In posts like these, the margin between positive and negative word counts is often narrow, making word-based sentiment analysis methods highly unreliable. In contrast, our model effectively identifies personal attitudes in these scenarios, where the blend of positive and negative language complicates the interpretation of genuine sentiment.

The final two articles, Articles 7 and 8, described company narratives using positive and negative lexicon words, respectively, but also included explicit positive and negative personal opinions and emotions. This set illustrates scenarios where both the word-based method and our method align in their sentiment assessments since the use of sentiment words coincides with the expressed personal opinions. In this regard, our sentiment measures, although based on a distinct definition, would still align with conventional methods in some cases.

Combining all examples highlights the limitations of relying solely on word frequency to accurately capture semantic nuances and personal sentiments. Our analysis demonstrates that articles can contain a mix of positive and negative words while maintaining an overall neutral sentiment. Conversely, articles with predominantly negative or positive words can still convey opposing personal opinions or emotions. This illustrates the critical need for more nuanced sentiment analysis methods that go beyond simple word counts to truly identify and understand underlying personal expressions—a gap our method aims to fill.

4.4 Summary Statistics

Figure 2 visualizes the distribution of three sentiment measures, where $\text{LM}_{0.25}$ and $\text{LM}_{0.5}$ represent calculating LM_τ using 0.25 and 0.5, respectively. The 0.25 threshold is an intuitive choice for discretizing LM_{ratio}, i.e., an article with 69 positive words and 31 negative words would be classified as posiive, as $(69-31)/(69+31) = 0.38 > 0.25$. In contrast, an article with 60 positive words and 40 negative words would result in $(60-40)/(60+40) = 0.20 < 0.25$, thus being regarded as neutral. Higher thresholds such as $\tau = 0.5$ require an even larger margin of net positive words to classify an article as positive, and vice versa for negative sentiment. We use the 0.5 threshold to represent a stricter criterion, recognizing non-neutral sentiment only when a strong sentiment or tone is present.

From Figure 2, we observe a notable concentration of neutral opinions among WSJ, NYT and FT articles, despite all of them being labeled in the "opinion" sections. The mean of Senti-GPT is only -0.056, with a standard deviation of 0.345, and a high proportion of neutral classifications (87.8%). In contrast, the $\text{LM}_{0.25}$ and $\text{LM}_{0.5}$ exhibit more dispersed distributions across negative and positive sentiments. Specifically, $\text{LM}_{0.25}$ has a mean of -0.382, a standard deviation of 0.596, and only 34.2% neutral classifications. Even under a more strict setting to qualify non-neutral sentiment classifications, $\text{LM}_{0.5}$ still exhibits fewer neutral sentiments than Senti-GPT does. Specifically, $\text{LM}_{0.5}$ has a mean of -0.263, a standard deviation of 0.716, and 57.5% neutral classifications.

[Figure 2 about here.]

Figure 3 illustrates the evolving relationship between the two sentiment series by plotting the rolling correlation between Senti-GPT and $\text{LM}_{0.25}$ using a twelve-month window. To facilitate comparison, we average the daily sentiment series into weekly data to calculate the correlation per week. The figure reveals frequent periodic reversals in correlation, indicating that the two series are independent. This suggests that, with a focus on personal opinions and emotions, Senti-GPT captures a distinct rationale compared to the word-based approach.

[Figure 3 about here.]

4.5 Optimism and Pessimism of Stock Market

Building on the framework introduced by De Long et al. (1990), most models of sentiment theory posit the existence of two types of agents: noise traders and sophisticated investors who trade against them. Consistent with this framework, we test the implications of sentiment theory by examining whether opinions or emotions expressed in the Wall Street Journal, New York Times, and Financial Times contribute to irrational optimism or pessimism, thereby inducing noisier trading and price distortions

As sentiment theory predicts, noise traders buy (sell) stocks from (to) arbitrageurs after encountering positive (negative) narratives. If the theory holds, we should see higher (lower) abnormal returns during the publication period following positive (negative) narratives. The testable hypotheses are:

H1: Publication-period cumulative abnormal return is positively associated with positive sentiment of commentary articles.

H2: Publication-period cumulative abnormal return is negatively associated with negative sentiment of commentary articles.

However, new fundamentals in media coverage is naturally expected to influence short-term asset prices, which aligns with predictions from sentiment theory. To further distinguish between the sentiment theory and the information theory, we also investigate stock price reactions over a longer horizon. Specifically, sentiment theory predicts that short-term returns will revert in the long run, whereas information theory suggests they will persist. The testable hypotheses are:

H3: Post-publication cumulative abnormal return is negatively associated with positive sentiment of commentary articles.

H4: Post-publication cumulative abnormal return positively associated with negative sentiment of commentary articles.

Another key distinction between sentiment theory and information theory lies in their implications for trading volume. According to sentiment theory, noise traders are influenced by sentiment or narratives, even when these sentiments or narratives convey no new fundamental information. In this context, we should expect increasing aggregate disagreement between noise traders and sophisticated investors. Since higher disagreement can drive up subsequent trading volume, we would expect to observe higher abnormal

trading volume following both positive and negative narratives.[21] In contrast, information theory does not provide clear predictions regarding trading volume. For example, as argued in Tetlock (2007), additional soft information may lead to either disagreements or agreements among investors.

H5: Publication-period abnormal volume is positively associated with positive sentiment of commentary articles.

H6: Publication-period abnormal volume is positively associated with negative sentiment of commentary articles.

4.5.1 Media Sentiment and Stock Returns

We begin by testing Hypotheses 1 and 2 to evaluate whether the textual sentiment in commentary articles affects related stock prices. To clearly test the hypotheses, we focus on articles published on trading days. Our main regression specification is as follows:

$$CAR_i\ (t, t+2) = \alpha_{i,t} + \beta_1 Positive_{i,t} + \beta_2 Negative_{i,t} + \boldsymbol{X_{i,t-1}}\Gamma + \alpha_t + \alpha_j + \alpha_s + \varepsilon_{i,t}, \quad (10)$$

where $CAR_i\ (t, t+2)$ represents the cumulative abnormal return over trading days 0 to +2 around the commentary publication date (t). The CAR is calculated using the three Fama-French factors and momentum. $Positive_{i,t}$ and $Negative_{i,t}$ are set to 1 (and the other to 0) if firm i's average Senti-GPT on day t is positive or negative, respectively. Both are set to 0 if the average equals zero. Control variables ($\boldsymbol{X_{i,t-1}}$) include lagged cumulative abnormal returns (CAR over days -30 to -6 and CAR over days -5 to -1), lagged abnormal volume (the average trading volume for the week preceding the event, standardized based on the past 60 trading days before the filing date), and lagged volatility (the standard deviation of returns over the past 60 trading days). Day fixed effects (α_t), industry fixed effects (α_j), and article source fixed effects (α_s) are included.[22] Standard errors are double clustered at the firm and day level.

We also align LM indicators to the firm-day level by aggregating all commentaries for firm i on day t.[23] Specifically, $LM\ Positive_{i,t}$ ($LM\ Negative_{i,t}$) is an indicator variable equal to 1 if the aggregated $\text{LM}_{0.25}$ across all articles for firm i at day t is 1 (-1).

[21] See the review by Hong & Stein (2007) on disagreement, sentiment, and trading volume.

[22] The industry classification follows the 48 categories from Fama & French (1997). For detailed definitions, see https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_48_ind_port.html.

[23] We follow Garcia (2013) in aggregating articles, as LM indicators are word-based, and aggregation

$LM\ Positive_{i,t}^{fraction}$ ($LM\ Negative_{i,t}^{fraction}$) refers to the fraction of positive (negative) LM words across all commentaries for firm i at day t, based on the Loughran & McDonald (2011) word dictionary.

The results are presented in Table 5, Panel A. In Columns 1 to 3 show estimates without fixed effects. We observe that when the textual sentiment is positive (negative), the CAR increases by an average of 45.2 bps (decreases by 87.8 bps) compared to when the sentiment is neutral, as shown in Column 1.[24] LM indicators do not show significant explanatory power in Column 2. In Column 3, the fraction of LM positive words shows no significant relationship with publication-period CAR, while the fraction of LM negative words negatively predicts CAR. Specifically, a one-standard-deviation increase in $LM\ Negative^{faction}$ corresponds to a 5.9 bps decrease in CAR. This aligns with previous studies highlighting the informativeness of negative words (e.g., Tetlock et al., 2008; Loughran & McDonald, 2011).

[Table 5 about here.]

Throughout Columns 4 to 6, we further control for day fixed effects, industry fixed effects, and article source fixed effects. Both the explanatory power of Senti-GPT indicators (*Postive* and *Negative* in Column 4) and that of $LM\ Negative^{faction}$ (Column 6) remains robust. *LM Positive* positively predicts publication-period CAR, however, its magnitude is roughly one-third that of the positive sentiment indicator.

To assess whether Senti-GPT or LM measures provide incremental sentiment information in predicting CARs, we estimate a regression that jointly includes both Senti-GPT and LM measures. The results in Panel B of Table 5 demonstrate that while Senti-GPT retains its statistical significance, the explanatory power of the LM measures becomes insignificant in this joint specification. This finding suggests that Senti-GPT captures

better preserves the firm-day-level lexical characteristics. Simply averaging may reduce the explanatory power of LM indicators.

[24]For ease of interpreting the results, we set *Neutral* as the reference category. One may alternatively select *Positive* or *Negative* as the reference category when examining the neutral sentiment effect, yielding comparable empirical interpretations. Specifically, when *Positive* is set as the reference category, the coefficients for *Neutral* and *Negative* are -0.452 and -1.330, respectively. This indicates that articles with neutral sentiment exhibit an average CAR 45.2 bps lower than those with positive sentiment, while articles with negative sentiment show a 133.0 bps reduction in CAR. Conversely, when *Negative* is the reference category, the coefficients for *Positive* and *Neutral* are 1.330 and 0.878, respectively. Here, articles with positive sentiment show an average CAR 133.0 bps higher than those with negative sentiment, whereas articles with neutral sentiment show a 87.8 bps decrease compared to negative articles.

the genuine sentiment component more precisely, whereas the LM dictionary's predictive power appears to stem largely from a blend of sentiment-related and non-sentiment-related content with some noise.

We provide additional robustness checks in Appendix Table B1, where CAR is calculated over different time windows (Panels A and B) and LM indicators are based on different thresholds (Panel C). The explanatory power of Senti-GPT remains robust, while in most cases, the explanatory power of LM words is fully absorbed.

Next, we examine whether the observed sentiment-driven optimism and pessimism during the publication period reverts in subsequent trading days (Hypotheses 3 and 4). We replicate the tests from Columns 4 to 6 of Table 5, replacing the cumulative abnormal return (CAR) with an extended window ranging from 0 to 63 trading days (equivalent to three trading months).

[Figure 4 about here.]

Figure 4 plots the results. Panel (a) shows that stock returns respond positively (negatively) to positive (negative) commentaries during the first two trading weeks, but these initial reactions gradually revert over a longer horizon. The return reversal following positive (negative) commentaries becomes statistically significant after 36 (57) trading days. These findings suggest that Senti-GPT successfully align with the sentiment theory where stock prices initially overreact to both positive and negative sentiment in commentaries, followed by a gradual reversal in the subsequent weeks. In contrast to much of the prior literature that only reports reversals in negative sentiment, our results also capture reversals on the positive sentiment side. Notably, the magnitude of the reversal is approximately three times greater than the initial reaction observed during the publication period. A potential explanation is that, during the reversal phase, new commentaries reflect the ongoing reversal in stock prices, and the accompanying media sentiment further amplifies the magnitude of the reversal.[25]

In Panel (b), LM indicators exhibit much weaker explanatory power in both short and long terms. Specifically, while stock returns respond positively to the positive LM

[25]In untabulated tests, we find that articles published following periods of higher (lower) recent CAR are more likely to display positive (negative) sentiment. When we control for the lagged average post-publication sentiment of commentaries, the magnitude of the reversal remains significant, while lagged sentiment positively predicts CAR. This suggests that, in the reversal phase, a longer time horizon allows not only reversal-related commentary but also sentiment from broader sources to more fully influence returns, thereby amplifying the observed reversal.

indicator during the publication period, there is no sign of a subsequent reversal. In contrast, returns reversals caused by negative LM indicators emerge after 17 trading days, however, it shows no clear pessimism during the early publication period. We provide more robustness checks in Appendix Figure B.1, where LM indicators are constructed using different thresholds τ (Panels (b) to (e)), and the fraction of LM words (Panel (f)). We observe consistent patterns for LM indicators.

4.5.2 Media Sentiment and Trading Volume

To test whether textual sentiment increases abnormal trading volume (Hypotheses 5 and 6), we perform similar tests to those in Panel A of Table 5, using the average abnormal trading volume (AVOL) over days 0 to +2 around the commentary publication date as the dependent variable.

[Table 6 about here.]

The results are presented in Table 6, Panel A. Column 1 shows that both positive and negative sentiment, as identified by Senti-GPT, significantly increase abnormal trading volume. Specifically, when an article's sentiment is positive (negative), the AVOL rises by an average of 22.5 basis points (33.1 basis points) compared to neutral sentiment. These findings align with prior research on LM negative words (e.g., Loughran & McDonald, 2011) and corroborate the positive relationship between trading volume and disagreement established by Cookson & Niessner (2020). In contrast, regressions using LM measures yield mixed results. Column 2 indicates that LM measures show no significant association with publication-period abnormal volume. Column 3 reveals that while the fraction of LM negative words remains unrelated to publication-period AVOL, the fraction of LM positive words predicts higher abnormal volume. Specifically, a one-standard-deviation increase in the fraction of positive LM words corresponds to a 2.4 basis point increase in AVOL.

The findings from Columns 1 and 2 are reinforced in Columns 4 and 5, which incorporate day fixed effects, industry fixed effects, and article source fixed effects. However, after including these fixed effects, the explanatory power of *LM positive* weakens, as demonstrated in Column 6.

To explore whether Senti-GPT or LM measures provide incremental sentiment information in predicting AVOL, we estimate a regression that jointly includes both Senti-GPT

and LM measures. The results in Panel B of Table 6 show that both positive and negative sentiment, as identified by Senti-GPT, positively predict AVOL, consistent with predictions from sentiment theory. However, once Senti-GPT measures are included, the negative coefficient on negative LM words turns significant. This finding suggests that while negative LM words exhibit explanatory power, they do not appear to capture investor sentiment in a manner consistent with the predictions of sentiment theory.

Robustness checks, presented in Appendix Table B1, examine AVOL calculated over different time windows (Panels A and B), while Panel C tests LM indicators based on different thresholds. The explanatory power of Senti-GPT remains robust, while only negative LM words negatively predict AVOL, which is inconsistent with predictions from sentiment theory.

In summary, the results from Sections 4.5.1 and 4.5.2 consistently demonstrate that textual sentiment is associated with both stock prices and trading volume, aligning with predictions from sentiment theory. In contrast, LM measures yield mixed results, with their explanatory power highly dependent on the specific construction of the measure. While some LM measures are associated with returns, none consistently explain short-term price reactions, long-term price reversals, and short-term abnormal trading volume. These findings suggest that textual sentiment offers a more reliable proxy for textual sentiment, whereas the explanatory power of LM measures is unlikely to stem from their ability to capture genuine sentiment.

4.6 Betting on Sentiment

To assess the economic significance of the relationship between textual sentiment and stock prices, we implement a naïve backtesting analysis that compares our approach with conventional word-based methodologies. We focus on analyzing strategies that go long (short) on firms associated with positive (negative) sentiment signals over short horizons. The primary motivation is that capitalizing on immediate, anticipatory price movements may be more attractive for sophisticated investors ("smart money") than waiting for the potentially slow and uncertain reversion of noise traders' beliefs to their mean (Shleifer & Vishny, 1990). In the same time, long-term reversion often involves taking positions contingent on price changes occurring far from the initial signal date, a process inherently subject to greater uncertainty.

Similar to the regression analysis, our backtest spans the period from January 1992 to December 2024, encompassing the identical firms considered in previous regression analysis. We construct long-short portfolios based on two distinct sentiment measures: Senti-GPT and $\text{LM}_{0.25}$. Specifically, At the market closing time of day t, we establish a long portfolio consisting of stocks exhibiting positive sentiment on day t and a short portfolio comprising stocks with negative sentiment on the same day, using adjusted closing prices for portfolio construction. We then execute positions accordingly using adjusted closing prices on day $t+1$ or the nearest subsequent trading day in cases where the subsequent day following the sentiment signal is a non-trading day. On days with no trading signals, the portfolio earns the 3-month US Treasury rate.

The results, presented in Table 7, demonstrate that Senti-GPT consistently outperforms $LM_{0.25}$ across three sample periods: 1992-2024 (full sample), 2001-2024 (post-tech bubble), and 2009-2024 (post-financial crisis). We first consider a non-transaction cost scenario to highlight the difference in trading results. During 2001-2024, for instance, the Senti-GPT portfolio generates an average annual return of 28.0%, significantly exceeding $\text{LM}_{0.25}$'s 16.2%. This performance advantage persists in risk-adjusted metrics, with Senti-GPT achieving Sharpe ratios between 0.564 and 0.687 across periods, compared to $\text{LM}_{0.25}$'s range of 0.247-0.376. The terminal wealth metrics further underscore this difference: a \$1 investment in the Senti-GPT portfolio grows to \$289.495 by 2024 in the full sample, more than fifteen times $\text{LM}_{0.25}$'s \$19.18. Importantly, although both strategies exhibit high volatility, particularly over the full sample period (1992–2024) which spans two major market crises, their performance differentials remain significant, implying that the divergence in returns is driven by superior signal efficacy rather than differences in risk exposure.

[Table 7 about here.]

We further confirm that Senti-GPT's superiority is not attributable to the choice of threshold τ by replicating the word-based trading backtest over the 1992-2024 period. As shown in Table 8, traditional word-based approaches consistently underperform across all threshold selections. These strategies yield Sharpe ratios ranging from 0.214 to 0.409, with terminal values of a \$1 investment never exceeding \$29.183 in any configuration - substantially below the \$289.495 achieved by the Senti-GPT benchmark. Notably, while

varying LM thresholds exhibit similar volatility levels, they fail to deliver corresponding improvements in returns. This pattern suggests that these models cannot be significantly improved through parameter tuning. The rightmost column, representing Senti-GPT, serves as a benchmark, outperforming all LM variants by significant margins. This stark performance again, signals that traditional word-based methods may be fundamentally limited in capturing the nuanced, context-dependent nature of textual sentiment.

[Table 8 about here.]

Figure 5 compares the cumulative performance of two long-short, sentiment-driven strategies with the S&P 500 buy-and-hold strategy. Initial wealth is normalized to \$1 for all strategies. The Y-axis is log-scaled to facilitate comparison. To reflect a more practical trading environment, in addition to a frictionless setting (zero transaction costs), we apply two levels of transaction costs: 2.5 basis points per transaction (5 bps per round trip) and 5 basis points per transaction (10 bps per round trip).

Overall, the Senti-GPT strategy delivers substantially higher compounded returns than the dictionary-based portfolio, both with and without transaction costs. This sustained outperformance, consistent with the regression results, underscores the economic value of robust sentiment measures in exploiting sentiment-driven mispricings.

The shaded red regions denote three major economic downturns: the dot-com crash (March–November 2001), the Global Financial Crisis (December 2007–June 2009), and the COVID-19 pandemic shock (February–April 2020). These recession periods serve as natural stress tests. During these turbulent episodes, the Senti-GPT strategy maintains strong relative performance, whereas the word-based strategy exhibits pronounced vulnerability, particularly when compared with the S&P 500 benchmark. This contrast highlights the inability of word-based approaches to capture the nuanced sentiment signals embedded in financial texts.

Both strategies are sensitive to transaction costs. Yet even under realistic frictions, Senti-GPT continues to outperform the benchmark, while the dictionary-based method struggles to remain profitable—barely generating returns even at the low-cost setting (5 bps per round trip). These findings suggest that contextual sentiment analysis via large language models yields signals that remain tradable after costs, in sharp contrast to traditional bag-of-words methods that rapidly lose efficacy.

This performance drop after accounting for transaction costs can be attributed to two primary factors. First, our trading implementation follows a deliberately simple design to ensure fair comparison between sentiment methodologies. Positions are established immediately upon receiving sentiment signals and are liquidated uniformly the following day, without any signal weighting, timing optimization, or extended holding periods. While this naive approach facilitates transparency, it could be substantially improved through more sophisticated trading designs. Second, the relatively limited media coverage from our selected sources (WSJ, NYT, and FT) generates few daily signals for many less popular firms, constraining diversification and reducing risk-adjusted returns. With broader text sources and more comprehensive firm coverage, the strategy could in principle produce a larger number of signals, leading to improved diversification and enhanced performance. Such optimizations, however, lie beyond the scope of this study. Our primary objective is to demonstrate the fundamental economic advantages of attitude-focused sentiment measures over traditional word-based approaches. The observed performance differential under this controlled setting underscores the superior information content captured by contextual sentiment analysis.

[Figure 5 about here.]

To further assess the sources of return for our investment strategy and to determine if it generates statistically significant abnormal returns (alpha), we employ the Fama-French five-factor model (Fama & French, 2015). The core logic is to regress the strategy's excess returns against these five systematic risk factors. The general specification of the model is:

$$R_t - R_{f,t} = \alpha + \beta(R_{M,t} - R_{f,t}) + s\,\mathrm{SMB}_t + h\,\mathrm{HML}_t + r\,\mathrm{RMW}_t + c\,\mathrm{CMA}_t + \varepsilon_t \quad (11)$$

where $R_t - R_{f,t}$ is the excess return of our long-short strategy over the risk-free rate $R_{f,t}$ in month t, spanning the period of January 1992 to December 2024. α is the intercept term, representing the abnormal return. A statistically significant positive alpha indicates that the strategy has outperformed the benchmark factors. $R_{M,t} - R_{f,t}$, SMB_t, HML_t, RMW_t, and CMA_t are the returns of the five factors in month t. ε_t is the idiosyncratic error term for month t. The five factors represent distinct risk dimensions in the market: $R_{M,t} - R_{f,t}$ captures the market risk premium; SMB_t (Small Minus Big) represents the

size premium between small and large-cap stocks; HML_t (High Minus Low) measures the value premium between high and low book-to-market stocks; RMW_t (Robust Minus Weak) captures the profitability premium between firms with robust and weak operating profitability; and CMA_t (Conservative Minus Aggressive) represents the investment premium between firms with conservative and aggressive investment patterns. All factor data are downloaded from the data library maintained by Kenneth R. French. For this analysis, we utilized monthly returns for both our strategy and the Fama-French factors, ensuring that any potential sentiment effects, which our strategy aims to capture, are aligned with the appropriate resolution of the factor data.

The Ordinary Least Squares (OLS) regression results, presented in Table 9, provide the strategy's performance characteristics. The estimated monthly alpha is 2.457% and is statistically significant at the 0.1% level (p-value = 0.001). This suggests that, after accounting for exposures to the five systematic risk factors, our strategy generated an average excess return of approximately 2.457% per month. The provided annualized alpha of 0.295 (or 29.5%) further highlights this robust positive performance, strongly suggesting that the strategy delivers substantial abnormal returns not explained by common risk factors.

The sentiment-based strategy exhibits a statistically significant coefficient for the market risk premium (Mkt-RF) of -0.630 (p-value = 0.001). This negative beta indicates a tendency for the strategy to perform counter to the broader market, a characteristic that can be attributed to active portfolio selection. In contrast, the coefficients for the other Fama-French factors: SMB (0.186), HML (0.092), RMW (0.023), and CMA (-0.113) are all statistically insignificant, with p-values of 0.469, 0.757, 0.945, and 0.800, respectively. This lack of significance implies that the strategy's returns are not systematically driven by exposures to size, value, profitability, or investment patterns as defined by these factors.

[Table 9 about here.]

In summary, the results in this section clearly demonstrate the superiority of attitude-oriented sentiment in constructing profitable and robust sentiment-driven portfolios. Portfolios based on Senti-GPT consistently deliver higher returns, stronger risk-adjusted performance, and greater value creation across all holding periods. This outperformance

highlights Senti-GPT's ability to capture nuanced, context-specific personal opinions and emotions that are most effective in fueling spreadable narratives within the investor community. In contrast, the underperformance of word-based approaches underscores their limitations, as they rely solely on word frequency and lack the semantic depth required to identify and interpret people's attitudes in text.

5 Conclusion

While many existing studies find that financial texts can move stock prices, empirical evidence on the impact of sentiment in texts remains limited, largely due to the absence of a clear definition and, consequently, a precise measurement of investor sentiment. In this paper, we address this empirical gap by proposing a novel sentiment classification framework that includes a clear definition of sentiment, a specialized corpus aligned with this definition, and a fine-tuned LLM trained on the corpus.

Using our fine-tuned sentiment model, we analyze the relationship between sentiment in articles in the Wall Street Journal, the New York Times, and the Financial Times and stock prices. Our analysis reveals three key findings. First, positive (negative) sentiment in commentary articles predicts higher (lower) cumulative abnormal returns, with significant price reversals occurring after both positive and negative sentiment shocks. Second, both types of sentiment measure significantly increase abnormal trading volume. These results provide empirical support De Long et al. (1990). In contrast, we find no robust relationship between the sentiment generated by traditional word-based approaches and market responses. Furthermore, a trading strategy based on our sentiment classification substantially outperforms the word-based trading strategy, highlighting the superior economic relevance of our approach.

Beyond these facts, our contribution is methodological. We develop a compact, article-level fine-tuning framework that isolates sentiment, the author's evaluative stance, from fundamentals discussed in the same text. Our methodology demonstrates that fine-tuning an LLM for a specific task can outperform general-purpose models such as ChatGPT. Our results underscore that measurement quality is first-order for detecting sentiment-driven price pressure and its unwind. They invite work extending the approach to improve the quantification of variables of interest from texts in various economic set-

tings.

References

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

Azimi, M., & Agrawal, A. (2021). Is positive sentiment in corporate annual reports informative? evidence from deep learning. *The Review of Asset Pricing Studies*, *11*, 762–805.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance*, *61*, 1645–1680.

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

Bin Hasan, S., Kumar, A., & Taffler, R. (2024). Investor emotions and asset prices. *SSRN Electronic Journal*, . Available at SSRN: https://ssrn.com/abstract=5020654.

Bybee, J. L. (2023). The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, *70*, 223–260.

Chen, H., De, P., Hu, Y., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*, 1367–1403.

Chen, H., & Hwang, B.-H. (2022). Listening in on investors' thoughts and conversations. *Journal of Financial Economics*, *145*, 426–444.

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

Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2024). The social signal. *Journal of Financial Economics*, *158*, 103870.

Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2025). Market signals from social media. *Available at SSRN*, .

Cookson, J. A., & Niessner, M. (2020). Why don't we agree? evidence from a social network of investors. *The Journal of Finance*, *75*, 173–228.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*, 1375–1388.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*, 703–738.

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

Dougal, C., Engelberg, J., Garcia, D., & Parsons, C. A. (2012). Journalists and the stock market. *The Review of Financial Studies*, *25*, 639–679.

Engelberg, J., Manela, A., Mullins, W., & Vulicevic, L. (2025). Entity neutering. *Available at SSRN*, .

Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. *The Review of Financial Studies*, *33*, 1184–1216.

Fama, E. F., & French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, *43*,

153–193.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*, 1–22.

Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. (2024). Language is primarily a tool for communication rather than thought. *Nature*, *630*, 575–586.

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

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

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*, 535–574.

Goetzmann, W. N., Kim, D., & Shiller, R. J. (2024). *Emotions and subjective crash beliefs*. Technical Report National Bureau of Economic Research.

Goldman, E., Gupta, N., & Israelsen, R. (2024). Political polarization in financial news. *Journal of Financial Economics*, *155*, 103816.

Graeber, T., Roth, C., & Zimmermann, F. (2024). Stories, statistics, and memory. *The Quarterly Journal of Economics*, *139*, 2181–2225.

He, S., Lv, L., Manela, A., & Wu, J. (2025). Chronologically consistent large language models. *arXiv preprint arXiv:2502.21206*, .

Hong, H., & Stein, J. C. (2007). Disagreement and the stock market. *Journal of Economic Perspectives*, *21*, 109–128.

Huang, A. H., Wang, H., & Yang, Y. (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*, 806–841.

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review*, *89*, 2151–2180.

Huberman, G., & Regev, T. (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *The Journal of Finance*, *56*, 387–396.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, *110*, 712–729.

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

Ke, S. (2024). Analysts' belief formation in their own words. *Available at SSRN 5025830*, .

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kim, A., Muhn, M., & Nikolaev, V. (2023). From transcripts to insights: Uncovering corporate risks using generative AI. *arXiv preprint arXiv:2310.17721*.

Kim, A., Muhn, M., & Nikolaev, V. V. (2024). Bloated disclosures: Can chatgpt help investors process information? *Chicago Booth Research Paper*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? Return predictability

and large language models. *arXiv preprint arXiv:2304.07619*.

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

OpenAI (2023). Gpt-4 technical report. arXiv:2303.08774.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*, e2308950121.

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, *107*, 967–1004.

Shleifer, A., & Vishny, R. W. (1990). Equilibrium short horizons of investors and firms. *The American Economic Review*, *80*, 148–153.

Soo, C. K. (2018). Quantifying sentiment with news media across local housing markets. *The Review of Financial Studies*, *31*, 3689–3719.

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

Tetlock, P. C. (2014). Information transmission in finance. *Annu. Rev. Financ. Econ.*, *6*, 365–384.

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

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024). Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*, .

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zhou, G. (2018). Measuring investor sentiment. *Annual Review of Financial Economics*, *10*, 239–259.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., & Han, J. (2023). Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

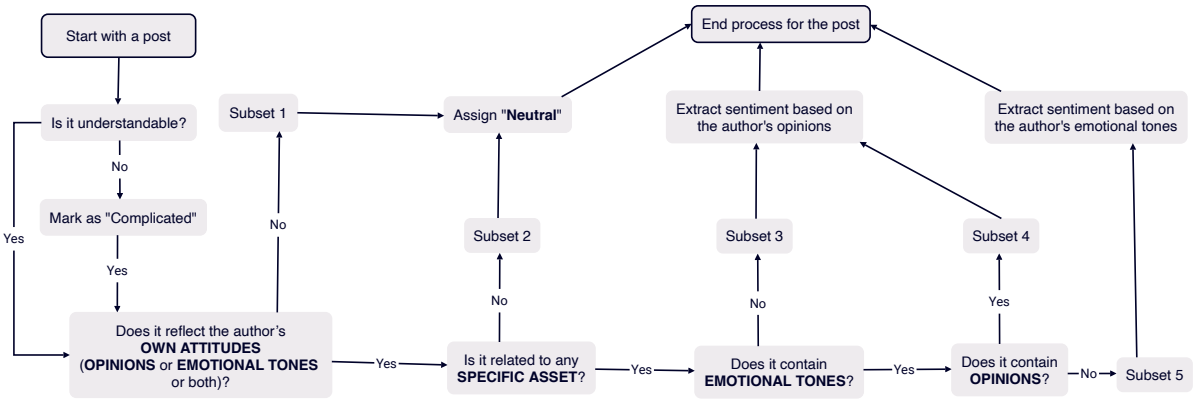


Figure 1: Flowchart of the Human Annotation

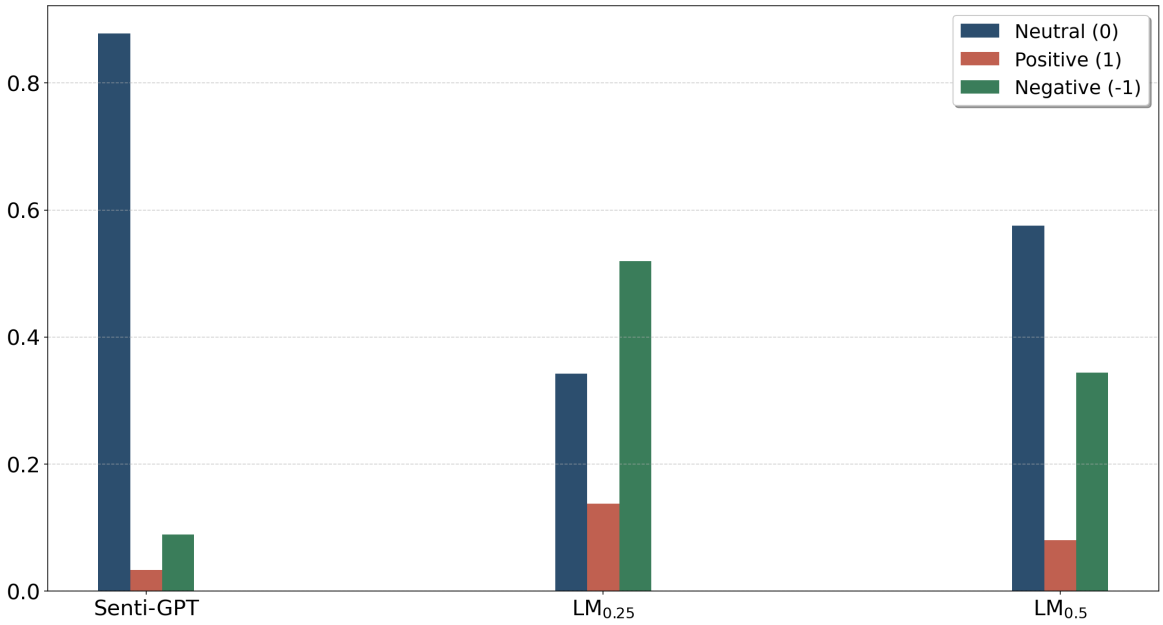


Figure 2: Proportions of the Three Sentiment Classifications

This figure shows the proportions of the three sentiment categories (Positive, Neutral, and Negative) for the three sentiment indicators. $LM_{0.25}$ and $LM_{0.5}$ are calculated using the word-based method with thresholds $\tau = 0.25$ and $\tau = 0.5$, respectively.

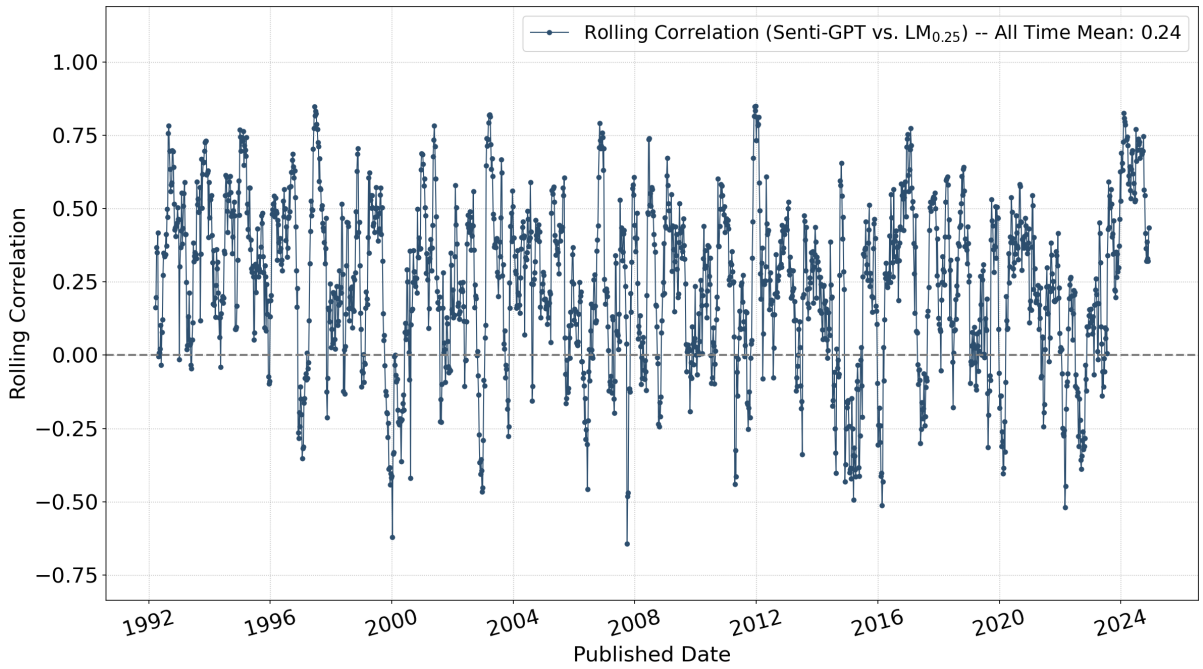
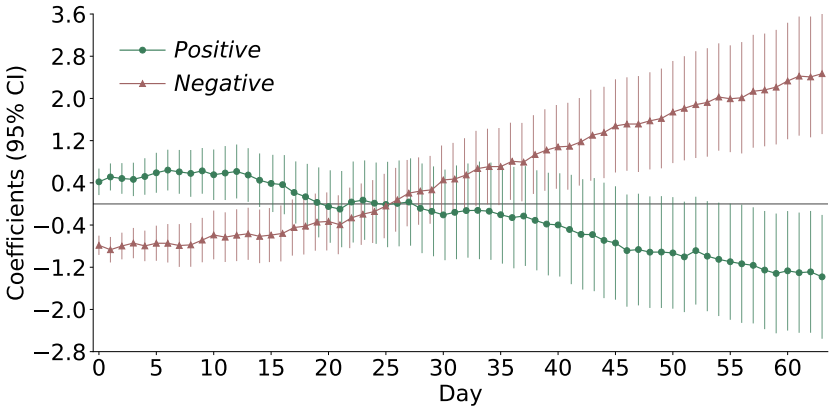
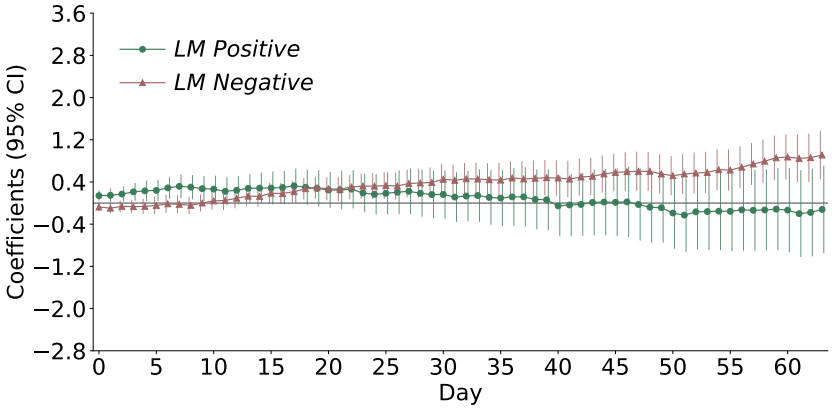


Figure 3: Rolling Correlation between Sentiment Classification and $LM_{0.25}$

This figure shows the rolling twelve-month correlation between Senti-GPT and $LM_{0.25}$ from January 1992 to December 2024.



(a) Senti-GPT indicators



(b) LM indicators

Figure 4: Media Sentiment and Stock Returns

This figure plots the estimated coefficients (95% confidence intervals) of genuine sentiment indicators in Panel (a) and LM indicators in Panel (b), where the dependent variable is the CAR as of days 0 to days 63. The same controls and fixed effects employed in Equation 10 are used. Standard errors are clustered by firm.

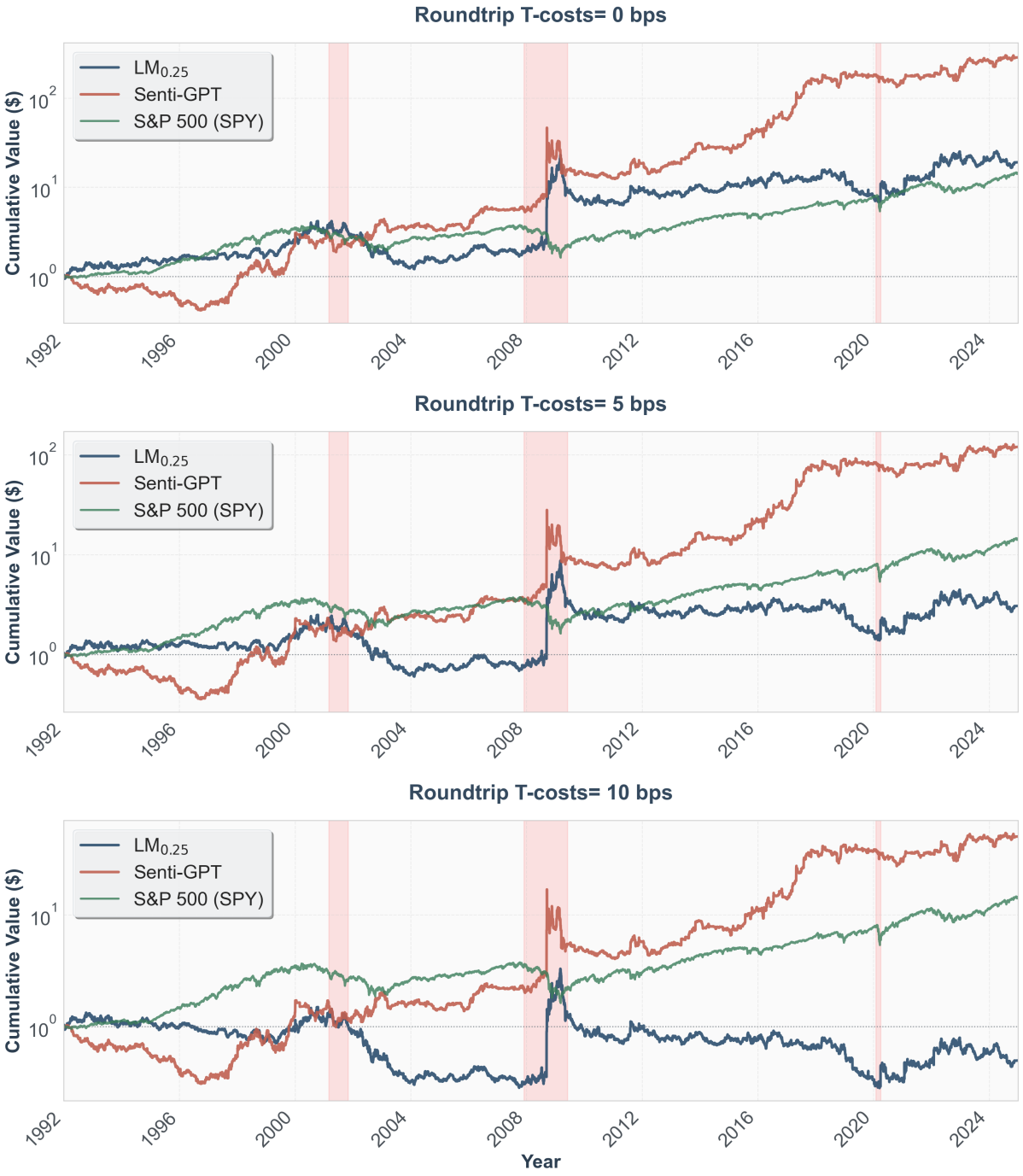


Figure 5: Comparison of Final Values Across Multiple Holding Periods

This figure illustrates the cumulative returns of three portfolios: the Senti-GPT portfolio, the $LM_{0.25}$ portfolio, and the SP 500 buy-and-hold portfolio (SPY). Each line represents cumulative final wealth starting from 1, spanning the period from January 1992 to December 2024. The Y-axis is log-scaled to facilitate comparison. The three subplots show performance under different transaction cost assumptions: 0 bps, 5 bps, and 10 bps per roundtrip, respectively. The shaded red regions indicate major recession periods: the dot-com crash (March-November 2001), the Global Financial Crisis (December 2007-June 2009), and the COVID-19 pandemic (February-April 2020).

Table 1: Confusion Matrix

	Pred -1	**Pred** 0	**Pred** 1
Actual -1	$N_{-1,-1}$	$N_{-1,0}$	$N_{-1,1}$
Actual 0	$N_{0,-1}$	$N_{0,0}$	$N_{0,1}$
Actual 1	$N_{1,-1}$	$N_{1,0}$	$N_{1,1}$

This table presents the confusion matrix for a three-sentiment classifier.

Table 2: Zero-shot Model Evaluation Based on Libra

A. Balanced posts	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 Score	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*Accuracy*
Sentiment-GPT	**85.11**	**85.11**	85.11	**91.08**	90.43	**91.75**	**95.35**	**96.06**	**94.66**	**90.67**
GPT-5	63.82	82.91	51.87	86.94	81.43	93.24	84.51	77.64	92.72	80.17
GPT-4.1	65.32	62.32	68.62	77.38	88.20	68.93	84.93	80.17	90.29	76.17
GPT-4o	65.08	64.74	65.43	77.60	88.75	68.93	83.33	76.00	92.23	75.83
GPT-4-turbo	60.38	61.20	59.57	78.24	90.45	68.93	78.54	70.38	88.83	72.83
FinBERT-tone	49.12	33.87	89.36	36.93	65.43	25.73	14.85	73.91	8.25	39.67
N	200	200	200	200	200	200	200	200	200	600

B. Ordinary posts	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 Score	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*Accuracy*
Sentiment-GPT	**95.38**	**97.89**	**93.00**	**97.06**	**95.19**	**99.00**	**98.51**	**98.02**	**99.00**	**97.00**
GPT-5	75.00	88.00	65.35	89.10	83.93	94.95	88.26	83.19	94.00	84.67
GPT-4.1	76.19	72.73	80.00	82.61	90.48	76.00	90.29	87.74	93.00	83.00
GPT-4o	78.39	78.79	78.00	86.17	92.05	81.00	87.32	82.30	93.00	84.00
GPT-4-turbo	72.45	73.96	71.00	82.98	88.64	78.00	86.11	80.17	93.00	80.67
FinBERT-tone	50.00	35.42	85.00	42.18	65.96	31.00	17.70	76.92	10.00	42.00
N	100	100	100	100	100	100	100	100	100	300

C. Complicated posts	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 Score	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*Accuracy*
Sentiment-GPT	**74.03**	**72.04**	76.14	**85.31**	85.71	**84.91**	**92.31**	**94.12**	90.57	**84.33**
GPT-5	48.44	73.81	36.05	84.98	79.20	91.67	81.17	72.93	91.51	75.67
GPT-4.1	52.97	50.52	55.68	72.13	85.71	62.26	80.17	73.81	87.74	69.33
GPT-4o	50.28	49.45	51.14	68.54	84.72	57.55	79.84	70.80	91.51	67.67
GPT-4-turbo	46.86	47.13	46.59	73.14	92.75	60.38	72.00	62.50	84.91	65.00
FinBERT-tone	48.26	32.42	94.32	31.43	64.71	20.75	12.07	70.00	6.60	37.33
N	100	100	100	100	100	100	100	100	100	300

This table reports the zero-shot textual sentiment extraction performance (%) of Sentiment-GPT and state-of-the-art LLMs based on Libra. Panels A, B, and C present results for balanced posts, ordinary posts, and complicated posts, respectively. Ordinary posts and complicated posts are complementary subsets of balanced posts. Performance metrics are defined in Equations 1 to 7. *Accuracy* is calculated as the number of correctly classified posts divided by the total number of posts in the testing sample. *Precision* represents the proportion of retrieved instances that are relevant, while *Recall* captures the proportion of relevant instances correctly retrieved. The *F_1 Score* is the harmonic mean of *Precision* and *Recall*. N denotes the sample size. Bolded scores indicate that Sentiment-GPT outperforms other models for that metric.

Table 3: Zero-shot Model Evaluation for Generalized Performance

A. Descriptive news and non-targeted commentaries		
	Descriptive news	Non-targeted commentaries
	Accuracy (%)	*Accuracy* (%)
Sentiment-GPT	**92.00**	**89.33**
GPT-5	86.00	84.67
GPT-4.1	84.67	82.67
GPT-4o	80.67	81.33
GPT-4-turbo	82.67	77.33
FinBERT-tone	79.33	80.67
N	150	150

(*continued*)

B. English-translated Libra

	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 Score	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*Accuracy*
Sentiment-GPT	**82.78**	**86.13**	**79.68**	**91.92**	88.05	**96.14**	**92.38**	**93.53**	**91.26**	**89.33**
GPT-5	65.20	78.79	55.61	86.44	82.46	90.82	85.65	79.58	92.72	80.50
GPT-4.1	67.59	70.11	65.24	85.57	88.21	83.09	84.21	79.65	89.32	79.67
GPT-4o	71.36	62.06	83.96	78.99	94.00	68.12	85.86	87.82	83.98	78.50
GPT-4-turbo	71.30	62.86	82.35	77.94	95.77	65.70	87.56	86.32	88.83	78.83
FinBERT-tone	49.12	36.55	74.87	46.11	64.91	35.75	40.13	60.19	30.10	46.00
N	200	200	200	200	200	200	200	200	200	600

C. Social media posts

	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 Score	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*Accuracy*
Sentiment-GPT	**84.21**	**78.05**	91.43	**89.86**	91.18	**88.57**	**89.23**	**96.67**	82.86	**87.62**
GPT-5	77.14	77.14	77.14	85.31	83.56	87.14	86.13	88.06	84.29	82.86
GPT-4.1	73.08	66.28	81.43	82.71	87.30	78.57	82.44	88.52	77.14	79.05
GPT-4o	70.21	55.93	94.29	76.27	93.75	64.29	73.68	95.45	60.00	72.86
GPT-4-turbo	72.32	59.81	91.43	74.34	97.67	60.00	80.00	86.67	74.29	75.24
FinBERT-tone	53.91	38.75	88.57	38.38	65.52	27.14	30.77	66.67	20.00	45.24
N	70	70	70	70	70	70	70	70	70	210

D. Long-form media articles

	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 Score	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*F_1 Score*	*Precision*	*Recall*	*Accuracy*
Sentiment-GPT	**81.58**	**100.00**	68.89	**91.67**	86.27	**97.78**	**91.84**	84.91	**100.00**	**88.89**
GPT-5	74.70	81.58	68.89	90.32	87.50	93.93	87.23	83.67	91.11	84.44
GPT-4.1	58.12	47.22	75.56	64.86	82.76	53.33	70.89	82.35	62.22	63.70
GPT-4o	67.19	51.81	95.56	72.22	96.30	57.78	68.57	96.00	53.33	68.89
GPT-4-turbo	64.96	52.78	84.44	71.05	87.10	60.00	75.32	90.63	64.44	69.63
FinBERT-tone	52.17	36.21	93.33	22.64	75.00	13.33	32.14	81.82	20.00	42.22
N	45	45	45	45	45	45	45	45	45	135

This table reports the zero-shot textual sentiment extraction performance (%) of Sentiment-GPT and state-of-the-art LLMs for generalized capabilities. In Panel A, we specifically test whether the models can follow our definition by assessing their performance on descriptive news and non-targeted commentaries. The tests in Panels B, C, and D are based on the English translation of Libra, social media posts, and long-form media articles, respectively. Performance metrics are defined in Equations 1 to 7. *Accuracy* is calculated as the number of correctly classified posts divided by the total number of posts in the testing sample. *Precision* represents the proportion of retrieved instances that are relevant, while *Recall* captures the proportion of relevant instances correctly retrieved. The *F_1 Score* is the harmonic mean of *Precision* and *Recall*. N denotes the sample size. Bolded scores indicate that Sentiment-GPT outperforms other models for that metric.

Table 4: Comparison of Sentiment Analysis Methods

Article	Narrative	Attitude	P_{count}	N_{count}	$\text{LM}_{0.25}$	Senti-GPT
1	-1	None	1	16	-1	0
2	1	None	18	0	1	0
3	-1	1	9	17	-1	1
4	1	-1	18	7	1	-1
5	1 & -1	-1	15	14	0	-1
6	1 & -1	1	11	14	0	1
7	-1	-1	3	17	-1	-1
8	1	1	20	2	1	1

This table summarizes the characteristics used to create each article and their corresponding sentiment assessments. The "Narrative" column indicates the use of negative (-1) or positive (1) words in generating content related to company fundamentals, news, and financial disclosures. The "Attitude" column represents whether the article includes explicit negative (-1) or positive (1) opinions or emotions. P_{count} and N_{count} represent the word counts of positive and negative words used in the article, respectively. $\text{LM}_{0.25}$ is the LM indicator using $\tau = 0.25$. Lastly, the Sentiment-GPT column shows the sentiment classification generated by our Sentiment GPT.

Table 5: Publication-period Abnormal Returns and Sentiment

A. Baseline results						
	Dependent Variable: *CAR (0, +2)*					
	(1)	(2)	(3)	(4)	(5)	(6)
Positive	0.452***			0.470***		
	(0.143)			(0.153)		
Negative	-0.878***			-0.837***		
	(0.121)			(0.132)		
LM Positive		0.124*			0.164**	
		(0.068)			(0.070)	
LM Negative		-0.082*			-0.060	
		(0.049)			(0.058)	
$LM\ Positive^{fraction}$			0.027			0.029
			(0.022)			(0.023)
$LM\ Negative^{fraction}$			-0.059***			-0.056***
			(0.021)			(0.020)
Obs.	37,681	37,681	37,681	36,585	36,585	36,585
Adj. R^2	0.005	0.001	0.001	0.025	0.022	0.022
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	No	No	No	Yes	Yes	Yes
Industry FE	No	No	No	Yes	Yes	Yes
Source FE	No	No	No	Yes	Yes	Yes

(*continued*)

B. Explanatory power of LM measures

	Dependent Variable: *CAR (0, +2)*			
	(1)	(2)	(3)	(4)
Positive	0.439***	0.435***	0.451***	0.454***
	(0.144)	(0.141)	(0.153)	(0.152)
Negative	-0.872***	-0.867***	-0.830***	-0.825**
	(0.122)	(0.121)	(0.132)	(0.132)
LM Positive	0.078		0.120*	
	(0.070)		(0.072)	
LM Negative	0.011		0.024	
	(0.047)		(0.056)	
$LM\ Positive^{fraction}$		0.014		0.011
		(0.021)		(0.022)
$LM\ Negative^{fraction}$		-0.018		-0.020
		(0.019)		(0.019)
Obs.	37,681	37,681	36,585	36,585
Adj. R^2	0.005	0.005	0.025	0.025
Time FE	No	No	Yes	Yes
Industry FE	No	No	Yes	Yes
Source FE	No	No	Yes	Yes

This table presents estimates of OLS regressions of the cumulative abnormal return (CAR) in percentages over days 0 to +2 around the commentary publication date. The CAR is in calculated based on the three Fama and French factors and momentum. Panel A shows baseline results, while results in Panel B include both sentiment indicators and LM measures in each test. $Positive_{i,t}$ and $Negative_{i,t}$ are set to 1 (and the other to 0) if firm i's average Senti-GPT on day t is positive or negative, respectively. $LM\ Positive_{i,t}$ ($LM\ Negative_{i,t}$) is an indicator variable equal to 1 if the aggregated $LM_{0.25}$ across all articles for firm i at day t is 1 (-1). $LM\ Positive_{i,t}^{fraction}$ ($LM\ Negative_{i,t}^{fraction}$) refers to the fraction of positive (negative) LM words across all commentaries for firm i at day t, standardized to unit variance. Control variables include lagged cumulative abnormal returns (CAR over days -30 to -6 and CAR over days -5 to -1), lagged abnormal volume (the average trading volume for the week preceding the event, standardized based on the 60 days before the filing date), and lagged volatility (the standard deviation of returns over the past 60 trading days). Day fixed effects, industry fixed effects, and article source fixed effects are included. Standard errors are double clustered at the firm and day level. *, **, and *** indicates significance at the 10%, 5%, and 1% level, respectively.

Table 6: Publication-period Abnormal Trading Volume and Sentiment

A. Baseline results						
	Dependent Variable: *AVOL (0, +2)*					
	(1)	(2)	(3)	(4)	(5)	(6)
Positive	0.225***			0.207***		
	(0.037)			(0.036)		
Negative	0.331***			0.297***		
	(0.035)			(0.036)		
LM Positive		0.007			-0.022	
		(0.019)			(0.021)	
LM Negative		-0.005			-0.016	
		(0.016)			(0.015)	
$LM\ Positive^{fraction}$			0.024***			0.006
			(0.006)			(0.007)
$LM\ Negative^{fraction}$			-0.000			-0.004
			(0.006)			(0.006)
Obs.	37,681	37,681	37,681	36,585	36,585	36,585
Adj. R^2	0.084	0.077	0.077	0.180	0.175	0.175
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	No	No	No	Yes	Yes	Yes
Industry FE	No	No	No	Yes	Yes	Yes
Source FE	No	No	No	Yes	Yes	Yes

(*continued*)

B. Explanatory power of LM measures

	Dependent Variable: *AVOL (0, +2)*			
	(1)	(2)	(3)	(4)
Positive	0.215***	0.205***	0.202***	0.201***
	(0.036)	(0.035)	(0.036)	(0.036)
Negative	0.340***	0.338***	0.305***	0.305***
	(0.036)	(0.035)	(0.037)	(0.037)
LM Positive	0.007		-0.022	
	(0.019)		(0.021)	
LM Negative	-0.028*		-0.036**	
	(0.016)		(0.015)	
$LM\ Positive^{fraction}$		0.020***		0.002
		(0.006)		(0.007)
$LM\ Negative^{fraction}$		-0.012**		-0.014**
		(0.005)		(0.006)
Obs.	37,681	37,681	36,585	36,585
Adj. R^2	0.084	0.084	0.180	0.180
Time FE	No	No	Yes	Yes
Industry FE	No	No	Yes	Yes
Source FE	No	No	Yes	Yes

This table presents estimates of OLS regressions of the abnormal trading volume (AVOL) for the 3-day publication event period, where volume is standardized based on the 60 days prior to the commentary publication date. The main explanatory variables of interest are Sentiment and LM measures. $Positive_{i,t}$ and $Negative_{i,t}$ are set to 1 (and the other to 0) if firm i's average Senti-GPT on day t is positive or negative, respectively. $LM\ Positive_{i,t}$ ($LM\ Negative_{i,t}$) is an indicator variable equal to 1 if the aggregated $LM_{0.25}$ across all articles for firm i at day t is 1 (-1). $LM\ Positive_{i,t}^{fraction}$ ($LM\ Negative_{i,t}^{fraction}$) refers to the fraction of positive (negative) LM words across all commentaries for firm i at day t, standardized to unit variance. Control variables include lagged cumulative abnormal returns (CAR over days -30 to -6 and CAR over days -5 to -1), lagged abnormal volume (the average trading volume for the week preceding the event, standardized based on the 60 days before the filing date), and lagged volatility (the standard deviation of returns over the past 60 trading days). Day fixed effects, industry fixed effects, and article source fixed effects are included. Standard errors are double clustered at the firm and day level. *, **, and *** indicates significance at the 10%, 5%, and 1% level, respectively.

Table 7: Portfolio Performance Metrics Across Different Time Periods

	2001-2024		2009-2024		1992-2024	
	Senti-GPT	$LM_{0.25}$	Senti-GPT	$LM_{0.25}$	Senti-GPT	$LM_{0.25}$
Port Ret	0.280	0.162	0.206	0.089	0.244	0.161
Port Std	0.477	0.482	0.301	0.359	0.432	0.428
Port Sharpe	0.587	0.336	0.687	0.247	0.564	0.376
$1 Port FV	108.916	5.832	13.351	1.563	289.495	19.18
Long-Only Sharpe	0.498	-0.035	0.475	0.128	0.406	0.239
Short-Only Sharpe	0.571	0.417	0.703	0.282	0.571	0.391

This table presents the backtesting reslts that builds portfolios based on sentiment classifications derived from Senti-GPT and $LM_{0.25}$ for various time periods. All portfolios are euqally weighted. All reported returns, standard deviations, and Sharpe ratios are annualized.

Table 8: Performance Metrics Across Different Sentiment Thresholds

	$LM_{0.1}$	$LM_{0.3}$	$LM_{0.5}$	$LM_{0.7}$	Senti-GPT
Port Ret	0.141	0.175	0.158	0.089	0.244
Port Std	0.346	0.430	0.438	0.416	0.432
Port Sharpe	0.409	0.406	0.361	0.214	0.564
$1 Port FV	17.654	29.183	14.898	1.936	289.495
Long-Only Sharpe	0.197	0.298	0.260	0.316	0.406
Short-Only Sharpe	0.430	0.398	0.321	0.181	0.571

This table compares the performance of sentiment-based trading strategies using different τ values (ranging from 0.1 to 0.7) against the Senti-GPT strategy. The backtest covers the full sample period from 1992 to 2024. All performance metrics are annualized. The $1 Final Value (FV) indicates the terminal value of an initial $1 investment.

Table 9: Fama-French Five-Factor Regression Results

	Coefficient	t-statistic	p-value
const	2.457	3.367	0.001
Mkt-RF	-0.630	-3.482	0.001
SMB	0.186	0.725	0.469
HML	0.092	0.310	0.757
RMW	0.023	0.070	0.945
CMA	-0.113	-0.254	0.800
R^2: 0.036, Adj. R^2: 0.024, F: 2.935 (p = 0.013), N = 396			

This table presents the results of the Fama-French five-factor regression analysis for our sentiment-based investment strategy. The regression examines the strategy's excess returns against five systematic risk factors over the period from January 1992 to December 2024. $R_{M,t} - R_{f,t}$ captures the market risk premium; SMB_t (Small Minus Big) represents the size premium between small and large-cap stocks; HML_t (High Minus Low) measures the value premium between high and low book-to-market stocks; RMW_t (Robust Minus Weak) captures the profitability premium between firms with robust and weak operating profitability; and CMA_t (Conservative Minus Aggressive) represents the investment premium between firms with conservative and aggressive investment patterns. The estimated annualized alpha is 29.49 percent.

Appendix A A Test for Look-ahead Bias

Table A1: Test for Look-ahead Bias Based on Libra

A. Balanced posts										
	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 *Score*	*Precision*	*Recall*	F_1 *Score*	*Precision*	*Recall*	F_1 *Score*	*Precision*	*Recall*	*Accuracy*
Qwen2.5-72B	60.59	67.32	55.08	83.90	84.73	83.09	81.45	80.86	82.04	74.00
N	200	200	200	200	200	200	200	200	200	600
B. Ordinary posts										
	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 *Score*	*Precision*	*Recall*	F_1 *Score*	*Precision*	*Recall*	F_1 *Score*	*Precision*	*Recall*	*Accuracy*
Qwen2.5-72B	78.31	100.00	64.36	92.68	89.62	95.96	89.42	86.11	93.00	84.33
N	100	100	100	100	100	100	100	100	100	300
C. Complicated posts										
	Neutral (%)			Positive (%)			Negative (%)			Overall (%)
	F_1 *Score*	*Precision*	*Recall*	F_1 *Score*	*Precision*	*Recall*	F_1 *Score*	*Precision*	*Recall*	*Accuracy*
Qwen2.5-72B	43.68	43.18	44.19	75.12	79.38	71.30	73.43	75.24	71.70	63.67
N	100	100	100	100	100	100	100	100	100	300

This table reports the zero-shot textual sentiment extraction performance (%) of Qwen2.5-72B based on Libra. Panels A, B, and C present results for balanced posts, ordinary posts, and complicated posts, respectively. Ordinary posts and complicated posts are complementary subsets of balanced posts. Performance metrics are defined in Equations 1 to 7. *Accuracy* is calculated as the number of correctly classified posts divided by the total number of posts in the testing sample. *Precision* represents the proportion of retrieved instances that are relevant, while *Recall* captures the proportion of relevant instances correctly retrieved. The F_1 *Score* is the harmonic mean of *Precision* and *Recall*. N denotes the sample size.

Appendix B Robustness Checks

Appendix B.1 Additional Figures

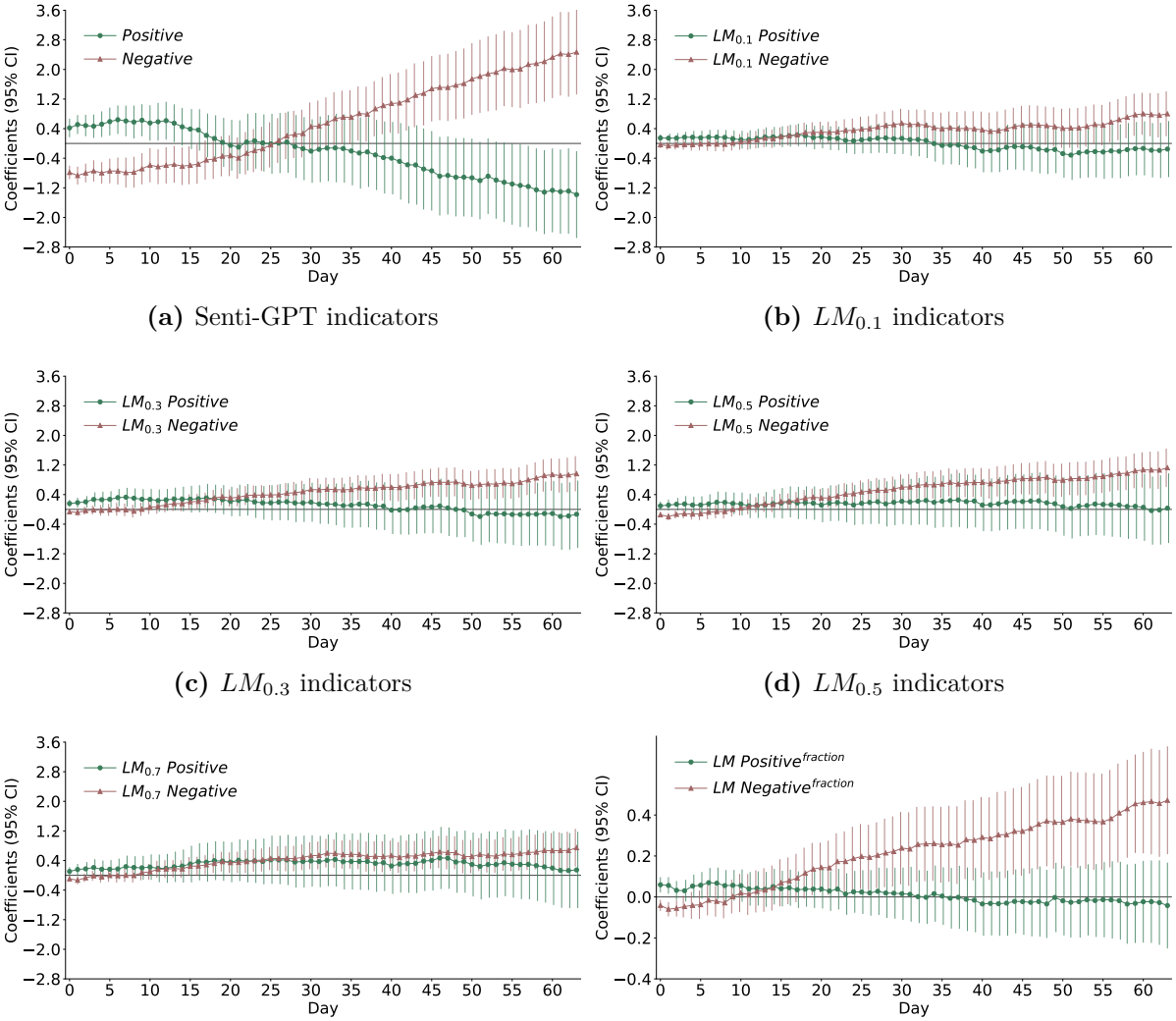


Figure B.1: Robustness Check: Post-publication Abnormal Return and Sentiment

This figure plots the estimated coefficients (95% confidence intervals) of genuine sentiment indicators in Panel (a), $LM_{0.1}$ indicators in Panel (b), $LM_{0.3}$ indicators in Panel (c), $LM_{0.5}$ indicators in Panel (d), $LM_{0.7}$ indicators in Panel (e), and the fraction of LM words in Panel (f), where the dependent variable is the CAR as of days 0 to days 63. The same controls and fixed effects employed in Equation 10 are used. Standard errors are clustered by firm.

Appendix B.2 Additional Tables

Table B1: Robustness Check: Publication-period Abnormal Returns and Sentiment

A. Two trading days					
	Dependent Variable: *CAR (0, +1)*				
	(1)	(2)	(3)	(4)	(5)
Positive	0.501***			0.479***	0.464***
	(0.135)			(0.135)	(0.133)
Negative	-0.880***			-0.867***	-0.864***
	(0.121)			(0.121)	(0.122)
$LM_{0.25}$ *Positive*		0.144**		0.097	
		(0.063)		(0.064)	
$LM_{0.25}$ *Negative*		-0.096*		-0.008	
		(0.050)		(0.048)	
$LM\ Positive_{fraction}$			0.055***		0.037*
			(0.020)		(0.019)
$LM\ Negative_{fraction}$			-0.060***		-0.022
			(0.017)		(0.017)
Obs.	36,585	36,585	36,585	36,585	36,585
Adj. R^2	0.022	0.017	0.017	0.022	0.022
Controls	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Source FE	Yes	Yes	Yes	Yes	Yes

(*continued*)

B. Four trading days

	Dependent Variable: *CAR (0, +3)*				
	(1)	(2)	(3)	(4)	(5)
Positive	0.452***			0.421***	0.441***
	(0.161)			(0.162)	(0.161)
Negative	-0.777***			-0.761***	-0.771***
	(0.148)			(0.146)	(0.147)
$LM_{0.25}$ *Positive*		0.202**		0.161*	
		(0.084)		(0.085)	
$LM_{0.25}$ *Negative*		-0.068		0.009	
		(0.063)		(0.061)	
$LM\ Positive_{fraction}$			0.027		0.010
			(0.027)		(0.026)
$LM\ Negative_{fraction}$			-0.043*		-0.009
			(0.025)		(0.024)
Obs.	36,585	36,585	36,585	36,585	36,585
Adj. R^2	0.026	0.024	0.025	0.026	0.026
Controls	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Source FE	Yes	Yes	Yes	Yes	Yes

(*continued*)

C. LM indicators with different thresholds

	Dependent Variable: *CAR (0, +2)*								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Positive	0.470***					0.452***	0.454***	0.447***	0.462***
	(0.153)					(0.154)	(0.153)	(0.153)	(0.153)
Negative	-0.837***					-0.833***	-0.833***	-0.819***	-0.830***
	(0.132)					(0.132)	(0.131)	(0.131)	(0.132)
$LM_{0.1}$ *Positive*		0.145**				0.107			
		(0.067)				(0.068)			
$LM_{0.1}$ *Negative*		-0.036				0.044			
		(0.062)				(0.060)			
$LM_{0.3}$ *Positive*			0.205***				0.158**		
			(0.075)				(0.076)		
$LM_{0.3}$ *Negative*			-0.034				0.049		
			(0.056)				(0.054)		
$LM_{0.5}$ *Positive*				0.141*				0.086	
				(0.083)				(0.084)	
$LM_{0.5}$ *Negative*				-0.126**				-0.048	
				(0.055)				(0.052)	
$LM_{0.7}$ *Positive*					0.173*				0.113
					(0.090)				(0.091)
$LM_{0.7}$ *Negative*					-0.065				-0.003
					(0.059)				(0.057)
Obs.	36,585	36,585	36,585	36,585	36,585	36,585	36,585	36,585	36,585
Adj. R^2	0.025	0.022	0.022	0.022	0.022	0.025	0.026	0.025	0.025
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Source FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

This table presents estimates of OLS regressions of the cumulative abnormal return (CAR) in percentages over days 0 to +1 (Panel A) and days 0 to +3 (Panel B), respectively, around the commentary publication date. Panel C reports estimates from the same specifications as in Table 5, with LM indicators constructed based on different thresholds. The CAR is in calculated based on the three Fama and French factors and momentum. $Positive_{i,t}$ and $Negative_{i,t}$ are set to 1 (and the other to 0) if firm i's average Senti-GPT on day t is positive or negative, respectively. LM_τ $Positive_{i,t}$ (LM_τ $Negative_{i,t}$) is an indicator variable equal to 1 if the aggregated LM_τ across all articles for firm i at day t is 1 (-1). $LM\ Positive_{i,t}^{fraction}$ ($LM\ Negative_{i,t}^{fraction}$) refers to the fraction of positive (negative) LM words across all commentaries for firm i at day t, standardized to unit variance. Control variables include lagged cumulative abnormal returns (CAR over days -30 to -6 and CAR over days -5 to -1), lagged abnormal volume (the average trading volume for the week preceding the event, standardized based on the 60 days before the filing date), and lagged volatility (the standard deviation of returns over the past 60 trading days). Day fixed effects, industry fixed effects, and article source fixed effects are included. Standard errors are double clustered at the firm and day level. *, **, and *** indicates significance at the 10%, 5%, and 1% level, respectively.

Table B2: Robustness Check: Publication-period Abnormal Volume and Sentiment

A. Two trading days					
	Dependent Variable: *AVOL (0, +1)*				
	(1)	(2)	(3)	(4)	(5)
Positive	0.292***			0.285***	0.281***
	(0.047)			(0.046)	(0.046)
Negative	0.422***			0.431***	0.432***
	(0.045)			(0.046)	(0.046)
$LM_{0.25}$ *Positive*		-0.014		-0.015	
		(0.024)		(0.024)	
$LM_{0.25}$ *Negative*		-0.010		-0.038**	
		(0.017)		(0.017)	
$LM\ Positive_{fraction}$			0.011		0.006
			(0.008)		(0.008)
$LM\ Negative_{fraction}$			-0.004		-0.018**
			(0.007)		(0.007)
Obs.	36,585	36,585	36,585	36,585	36,585
Adj. R^2	0.170	0.163	0.163	0.170	0.170
Controls	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Source FE	Yes	Yes	Yes	Yes	Yes
B. Four trading days					
	Dependent Variable: *AVOL (0, +3)*				
	(1)	(2)	(3)	(4)	(5)
Positive	0.151***			0.149***	0.148***
	(0.032)			(0.032)	(0.031)
Negative	0.222***			0.227***	0.227***
	(0.030)			(0.030)	(0.030)
$LM_{0.25}$ *Positive*		-0.026		-0.026	
		(0.019)		(0.018)	
$LM_{0.25}$ *Negative*		-0.016		-0.031**	
		(0.013)		(0.013)	
$LM\ Positive_{fraction}$			0.002		0.000
			(0.006)		(0.006)
$LM\ Negative_{fraction}$			-0.004		-0.011**
			(0.005)		(0.005)
Obs.	36,585	36,585	36,585	36,585	36,585
Adj. R^2	0.184	0.180	0.180	0.184	0.184
Controls	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Source FE	Yes	Yes	Yes	Yes	Yes

(*continued*)

C. LM indicators with different thresholds

	Dependent Variable: *AVOL (0, +2)*								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Positive	0.207***					0.205***	0.203***	0.202***	0.206***
	(0.036)					(0.036)	(0.036)	(0.036)	(0.036)
Negative	0.297***					0.305***	0.302***	0.302***	0.299***
	(0.036)					(0.037)	(0.037)	(0.037)	(0.037)
$LM_{0.1}$ *Positive*		-0.028				-0.033			
		(0.021)				(0.020)			
$LM_{0.1}$ *Negative*		-0.024				-0.046***			
		(0.018)				(0.017)			
$LM_{0.3}$ *Positive*			-0.008				-0.006		
			(0.021)				(0.021)		
$LM_{0.3}$ *Negative*			-0.002				-0.021		
			(0.014)				(0.014)		
$LM_{0.5}$ *Positive*				-0.016				-0.010	
				(0.022)				(0.022)	
$LM_{0.5}$ *Negative*				-0.009				-0.025	
				(0.015)				(0.015)	
$LM_{0.7}$ *Positive*					-0.006				0.006
					(0.025)				(0.025)
$LM_{0.7}$ *Negative*					0.001				-0.008
					(0.016)				(0.016)
Obs.	36,585	36,585	36,585	36,585	36,585	36,585	36,585	36,585	36,585
Adj. R^2	0.180	0.175	0.175	0.175	0.175	0.180	0.180	0.180	0.180
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Source FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

This table presents estimates of OLS regressions of the abnormal trading volume (AVOL) over days 0 to +1 (Panel A) and days 0 to +3 (Panel B), respectively, around the commentary publication date. Panel C reports estimates from the same specifications as in Table 6, with LM indicators constructed based on different thresholds. Volume is standardized based on the 60 days prior to the commentary publication date. $Positive_{i,t}$ and $Negative_{i,t}$ are set to 1 (and the other to 0) if firm i's average Senti-GPT on day t is positive or negative, respectively. LM_τ $Positive_{i,t}$ (LM_τ $Negative_{i,t}$) is an indicator variable equal to 1 if the aggregated LM_τ across all articles for firm i at day t is 1 (-1). $LM\ Positive_{i,t}^{fraction}$ ($LM\ Negative_{i,t}^{fraction}$) refers to the fraction of positive (negative) LM words across all commentaries for firm i at day t, standardized to unit variance. Control variables include lagged cumulative abnormal returns (CAR over days -30 to -6 and CAR over days -5 to -1), lagged abnormal volume (the average trading volume for the week preceding the event, standardized based on the 60 days before the filing date), and lagged volatility (the standard deviation of returns over the past 60 trading days). Day fixed effects, industry fixed effects, and article source fixed effects are included. Standard errors are double clustered at the firm and day level. *, **, and *** indicates significance at the 10%, 5%, and 1% level, respectively.

Appendix C Additional Details of The Framework

Appendix C.1 How Our Approach Differs from Conventional Ones

In this section, we provide an additional comparison to show how sentiment extraction, integrated with our definition, differs from sentiments extracted using conventional methods. We consider a collection of annotated dataset that are widely used in the literature (e.g., see Wu et al., 2023), including "Financial Phrasebank" (FPB) , "Twitter Financial News Sentiment" (TFNS), "FiQA SA", and "News With GPT Instruction" (MWGI), FPB and FiQA SA are financial news sentiment analysis benchmarks where FiQA SA including additional social media headlines. TFNS is an annotated corpus of finance-related tweets. Meanwhile, MWGI is annotated by ChatGPT with prompts and explanations.

Table C1 presents a comparison of sentiment extractions between conventional corpora and those based on the definition. The column "Conventional Sentiment" refers to sentiment assigned based on current corpora, while the column "Classification" shows the sentiment after classification.[26] The column "Defined Sentiment" represents sentiment classified using our definition.

As can be seen, posts 1, 2, and 3 are typical factual descriptions, which reflect none of the author's own attitudes and therefore fit Subset 1—their sentiments, under our definition, are "neutral". In contrast, they are assigned as positive or negative because they include positive/negative words or appear to be "good news" or "bad news" based on conventional methods. Additionally, while post 4 shows a positive attitude, the attitude covers multiple assets, including: $SPY, $LABU, $BLUE, and $AMZN. Therefore, this post should be in Subset 2 and assigned as "neutral".

On the other hand, under the definition, post 5 presents a positive opinion on the upcoming earnings and should be in Subset 3. Post 6 contains both a negative opinion and a negative emotional tone and should be in Subset 4. Finally, while post 7 does not contain any "positive" words, which results in a neutral label by conventional methods, it shows an apparently positive emotional tone and a positive perspective from the author

[26] For the ease of presentation, we simplify terms such as "bullish" (or "bearish") and "strong positive" (or "strong negative") from the corpora to "positive" (or "negative").

and should be referred to Subset 5.

Table C1: Bias in Sentiment Extraction: Conventional Corpora vs. the Definition

Post	Data Source	Conventional Sentiment	Classification	Defined Sentiment
1) The tech giant cited the strong U.S. dollar, falling consumer demand and production slowdowns of PCs for its performance.	NWGI	negative	S1	neutral
2) The company 's order book stood at 1.5 bln euro $ 2.2 bln on September 30 , 2007 , up by 24.2 pct on the year , with international orders amounting to 365 mln euro $ 534.3 mln .	FPB	positive	S1	neutral
3) Pharmaceuticals group Orion Corp reported a fall in its third-quarter earnings that were hit by larger expenditures on R&D and marketing .	FPB	negative	S1	neutral
4) Added long at friday close $SPY $LABU $BLUE $AMZN.	FiQA SA	positive	S2	neutral
5) $AAPL Now I'm glad I got stopped out of my $113 weeklies when I did. Will re-enter next Monday for the #earnings run-up.	FiQA SA	neutral	S3	positive
6) 2020's Biggest Risks	TFNS	neutral	S4	negative
7) $MYOV here we go !!	TFNS	neutral	S5	positive

This table presents a comparison of sentiment extractions between conventional corpora and our definition. The column titled "Conventional Sentiment" refers to sentiment assigned based on current benchmarks, while the column labeled "Classification" shows the sentiment after classification. The column "Defined Sentiment" reports sentiment classifications derived from the definition.

Appendix D Posts from Eastmoney.com

A suitable source of raw textual data for constructing firm-level sentiment measures should adhere to several key criteria. Firstly, it is crucial that the data is firm-specific and presented in a consistent and organized manner. This structured approach is vital as it enables clear identification of the firms discussed in each post, providing essential context for annotators. While popular social media platforms like X (formerly Twitter), Seeking Alpha, TikTok (Douyin in mainland China), Reddit, Weibo, and WeChat have amassed

a vast repository of textual content.[27] However, the principle of "Dive into anything", as exemplified by Reddit, highlights the broad and unstructured nature of discussions on these platforms, which are not typically centered around specific securities. This diversity and scattered nature of topics complicate the task of categorizing and tracking content to consistently align with specific firms.

Furthermore, even with precise grouping of texts by corresponding firms, the availability of texts varies greatly across different entities due to differences in firm popularity. This disparity significantly challenges the ability to perform statistical analyses on panel data, as the uneven text distribution can skew results. Additionally, the popularity of the platform itself is crucial. It must serve as an "attention hub", attracting a large number of active content creators and consumers. This ensures that the sentiment expressed in the textual data encompasses a wide range of perspectives, accurately reflecting the broader public sentiment towards a firm and providing sufficient data volume for impactful analysis. For instance, the Snowball platform categorizes discussion content clearly,[28] facilitating easier navigation. However, it falls short in user engagement compared to larger social media platforms. The relatively low number of reads, comments, and likes on most posts on Snowball raises concerns about the strength and exposure of these discussions, potentially limiting their utility in sentiment analysis.

Finally, StockTwits warrants particular attention as a leading platform in finance-specific social media. It serves as a hub where users globally share portfolio updates, seek advice, and engage with fellow investors, linking their posts to specific stock tickers and voluntarily assigning them as "bearish" or "bullish" sentiments. This tagging system, while structuring data collection for individual firms, presents StockTwits as a potential natural corpus for analysis. However, there are notable limitations with using StockTwits as a source of textual data for our analysis. Firstly, neutral sentiment is a key classification that should not be omitted. If we only consider posts with a sentiment tag, the absence of a "neutral" sentiment tag may lead to a skewed representation of the overall picture. Posts

[27] Reddit is an online discussion forum where users can submit content, comment on posts, and vote on the popularity of submissions. Weibo is a Chinese microblogging platform similar to X (Twitter), which allows users to post short messages, share images and videos, and follow other users. WeChat is a mobile messaging and social media app in China that allows users to send text, voice, image, and video messages, as well as make payments, access mini-programs, and more. It has more than 1.2 billion monthly active users.

[28] Snowball is a Chinese online community forum that focuses on in-depth discussions about stocks, investing, and trading strategies.

without sentiment tags may include many neutral comments, and in reality, people do leave neutral comments. Therefore, utilizing only sentiment-tagged posts could result in an overrepresentation of extreme sentiments—either bullish or bearish. Secondly, even if we skip the self-classified sentiment system and aim to annotate all posts again by adding a "neutral" option, posts from StockTwits are still not suitable for a potential corpus. This is because, a large number of posts include multimedia elements, such as images, links, and forwards. These elements potentially reduce the proportion of information conveyed by text alone. For instance, many posts contain ironic, exaggerated, and humorous images. These images are not only attention-grabbing but also integral to conveying the authors' main ideas. Consequently, extracting sentiments from such posts based solely on text may lead to numerous misinterpretations.

We source the corpus, Andromeda, from Eastmoney.com. The dataset consists of a comprehensive collection of 252 million posts covering all listed firms in China, spanning from the inception date of each stock forum to the end of 2022. This level of completeness ensures that our analysis encompasses every available post related to firms of all sizes, industries, and styles on Eastmoney.com.

The annual distribution of total posts and the median number of posts per firm each year from 1997 to 2022. In the late 1990s, the total number of posts was relatively low, with approximately 3,771 posts in 1997. This figure gradually increased in the early 2000s, reaching 54,277 posts by 2007. However, a significant surge in post counts began around 2010, when firms collectively recorded 235,967 posts, coinciding with the expansion of online platforms and the proliferation of social media. The most pronounced growth occurred after 2015, peaking at an exceptional 64,246,958 posts in 2021. This rapid growth, particularly from 2017 onward, is likely due to the widespread adoption of mobile internet and the increasing integration of social media strategies into corporate communication. Although total posts slightly decreased to 55,634,513 in 2022, the overall upward trajectory over the 25-year period is clear.

Appendix D.1 Post Classification

This subsection presents examples of the post classification to ensure consistent annotation across the data.

Subset 1

Texts either not related to any individual asset or not reflecting the author's own expressions.

Example A.1.1.1: Analysts have pointed out that if CDRs are officially implemented, it would boost not only Chinese investors, listed firms, and the Chinese stock market but also the Chinese economy as a whole, potentially accelerating the maturation of China's capital market. It's been noted that related to this, major Chinese new economy firms listed in the US like Baidu, Tencent, and JD.com have all expressed interest and willingness to return to China's A-shares.
Sentiment: neutral.

Example A.1.1.2: Bullish on the market in long-term, decided to hold for the long haul.
Sentiment: neutral.

Subset 2

Texts related to an individual asset, articulated by the author but devoid of any opinions or emotional tones, such as inquiries about the asset.

Example A.1.2.1: Just wanna ask real quick, is today the last drop?
Sentiment: neutral.

Example A.1.2.2: Hasn't hit the limit down yet? Hurry up so I can start bottom fishing.
Sentiment: neutral.

Subset 3

Texts exclusively containing the author's opinions on an individual asset.

Example A.1.3.1: Planning to scoop up 10000-20000 shares of this stock soon. As for so-called value investments like Moutai, those are off the table for me. Sure, stocks don't rise without some speculation, but those so-called value stocks? Their bubble is way too big, downright scary. Even if I lose money on this stock, I'm not touching those big

bubbles. Of course, it'd be even better if this stock dips a bit more.

Sentiment: positive.

Example A.1.3.2: Time to buy the dip.

Sentiment: positive.

Subset 4

Texts exclusively reflecting the author's emotional tones towards an individual asset.

Example A.1.4.1: After all that hustle, it's only at the end the stock hits the limit up. Talk about underwhelming.

Sentiment: negative.

Example A.1.4.2: Midea's back in the green again...[29]

Sentiment: neutral.

Subset 5

Texts expressing the author's opinions and emotional tones towards an individual asset.

Example A.1.5.1: Breaking through the resistance level is tough!

Sentiment: neutral.

Example A.1.5.2: Up by a penny, let's clap!!!

Sentiment: negative.

Appendix D.2 Textual Unreadability

[29]We translate "red" in Chinese into "green", because "red" stands for increasing in Chinese stock market.

Appendix D.2.1 Unclear Expressions and Typos

Buying (Selling) does not directly imply a positive attitude.

Example A.2.1.1: I'm the one always buying booze in your live streams, picking up both Baiyu and Qingyu. Not a fan of yellow wine, but I've been getting used to it, consider it a form of support. I drink it, share it with others too. But this kind of stock manipulation really tarnishes the Guyue Longshan brand. Why do I still pick your stocks? Because I'm from Shaoxing.

Sentiment: negative.

Self-deprecation.

Example A.2.1.2: Tell a joke, bought at 2700 points, got trapped at 3400 points... [LOL]

Sentiment: negative.

Profanity, typo and jargon.

Example A.2.1.3: Big s(ell)cum, big sell-off.

Sentiment: negative.

Appendix D.2.2 Inconsistent multi-horizon sentiments

Example A.2.2.1: Don't catch falling knives, but I think an entry is coming soon on this.

Sentiment: negative.

Appendix E GPT-Generated Article

This section provides eight full-length GPT-generated articles discussing Apple, along with the prompts used to generate them. To facilitate comparison, each article includes sentiment classifications at the bottom, which are $LM_{0.25}$ and the sentiment classification generated by our Sentiment-GPT.

Prompt of Article 1

Negative Narrative with None Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using only the negative words provided in the file. [a]The content should focus solely on fundamentals, facts, figures (you can make up the appropriate numbers if necessary), or company news. However, do not directly express the author's personal views or attitudes anywhere in the text.

[a]The lexicon of Loughran & McDonald (2011) has be already attached to GPT before this instruction.

Main Body: Apple Inc. faces challenges as recent financial reports indicate a 5% decline in revenue due to disruptions in supply chains and increased competition in the tech sector. The company has experienced setbacks in its latest product launches, which have been met with mixed reviews and lower-than-expected sales figures. Regulatory pressures have also escalated, with antitrust investigations in multiple countries focusing on Apple's market practices, further complicating its operational landscape. Additionally, Apple's recent earnings have been impacted by significant fines for privacy breaches, contributing to a downturn in investor sentiment. The tech giant has acknowledged these obstacles in its latest communications, noting difficulties in meeting production targets for its newest iPhone model due to component shortages. Market analysts have expressed concerns over declining profit margins and the potential long-term effects of ongoing legal challenges on the company's financial health. Apple has also reported higher than usual employee turnover, attributing this to a demanding work environment and internal disagreements over strategic directions, which could hinder future innovation and growth. These factors collectively underscore the hurdles Apple currently faces in maintaining its market position and sustaining shareholder value in a rapidly evolving and increasingly competitive global market.

Article Sentiment: [$LM_{0.25}$: -1 vs. Senti-GPT: 0]

Prompt of Article 2

Positive Narrative with None Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using only the positive words provided in the file.[a] The content should focus solely

on fundamentals, facts, figures (you may make up appropriate numbers if necessary), or company news. However, do not directly express the author's personal opinions or emotions anywhere in the text.

[a]The lexicon of Loughran & McDonald (2011) has be already attached to GPT before this instruction.

Main Body: Apple Inc. continues to demonstrate financial robustness amid a dynamic tech landscape, with its latest quarterly earnings report revealing a 10% increase in revenue, largely driven by an uptick in global iPhone sales and an expanding services division. The company's innovative strategies have bolstered investor confidence, as seen in a 15% year-over-year rise in stock price. Apple's management has successfully leveraged strategic alliances and enhancements in technology to stay ahead of competitors, contributing to a favorable market position. Notably, the announcement of advancements in artificial intelligence and machine learning has sparked a surge in stock market performance, reflecting strong market optimism. Furthermore, Apple's commitment to efficiency is evident in its streamlined supply chain, which has effectively reduced costs and improved profit margins. The tech giant has also announced a breakthrough in battery technology, which is expected to significantly extend device longevity—an enhancement likely to boost consumer satisfaction and attract new customers. Collaborative efforts in environmental sustainability have also received acclaim, aligning with global movements towards eco-friendly business practices. These strategic advancements underscore Apple's ongoing commitment to innovation and market leadership, reinforcing its potential for sustained growth and profitability in the competitive tech industry.

Article Sentiment: [$LM_{0.25}$: 1 vs. Senti-GPT: 0]

Prompt of Article 3

Negative Narrative with Positive Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using only the negative words provided in the file.[a] The content should focus solely on fundamentals, facts, figures (you may make up appropriate numbers if necessary), or company news. Additionally, include positive personal opinions or emotions from the author in the text where appropriate.

[a]The lexicon of Loughran & McDonald (2011) has already been attached to GPT before this instruction.

Main Body: Despite the grim shadows cast by numerous challenges facing Apple Inc., I remain optimistic about the resilience and potential recovery of the company's stock price. Recent reports and data reflect a troubling period for Apple, marked by a 7% decline in smartphone sales, intensified by ongoing global supply chain disruptions which have delayed production and hindered the company's ability to fulfill orders swiftly. Adding to the woes, the technology giant has faced regulatory scrutiny that has resulted in hefty fines and threatens to disrupt its business model, particularly around the App Store which has been a significant revenue stream. Apple's recent foray into new markets has also been met with resistance, encountering strong competition from local providers entrenched within their regions. Furthermore, there have been concerns over data privacy issues that have damaged consumer trust, a critical component of Apple's brand equity. The company's issues are compounded by internal discord marked by high-profile exits from its executive team, signaling potential instability in its leadership. Despite these adverse conditions, Apple's history of navigating through tumultuous times with strategic adjustments and its ongoing investment in research and development provide a strong foundation for future success. The company's robust financial position, with substantial cash reserves, allows it flexibility in maneuvering through economic downturns and investing in new growth areas like artificial intelligence and augmented reality technologies. My positive outlook is based on Apple's proven track record of overcoming obstacles and emerging stronger. With its continued focus on innovation and expanding its product ecosystem, coupled with a loyal customer base, Apple is well-positioned to weather the current storms. The resilience of Apple's brand and its strategic initiatives aimed at resolving current challenges suggest that the current dip in stock price may present a buying opportunity for long-term investors who are willing to weather the storm alongside a company that has repeatedly demonstrated its capacity for reinvention and sustained growth.

Article Sentiment: [$LM_{0.25}$: -1 vs. Senti-GPT: 1]

Prompt of Article 4

Positive Narrative with Negative Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using only the positive words provided in the file.[a] The content should focus solely

on fundamentals, facts, figures (you may make up appropriate numbers if necessary), or company news. Additionally, include negative personal opinions or emotions from the author in the text where appropriate.

[a]The lexicon of Loughran & McDonald (2011) has already been attached to GPT before this instruction.

Main Body: Despite Apple Inc.'s impressive display of financial resilience and strategic innovation, I maintain a cautious stance on the company's stock price outlook due to broader market conditions and internal challenges that could potentially overshadow its achievements. Apple recently announced a record-breaking revenue increase of 15% from the previous quarter, driven by robust sales in its latest iPhone and services segments. The company also highlighted its success in advancing technological innovation, particularly through the development of AI-driven applications that have significantly enhanced user engagement and opened new revenue streams. Additionally, Apple has expanded its market reach with the successful launch of retail stores in key emerging markets, which has not only increased its global footprint but also strengthened its brand dominance. Moreover, the company's commitment to sustainability has been recognized with several awards, reflecting positively on its corporate responsibility initiatives. Apple's strategic alliances have also been pivotal in maintaining a competitive edge, fostering collaborations that have led to enhanced product offerings and market diversification. However, despite these positive developments, I remain skeptical about the company's ability to sustain this momentum in a volatile economic environment characterized by rising interest rates and geopolitical tensions. The tech giant's heavy reliance on consumer electronics makes it vulnerable to shifts in consumer spending, which is expected to decline as economic conditions tighten. Additionally, while Apple's advancements and expansions are commendable, they come at a time when the company faces increasing competition from rivals who are also stepping up their game in innovation and market penetration. In conclusion, while Apple continues to accomplish significant milestones and strengthen its market position, the looming economic challenges and fierce competition present substantial risks that could dampen its future stock performance. Investors should be wary and consider these factors when evaluating Apple's stock potential, as the external pressures might outweigh the internal strengths in the upcoming periods.

Article Sentiment: [$LM_{0.25}$: 1 vs. Senti-GPT: -1]

Prompt of Article 5

Negative & Positive Narrative with Negative Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using equally both the positive and negative words provided in the file.[a] The content should focus solely on fundamentals, facts, figures (you may make up appropriate numbers if necessary), or company news. Additionally, include negative personal opinions or emotions from the author in the text where appropriate.

[a]The lexicon of Loughran & McDonald (2011) has already been attached to GPT before this instruction.

Main Body: Apple Inc., the global technology leader, continues to demonstrate resilience despite facing a dynamic and competitive landscape. While recent quarterly results showed a moderated revenue growth of 3%, this performance underscores Apple's ability to sustain steady expansion even in a mature smartphone market. The company's unparalleled brand loyalty and ecosystem strength continue to drive stable demand for its products, with its installed base of active devices reaching an all-time high. Apple's innovation pipeline remains robust, with significant investments in next-generation technologies such as artificial intelligence, augmented reality, and silicon design—areas that are expected to unlock new growth opportunities. The upcoming launch of its highly anticipated mixed-reality headset and advancements in custom chips position Apple to redefine future computing paradigms. While global supply chain adjustments and regulatory discussions present near-term considerations, Apple's strong cash reserves ($166 billion in cash and marketable securities as of last quarter) and pricing power provide ample flexibility to navigate these challenges. The company's aggressive share buyback program and consistent dividend growth further highlight its commitment to delivering shareholder value. With its unmatched integration of hardware, software, and services, Apple is strategically pivoting toward higher-margin segments, including digital subscriptions and financial services. This diversification, coupled with its relentless focus on user privacy and security, reinforces Apple's long-term competitive advantage in an increasingly digital world. Despite its historical resilience and market dominance, I view Apple's near-term prospects with considerable skepticism. The combination of market saturation, regulatory challenges, and competitive pressures forms a daunting gauntlet that the company must navigate. Its once-unassailable position is showing signs of vulnerability, and

without a significant strategic pivot or groundbreaking innovation, Apple's stock could face downward pressure. Investors should be wary of the potential risks and consider the possibility that Apple's golden era of unbridled growth and exceptional returns may be waning. The coming quarters will be critical in determining whether Apple can regain its momentum or if it will succumb to the mounting challenges that threaten to overshadow its achievements.

Article Sentiment: [$LM_{0.25}$: 0 vs. Senti-GPT: -1]

Prompt of Article 6

Negative & Positive Narrative with Positive Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using equally both the positive and negative words provided in the file.[a] The content should focus solely on fundamentals, facts, figures (you may make up appropriate numbers if necessary), or company news. Additionally, include positive personal opinions or emotions from the author in the text where appropriate.

[a]The lexicon of Loughran & McDonald (2011) has already been attached to GPT before this instruction.

Main Body: Apple Inc., the tech behemoth known for its innovation and market dominance, is currently navigating a complex landscape fraught with challenges yet peppered with significant growth opportunities. Financially, Apple has shown resilience with a reported 8% increase in total revenue this quarter, driven largely by its diversified product lines including the latest launch of its eco-friendly MacBooks and expansion into digital health services. This growth has been bolstered by strategic alliances that have broadened its market reach and fortified its standing in emerging markets, showcasing the company's adeptness at capturing new consumer bases. However, the company has not been immune to the adversities plaguing the global tech industry. Apple has faced severe supply chain disruptions that have delayed product launches and constrained inventory levels, leading to a temporary dip in sales figures for its flagship iPhone series. Moreover, the ongoing global economic slowdown has curtailed consumer spending, directly impacting the affordability and sales of higher-priced consumer electronics. Regulatory challenges also loom large, with antitrust investigations in both the U.S. and Europe threatening to reshape its business operations and potentially limit its market power. Despite these headwinds, I maintain a positive outlook on Apple's stock price prospects.

The company's sustained investment in research and development is a testament to its commitment to innovation, particularly in areas like artificial intelligence and renewable energy, which are expected to be key drivers of future growth. Apple's robust financial position, marked by strong cash flows and a healthy balance sheet, provides it with the necessary resources to navigate current challenges and invest in growth opportunities. Furthermore, Apple's consistent track record of navigating through market downturns and emerging stronger positions it as a resilient investment for the long term. In conclusion, although Apple faces significant challenges in the form of supply chain issues and regulatory pressures, its strategic initiatives, ongoing product innovation, and strong financial health present a compelling case for potential investors. The company's ability to adapt and thrive in changing market conditions, combined with its aggressive push into new technological frontiers, reinforces my optimistic stance on its stock price trajectory, making it an attractive prospect for those looking to benefit from its future growth and industry leadership.

Article Sentiment: [$LM_{0.25}$: 0 vs. Senti-GPT: 1]

Prompt of Article 7

Negative Narrative with Negative Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using only negative words provided in the file.[a] The content should focus solely on fundamentals, facts, figures (you may make up appropriate numbers if necessary), or company news. Additionally, include negative personal opinions or emotions from the author in the text where appropriate.

[a]The lexicon of Loughran & McDonald (2011) has already been attached to GPT before this instruction.

Main Body: The outlook for Apple Inc.'s stock price is increasingly clouded by a series of challenges that cast a shadow over its future performance. Recent developments have exposed the tech giant to significant vulnerabilities, notably in its over-reliance on global supply chains that are currently snarled by persistent disruptions. These issues are not mere inconveniences but are seriously hindering the company's ability to meet product demand, especially for its latest iPhone models. Additionally, Apple is grappling with regulatory pressures that threaten hefty fines and stringent controls, particularly in the European Union where privacy laws are tightening. The situation is further exacerbated

by intensifying competition from rivals who are rapidly enhancing their technological offerings at lower costs, thereby putting severe pressure on Apple's market share and profit margins. Moreover, investor sentiment has been dampened by recent allegations of the company's involvement in restrictive trade practices, which could lead to legal battles and further financial liabilities. The company's once sterling reputation for innovation is now being overshadowed by these legal and operational risks, creating a climate of uncertainty and skepticism among shareholders. From a personal perspective, I view these accumulating issues with a degree of pessimism regarding Apple's short-term stock trajectory. The confluence of supply chain vulnerabilities, regulatory challenges, and fierce competition forms a potent threat to Apple's historical stock stability and growth prospects. Investors might find it prudent to brace for potential downturns and recalibrate their expectations as Apple navigates through these turbulent times. The company's next moves will be crucial in determining whether it can mitigate these risks or if it will succumb to the mounting pressures that currently beleaguer its operational and strategic landscape.

Article Sentiment: [$LM_{0.25}$: -1 vs. Senti-GPT: -1]

Prompt of Article 8

Positive Narrative with positive Personal Attitude

Generate a Wall Street Journal-style article discussing Apple's stock price outlook, using only positive words provided in the file.[a] The content should focus solely on fundamentals, facts, figures (you may make up appropriate numbers if necessary), or company news. Additionally, include positive personal opinions or emotions from the author in the text where appropriate.

[a]The lexicon of Loughran & McDonald (2011) has already been attached to GPT before this instruction.

Main Body: Apple Inc.'s stock outlook remains exceptionally promising, buoyed by a series of strategic successes and robust financial health. The company continues to demonstrate an admirable proficiency in innovation and market expansion, recently launching breakthrough products that have been met with critical acclaim and consumer enthusiasm. These innovations have not only enhanced Apple's product portfolio but have also fortified its competitive edge in the fiercely contested tech sector. Furthermore, Apple's strategic investments in renewable energy and environmental sustainability are

paying dividends, enhancing its corporate reputation and aligning with global shifts toward eco-friendly business practices. Financially, Apple boasts substantial profitability and a strong balance sheet, underscored by a recent 20% increase in quarterly revenues, largely attributed to its diversified revenue streams that include a growing services sector. The company's forward-thinking management team continues to excel in navigating market dynamics, ensuring steady growth and shareholder returns. Apple's commitment to enhancing shareholder value is evident in its consistent dividend payouts and effective stock buyback programs, which reflect its stable financial footing and optimistic future outlook. From a personal standpoint, I am highly optimistic about Apple's future stock performance. The company's continuous focus on innovation, coupled with its proactive approach to global market challenges and opportunities, positions it well for sustained growth. Apple's ability to stay ahead of technological trends and its unwavering commitment to customer satisfaction are pivotal in driving its stock upwards. For investors seeking a stable yet dynamic company with a clear vision for the future, Apple represents a compelling choice. Its track record of success and strategic direction suggest a bright and prosperous road ahead, making Apple's stock an attractive investment with promising prospects for appreciation.

Article Sentiment: [$LM_{0.25}$: 1 vs. Senti-GPT: 1]