Journal of
*Theoretical and Applied Electronic Commerce Research*

MDPI

*Article*

# Leveraging Large Language Models for Sentiment Analysis and Investment Strategy Development in Financial Markets

**Yejoon Mun and Namhyoung Kim ***

Department of Applied Statistics, Gachon University, 1342 Seongnam-daero, Sujung-gu, Seongnam 13120, Republic of Korea; ansd@gachon.ac.kr
* Correspondence: nhkim@gachon.ac.kr; Tel.: +82-31-750-5390

**Abstract:** This study investigates the application of large language models (LLMs) in sentiment analysis of financial news and their use in developing effective investment strategies. We conducted sentiment analysis on news articles related to the top 30 companies listed on Nasdaq using both discriminative models such as BERT and FinBERT, and generative models including Llama 3.1, Mistral, and Gemma 2. To enhance the robustness of the analysis, advanced prompting techniques—such as Chain of Thought (CoT), Super In-Context Learning (SuperICL), and Bootstrapping—were applied to generative LLMs. The results demonstrate that long strategies generally yield superior portfolio performance compared to short and long–short strategies. Notably, generative LLMs outperformed discriminative models in this context. We also found that the application of SuperICL to generative LLMs led to significant performance improvements, with further enhancements noted when both SuperICL and Bootstrapping were applied together. These findings highlight the profitability and stability of the proposed approach. Additionally, this study examines the explainability of LLMs by identifying critical data considerations and potential risks associated with their use. The research highlights the potential of integrating LLMs into financial strategy development to provide a data-driven foundation for informed decision-making in financial markets.

**Keywords:** large language models; sentiment analysis; prompt optimization; portfolio performance analysis; investment strategy



Academic Editors: Albert Y.S. Lam and Andy Chun

Received: 28 February 2025
Revised: 9 April 2025
Accepted: 16 April 2025
Published: 20 April 2025

**Citation:** Mun, Y.; Kim, N. Leveraging Large Language Models for Sentiment Analysis and Investment Strategy Development in Financial Markets. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 77. https://doi.org/10.3390/jtaer20020077



## 1. Introduction

Financial market information significantly influences asset price volatility. Unstructured text data from sources such as news articles and social media play a crucial role in shaping market dynamics through investor sentiment. Research findings indicate that pessimistic media coverage can lead to short-term declines in market prices and increases in trading volumes, highlighting the capacity of unstructured data to elucidate the relationship between investor sentiment and market volatility [1]. For example, on days when the name of a company is frequently mentioned in the news, a notable increase in the trading volume of its stock occurs. A Spearman correlation coefficient of 0.43 has been observed between the number of news mentions and absolute returns of the stock of that company, indicating a significant correlation [2]. These studies emphasize that news data transcend mere information delivery and exert a tangible influence on investor psychology and trading behavior.

As sentiment analysis techniques continue to evolve, the prediction of stock market movements using news data has become increasingly sophisticated. A growing body of research quantifies positive and negative sentiments in news articles and social media content to anticipate market reactions, subsequently leveraging these insights to develop investment strategies [3,4]. Specifically, positive sentiment is often associated with stock price increases, whereas negative sentiment is correlated with price declines. In addition, phenomena such as gradual stock price declines following adverse news or abrupt price swings without any publicly released news, —often followed by subsequent reversals, —have been documented. This highlights the pressing need for systematic research and the application of these relationships in financial contexts [5].

However, conducting financial sentiment analysis manually requires not only professional knowledge of finance but also a comprehensive understanding of the overall market and individual companies. For example, constructing a finance-specific sentiment analysis dataset involved extracting approximately 4840 sentences from financial news articles, which were then manually labeled as positive, neutral, or negative by 16 finance experts. This process required substantial time and costs for labeling and review [6].

Traditional machine learning algorithms necessitate this labeling process. In contrast, leveraging advanced large language models (LLMs) offers the significant advantage of eliminating labeling costs. When combined with automated sentiment analysis techniques, LLMs present a promising avenue for enhancing the efficiency and accuracy of financial sentiment analysis, ultimately contributing to more informed investment decisions. LLMs such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), and Llama exhibit exceptional contextual understanding capabilities, enabling them to capture the subtle nuances of sentiment in natural language effectively [7,8]. This proficiency not only enhances general sentiment analysis but also demonstrates strong performance, particularly in financial sentiment analysis. For instance, studies utilizing ChatGPT to analyze the sentiment of news headlines to predict stock returns [9] and those that combine news sentiment with macroeconomic indicators using BERT to forecast stock indices [10] clearly illustrate the value of LLMs as powerful tools for stock market prediction.

However, the large-scale analysis of news data using commercial LLMs (such as GPT-4, Claude 3, and Gemini 1.5) incurs significant costs, which can be a barrier to their widespread adoption despite their high accuracy. This cost burden increases substantially with the volume of analyzed data. To address these financial constraints, this study explores the use of open-source LLMs as an alternative. Although open-source LLMs may have limitations in terms of model size and performance compared with their commercial counterparts, which could potentially lead to lower investment performance in financial sentiment analysis [11], they offer a cost-effective solution.

Various prompting techniques, such as Chain-of-Thought (CoT), Super In-Context Learning (SuperICL), and Bootstrapping, were employed in the experiments to mitigate the limitations of open-source LLMs and enhance the financial sentiment analysis accuracy. This approach aims to maximize performance, even with relatively small model sizes, while effectively managing the complexities inherent in financial data, thereby improving the reliability of the sentiment analysis results.

Accordingly, this study addresses the following research questions:

1. Can open-source LLMs provide stability and profitability for portfolios based on financial sentiment analysis?

2. Can performance enhancement techniques such as CoT, SuperICL, and LLM Bootstrapping improve the portfolio performance of open-source LLMs in financial sentiment analysis?
3. How do the reliability and explainability of generative LLMs' analysis results contribute to the subsequent development of investment strategies and advancements in research?

This study examines the impact of LLM-based sentiment analysis on investment strategy performance in the US stock market. Portfolio strategies that reflect the characteristics of the US market are designed, and investment performance is evaluated using profitability and stability indicators. The research findings provide empirical evidence that LLM-based sentiment analysis can contribute significantly to the formulation of effective investment strategies in financial markets, facilitating the development of LLMs as reliable and practical tools for application in the financial sector.

This study offers the following key contributions: First, while previous research has predominantly focused on evaluating the performance of commercial generative LLMs [7–9], this study leverages open-source generative LLMs, enhancing accessibility while maintaining profitability. Second, in contrast to existing studies that have typically employed a single prompting technique [8,9,12,13], this study employs a diverse range of prompting techniques to maximize the value derived from sentiment analysis.

The remainder of this paper is organized as follows. Section 2 explores the development of key language models utilized in text data analysis and their significance in sentiment analysis. Specifically, it examines research on sentiment analysis within the financial domain, highlighting state-of-the-art performance enhancement techniques such as CoT, SuperICL, and Bootstrapping. Section 3 outlines the construction of portfolios using LLM-based sentiment analysis, detailing the various performance metrics and experimental methods employed for evaluation. Section 4 provides a comprehensive description of the data collection, preprocessing, and analysis procedures implemented in the study, ensuring transparency and reproducibility. Section 5 discusses the performance and effectiveness of sentiment analysis-based strategies, drawing insights from the analysis results to evaluate their practical implications. Section 6 investigates the interpretability of LLMs using techniques to elucidate sentiment predictions, emphasizing the importance of transparency and reliability in financial decision-making. Finally, Section 7 summarizes the main findings of the study and offers suggestions for future research, emphasizing the potential for further advancements in the field.

## 2. Related Works

### *2.1. Language Models*

Language models predict the continuity of words in text data and play a crucial role in natural language processing (NLP) tasks. Initially, statistical approaches based on frequency and conditional probabilities were employed to determine the relationships between words. However, these models exhibited limitations, such as low generalization performance when data were scarce and challenges related to the curse of dimensionality. Specifically, n-gram-based models were effective in short contexts but struggled with long contexts and the prediction of novel word combinations [14,15]. To address these shortcomings, the concepts of n-gram similarity and distance were proposed, which effectively quantified string similarity and, as special cases of edit distance and the longest common subsequence, outperformed traditional unigram methods. However, these approaches still faced difficulties in processing complex and lengthy contexts [16].

With the advent of machine learning, language models have evolved beyond traditional statistical methods. Tree-based models [17] and support vector machines (SVMs) [18] have been applied to tasks such as text classification, whereas deep learning models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [19] have been developed to maintain context over time and improve the prediction accuracy. However, RNN-based models struggled with long and complex contexts, leading to the development of the Transformer architecture [20]. The Transformer effectively learns from large-scale data through parallel processing and multi-layer attention mechanisms, laying the groundwork for LLMs such as BERT and GPT.

BERT is designed to understand the bidirectional context in text and has been used extensively in various NLP tasks, including sentiment analysis [21]. Through pre-training and fine-tuning, BERT learns the contextual information bidirectionally within a sentence, demonstrating high performance across a range of NLP applications.

In contrast, GPT is a unidirectional language model that sequentially generates text from left to right to understand and predict the context [22]. GPT-3, with approximately 175 billion parameters, exhibits remarkable versatility in performing new tasks with minimal examples [9]. Building on this foundation, GPT-4 has exhibited exceptional performance in processing multimodal inputs and handling complex question-answering tasks, achieving outstanding results in applications such as text generation, translation, and sentiment analysis. Furthermore, research has shown that while traditional machine learning models such as SVM excel in concise text classification, GPT-4 significantly outperforms these methods in context-rich sentiment analysis, offering detailed insights and superior F1-scores [23].

Recent advancements have highlighted the adaptability of GPT-based models in addressing domain-specific challenges. For instance, ElliottAgents integrates GPT with the Elliott wave principle and deep reinforcement learning to improve the interpretability and prediction accuracy of financial market trends [24]. FinSoSent, which is another application of GPT-based models, focuses on financial sentiment analysis by fine-tuning pretrained models on domain-specific data, achieving superior performance in analyzing financial social media content [25].

Furthermore, open-source LLMs provide researchers and developers with the flexibility to modify and apply models to specific tasks, including sentiment analysis. Llama 3, which was developed by Meta and is the latest version of the Llama series, is an open-source model that balances performance and efficiency. It is designed to deliver excellent performance with minimal resources, offering a more efficient research environment for language model studies. Notably, Llama 3 has shown improved results in various text generation and sentiment analysis tasks compared with existing models [10].

Mistral AI, which is a France-based AI research institute established in 2023, focuses on developing high-performance open-source models. Its flagship model, Mistral 7B, features 7.3 billion parameters and outperforms existing models such as Llama 2 13B. This model leverages the Grouped-Query Attention (GQA) and Sliding Window Attention (SWA) techniques to enhance the inference speed and handle long sequences efficiently [26].

Google has released Gemma 2, which is a lightweight, cutting-edge, open model product line based on the research and technology of the Gemini model. The Gemma models are available in versions with 2 billion and 7 billion parameters, including both pretrained models and instruction-tuned variants. These models exhibit excellent performance across various academic benchmarks, including language understanding, reasoning, and safety and have been developed to promote responsible AI use cases [27].

### 2.2. Sentiment Analysis

Sentiment analysis is a technique that is used to identify positive, negative, and neutral emotions in text data and is frequently applied in financial domains, such as the stock market to understand the market conditions and psychological trends of investors. Initially, financial experts manually read and classified the sentiments of text from sources such as news articles and social media posts. Although this method maintains high accuracy, it is limited in handling large volumes of data and may suffer from inconsistencies owing to the subjective nature of analyst interpretations. Consequently, cross-analyses by multiple professional analysts are often necessary to address these issues [6].

To overcome these limitations, researchers have turned to machine learning-based sentiment analysis, which is designed to process large amounts of text data more efficiently. Representative models in this domain include SVMs, which automatically classify text data as positive, negative, or neutral based on classifiers that are trained on financial datasets [28]. In addition, some studies have successfully predicted stock market volatility by analyzing investor sentiment in online forums. However, machine learning-based methods often struggle to grasp the contextual meanings of text fully and exhibit limitations in analyzing subtle expressions of emotion [29].

Recent advancements in LLMs have significantly enhanced the sophistication and efficiency of sentiment analysis. LLMs possess exceptional contextual understanding capabilities and excel in analyzing nuanced sentiments within sentences, allowing for a more accurate reflection of sentiment states in financial markets [7,8]. For example, BERT analyzes bidirectional contexts to extract positive and negative sentiments from news articles meticulously, making it a valuable tool for capturing real-time investor psychology [30]. Moreover, generative language models such as GPT-3 are employed not only for sentiment analysis but also for predicting trends in emotional changes, with studies demonstrating their ability to forecast market trends based on social media data from financial platforms [9].

Amidst these developments, the Llama 2 model has garnered attention owing to its specialized applications in financial data analysis. By fine-tuning Llama 2 to classify the sentiments of financial news as positive, negative, or neutral, researchers have applied it to algorithmic trading and portfolio management to reflect investor psychology accurately [12]. Furthermore, Llama 3-based systems have been developed for financial news summarization, named entity recognition, and sentiment analysis to support investor decision-making. Recent research [13] demonstrates that Llama 3 effectively leverages domain-specific knowledge, outperforming baseline models in financial sentiment analysis. However, its performance is sensitive to sentence length, exhibiting strong performance on short texts but encountering challenges with longer texts, thus highlighting a key area for improvement.

### 2.3. Prompting Performance Enhancement Techniques

Various techniques have been employed to maximize the performance of sentiment analysis using LLMs. These techniques play a crucial role in enhancing the precision and reliability of the analysis by improving contextual understanding and logical reasoning and reinforcing the credibility of the results. Representative techniques include CoT, ICL, SuperICL, and LLM Bootstrapping.

CoT facilitates a deeper understanding of context by guiding LLMs through a step-by-step reasoning process. This method supports a logical thinking structure that can discern differences in sentiments, thereby enabling more accurate analysis of complex emotions within text. CoT enhances the precision of models across various tasks, particularly in sentiment analysis, which requires nuanced judgments. By promoting a logical flow, CoT allows the model to interpret emotions deductively and systematically based on the given context, rather than merely listing words [31].

ICL is designed to improve the precision of sentiment analysis by enabling the model to understand the context within the input data without requiring additional model training [7]. ICL provides example responses to the model, which then generates results based on these examples. This technique allows the model to make predictions by leveraging real-time context without relying solely on past data. By supplying domain-specific examples, ICL guides the model to achieve better performance, particularly when clear examples are provided, leading to higher accuracy in sentiment analysis.

SuperICL extends the concept of ICL by utilizing responses from specialized smaller models such as FinBERT, instead of user-provided examples, to enhance the accuracy [32]. This technique is particularly beneficial for data for which clear answers are not readily available, such as financial texts. Financial data often contain subtle emotional variations that are influenced by word choice and context, making SuperICL highly effective in these specialized cases. It achieves excellent performance even with data that include complex emotional expressions, thereby enhancing the reliability of sentiment analysis.

LLM Bootstrapping leverages the characteristics of LLMs to generate diverse responses to the same query through a sampling method, reinforcing the credibility of the results via a majority vote approach [8]. This technique ensures consistent results by comparing multiple repeated responses and selecting the most reliable response. In sentiment analysis experiments using Reddit microblogging financial text data, generating eight repeated responses to the same question yielded clear majority results for approximately 85% of the data. Consequently, in this study we employ a five-repeat approach to generate responses with the aim of increasing the computational efficiency while maintaining result consistency. This method significantly reduces the variability of model-generated responses and ensures the reliability of sentiment analysis.

Collectively, these techniques enhance the performance of LLM-based financial sentiment analysis by improving the contextual understanding and logical reasoning and reinforcing result credibility. In particular, SuperICL effectively addresses the limitations of sentiment analysis for complex and ambiguous text data. By combining these techniques, sentiment analysis using LLMs can yield more sophisticated and reliable results.

## 3. Investment Strategy Design and Evaluation Methodology

This study evaluates the performance of sentiment analysis in financial markets using LLMs by constructing portfolios for each model and comparing the metrics related to profitability and volatility. This section introduces portfolio construction strategies and evaluation metrics.

### *3.1. Portfolio Construction Based on LLM Sentiment Analysis*

A portfolio was constructed using LLMs to conduct sentiment analysis of news articles. Investment strategies were implemented in the US market based on the sentiment analysis results for specific companies from the previous day. This study employed three investment strategies, each executed on the subsequent trading day based on sentiment analysis predictions. In the long strategy, a portfolio was constructed by purchasing all stocks predicted to exhibit positive sentiment at the opening price and selling them at the closing

price on the following trading day. Conversely, the short strategy involved short-selling stocks predicted to exhibit negative sentiment at the opening price and repurchasing them at the closing price on the following trading day. The long–short strategy combined these approaches, simultaneously taking long positions in stocks predicted to exhibit positive sentiment and short positions in stocks predicted to exhibit negative sentiment. This portfolio construction approach was based on the methodology proposed by Lopez-Lira and Tang [9], incorporating existing research on portfolio formation strategies [33].

Daily sentiments were calculated for the top 30 major companies by market capitalization on Nasdaq in the US market. Sentiments were classified as positive, negative, or neutral based on news headlines. Companies classified as having positive sentiment were included in the long strategy, whereas those with negative sentiment were incorporated into the short strategy. Companies deemed to have neutral sentiment were excluded from both strategies. This approach assumes that positive sentiment is predictive of future stock price increases, while negative sentiment is indicative of potential price decreases. Therefore, companies with neutral sentiment, lacking a clear directional signal, were not included in either the long or short strategy. The overall performance of these strategies was assessed by averaging the returns of the companies within each position. This methodology validated the effectiveness of sentiment analysis-based investment strategies and demonstrated their potential in enhancing portfolio performance [3–5].

### *3.2. Portfolio Performance Measurement Metrics*

Unlike traditional sentiment analysis studies that evaluate models based on accuracy and the F1-score, this study assesses model effectiveness through investment performance. A significant challenge in financial sentiment analysis is the inherent difficulty of precise labeling. Unlike standard NLP tasks with well-defined ground truths, sentiment in financial news is often ambiguous, context-dependent, and subject to interpretation, making accurate labeling difficult. Moreover, sentiment does not always directly correlate with immediate stock price movements, further limiting the relevance of conventional classification metrics for evaluating model effectiveness.

Given these challenges, this study adopts portfolio performance as the primary evaluation metric, following the methodology of previous research [9]. This approach ensures a realistic assessment of how sentiment analysis translates into financial decision-making, rather than relying on abstract classification accuracy. Furthermore, directly mapping positive and negative sentiment labels to stock price movements is often inappropriate due to the inherent complexities of financial markets. Therefore, instead of relying on standard classification metrics, this study employs established portfolio performance indicators to provide more practical and meaningful evaluation of sentiment analysis models in an investment context.

The daily average return represents the average return generated by the portfolio on a daily basis. The overall performance of the portfolio can be assessed by calculating the mean of each daily return, which is mathematically expressed as follows:

$$\text{Daily average return} = \frac{1}{N}\sum_{i=1}^{N} r_i = \bar{r}, \tag{1}$$

where $r_i$ denotes the return on day $i$ and $N$ is the total number of trading days.

The daily standard deviation serves as an indicator of portfolio return volatility. It is calculated by determining the standard deviation of the daily returns, which helps to evaluate the stability of the portfolio:

$$\text{Daily standard deviation} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(r_i - \bar{r})^2} = \sigma_d. \tag{2}$$

The Sharpe ratio is a crucial metric to assess returns relative to risk. It is calculated by subtracting the risk-free rate from the average return and dividing the result by the standard deviation of the returns. This ratio serves as a key indicator for identifying portfolios that provide higher returns for the same level of risk. In this study, the risk-free rate was assumed to be 0%. The Sharpe ratio was computed based on daily returns and subsequently annualized by multiplying it by the square root of 252, reflecting the average number of trading days per year. This standard method transforms daily volatility into annual volatility, thereby enhancing the comparability and interpretability of the Sharpe ratio.

$$\text{Sharpe ratio} = \frac{\bar{r} - r_f}{\sigma_d} \times \sqrt{252}\,, \tag{3}$$

where $r_f$ is the risk-free rate.

The maximum drawdown (MDD) quantifies the maximum loss ratio experienced by the portfolio, indicating the extent to which the asset value has declined from its peak to its trough during the evaluation period. It is defined as follows:

- Peak value: The highest asset value reached by the portfolio during a specific period.
- Trough value: The lowest asset value reached after the peak within the same period.

The MDD is calculated using the following formula:

$$\text{MDD} = \frac{Peak\ value - trough\ value}{Peak\ value}\,. \tag{4}$$

The final return represents the total return generated by the portfolio at the end of the study period. It is calculated by setting the initial investment amount to 1, thereby simplifying the analysis to illustrate the change in asset value relative to initial capital:

- Final portfolio value: The asset value of the portfolio at the end of the study period.
- Initial investment: The initial investment amount at the start of the study period (=1).

The final return is calculated using the following formula:

$$\text{Final return} = (final\ portfolio\ value - 1) \times 100. \tag{5}$$

## 4. Experimental Design

### *4.1. Experimental Data*

US market data were collected for the top 30 companies by market capitalization on Nasdaq from 2021 to 2023. News data were gathered using the GNews application programming interface (API) library, which analyzes changes in sentiment through daily news articles related to major companies. The GNews API facilitates convenient retrieval of the latest news from various sources via API calls, collecting significant English-language articles that are highly relevant to each company daily. Initially, 21,599 news articles were obtained; however, after applying a filtering process, 13,696 articles were used for analysis. Articles were excluded if the subsequent day was not a trading day, as this would prevent accurate assessment of market impact. Beyond this, no machine learning techniques or subjective value judgments were employed to filter the articles. Table 1 provides an

overview of the companies included in the dataset, along with their ticker symbols and the number of news articles gathered.

**Table 1.** Number of news articles collected for top 30 Nasdaq companies (2021–2023).

| No. | Ticker | Company Name (Location) | News Count | No. | Ticker | Company Name (Location) | News Count | No. | Ticker | Company Name (Location) | News Count |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AAPL | Apple Inc. (Cupertino, CA, USA) | 598 | 11 | BKNG | Booking Holdings Inc. (Norwalk, CT, USA) | 763 | 21 | META | Meta Platforms, Inc. (Menlo Park, CA, USA) | 742 |
| 2 | ADBE | Adobe Inc. (San Jose, CA, USA) | 146 | 12 | CMCSA | Comcast Corporation (Philadelphia, PA, USA) | 77 | 22 | MSFT | Microsoft Corporation (Redmond, WA, USA) | 583 |
| 3 | AMAT | Applied Materials, Inc. (Santa Clara, CA, USA) | 354 | 13 | COST | Costco Wholesale Corporation (Issaquah, WA, USA) | 743 | 23 | NFLX | Netflix, Inc. (Los Gatos, CA, USA) | 255 |
| 4 | AMD | Advanced Micro Devices, Inc. (Santa Clara, CA, USA) | 745 | 14 | CSCO | Cisco Systems, Inc. (San Jose, CA, USA) | 212 | 24 | NVDA | NVIDIA Corporation (Santa Clara, CA, USA) | 426 |
| 5 | AMGN | Amgen Inc. (Thousand Oaks, CA, USA) | 81 | 15 | GOOG | Alphabet Inc. (Mountain View, CA, USA) | 339 | 25 | PDD | PDD Holdings Inc. (Shanghai, China) | 426 |
| 6 | AMZN | Amazon.com, Inc. (Seattle, WA, USA) | 731 | 16 | GOOGL | Alphabet Inc. (Mountain View, CA, USA) | 381 | 26 | PEP | PepsiCo, Inc. (Harrison, NY, USA) | 756 |
| 7 | ARM | Arm Holdings plc (Cambridge, UK) | 759 | 17 | HON | Honeywell International Inc. (Charlotte, NC, USA) | 758 | 27 | QCOM | Qualcomm Incorporated (San Diego, CA, USA) | 715 |
| 8 | ASML | ASML Holding N.V. (Veldhoven, Netherlands) | 460 | 18 | INTU | Intuit Inc. (Mountain View, CA, USA) | 195 | 28 | TMUS | T-Mobile US, Inc. (Bellevue, WA, USA) | 437 |
| 9 | AVGO | Broadcom Inc. (San Jose, CA, USA) | 122 | 19 | ISRG | Intuitive Surgical, Inc. (Sunnyvale, CA, USA) | 85 | 29 | TSLA | Tesla, Inc. (Austin, Texas, USA) | 692 |
| 10 | AZN | AstraZeneca plc (Cambridge, UK) | 242 | 20 | LIN | Linde plc (Guildford, UK) | 766 | 30 | TXN | Texas Instruments Incorporated (Dallas, Texas, USA) | 107 |

The stock price data were collected from the Yahoo Finance Library. The daily closing prices of each company were tracked throughout the study period to analyze stock price volatility. These data were then combined with the results of news sentiment analysis (Table 1) to assess their impact on portfolio performance. Yahoo Finance is a reliable source that is widely used for collecting financial data, enabling the evaluation of daily price volatility through the opening and closing prices of each company.

### *4.2. LLMs Used in Sentiment Analysis*

This study focused on comparing LLM-based sentiment analysis by utilizing pre-trained financial models. Traditional machine learning models such as SVM, Naïve Bayes, LSTM, and CNN were initially considered for sentiment analysis. However, sentiment labeling for news articles is inherently challenging due to subjective interpretations, context dependency, and the lack of a standardized labeling criterion. Additionally, labeling

methods based on returns were found to be inadequate, making these traditional approaches unsuitable for this study. Instead, we adopted LLMs, which can provide more contextualized and nuanced sentiment analysis without requiring precise labels.

Sentiment analysis was conducted using BERT-based models and open-source generative LLMs. The selection of various LLMs was based on the fact that each model is optimized for specific domains and tasks, allowing them to exhibit diverse performance levels. The LLM models used in the study are summarized in Table 2.

**Table 2.** Summary of LLMs.

| Model Type | Model Name | Manufacturer (Location) | Release Year | Number of Parameters | Advantages | Disadvantages | Features |
|---|---|---|---|---|---|---|---|
| Discriminative model | BERT-Sentiment | NLP Town (Lubbeek, Belgium) | 2021 | 167.36 M | Excellent for multilingual sentiment analysis and classifying sentiments as positive, negative, or neutral | Optimized for short texts such as Twitter; limited in analyzing long texts | Based on BERT, optimized for sentiment analysis of user reviews and social media texts |
| | FinBERT-tone | Hong Kong University of Science and Technology (Kowloon, Hong Kong) | 2021 | 109.75 M | Excellent performance in English sentiment analysis specialized for the financial domain | Performance may degrade when applied to texts outside the financial field | Based on BERT, fine-tuned to analyze sentiments (positive, negative, or neutral) in financial texts |
| | RoBERTa-Finance | Concordia University (Montreal, QC, Canada) | 2021 | 355.36 M | Optimized performance for analyzing financial news and reports | Limited when applied to texts outside the financial domain | Trained RoBERTa model for financial text sentiment analysis |
| Generative model | Llama 3.1 3.2 | Meta (Menlo Park, CA, USA) | 2024 | 3 B, 8 B | High performance provided as open-source for free | May be inferior in performance compared with some competitive models | Uses a tokenizer with 128,000 tokens and Grouped-Query Attention |
| | Mistral | Mistral AI (Paris, France) | 2023 | 7 B | High performance with a lightweight model | Limited in complex tasks compared with ultra-large models | Suitable for real-time data analysis and conversational AI applications |
| | Gemma 2 | Google (Mountain View, CA, USA) | 2024 | 27 B | Excellent sentiment analysis capabilities across various languages and cultural contexts | May be limited in certain advanced analytical tasks | Outstanding performance in text data processing and long text analysis |

Discriminative and generative language models are employed in various NLP tasks to leverage their unique characteristics and strengths. Discriminative models focus on analyzing and understanding text, specializing in specific domain tasks such as text classification, sentiment analysis, and question-answering (Q&A). Notable examples include BERT-Sentiment, FinBERT-tone, and RoBERTa-Finance from the BERT family. These models provide high accuracy and reliability in the financial domain, excelling in the analysis of structured data such as financial news, corporate reports, and analyst reports. Discriminative models typically have relatively lightweight structures and exhibit high efficiency through fine-tuning processes that are tailored to specific tasks.

In contrast, generative language models excel in versatile and creative tasks such as text generation, summarization, and conversational applications. Generative models such as Llama 3.1, 3.2, Mistral, and Gemma 2 learn from large-scale data and demonstrate outstanding performance in long contexts, multilingual support, and multimodal tasks.

Specifically, generative models can understand contextual meanings within extended contexts to generate text or perform complex information extraction tasks. Although discriminative models offer high performance in tasks optimized for specific domains, generative models provide differentiated value in terms of versatility and creative applications, necessitating appropriate selection based on usage purposes and data characteristics.

BERT-Sentiment is a pre-trained language model optimized for multilingual sentiment analysis that exhibits excellent performance in five-level emotion classification tasks, including positive, negative, and neutral sentiments. In particular, it excels in analyzing social media data and user reviews, which are frequently utilized for sentiment analysis of short text data such as Twitter. For Twitter sentiment analysis, a dataset of 1,578,627 manually annotated tweets was used to evaluate the accuracy, precision, recall, and F1-score. In this study, BERT-Sentiment achieved an accuracy of 81.23% and an F1-score of 0.80, demonstrating performance comparable to that of RoBERTuito. In contrast, RuBERT recorded a relatively lower accuracy (78.44%) and F1-Score (0.78) [34].

FinBERT-tone is a pre-trained language model that specializes in financial text sentiment analysis. Developed by a research team at the Hong Kong University of Science and Technology (HKUST), this model provides high accuracy in classifying sentiments in financial documents as positive, negative, or neutral. FinBERT-tone is based on BERT and extends the concept of Liu et al.'s FinBERT model by optimizing the performance in the English financial domain [35]. The primary training data include large-scale financial text datasets such as Reuters News and Corporate Reports, encompassing data from financial news, corporate disclosures, and stock reports. This model understands the contextual meanings of financial texts and exhibits excellent performance in sentiment analysis, making it highly applicable to financial services, investment psychology analysis, and market trend evaluation.

RoBERTa-Finance is a RoBERTa Large-based model specialized for financial text analysis, developed by Mohammad Soleimanian's research team at Concordia University. The model is optimized for accurately classifying sentiments as positive, negative, or neutral. It was trained using a diverse dataset that included ESG (Environmental, Social, and Governance) news, corporate disclosures, CSR (Corporate Social Responsibility)-related news, and stock reports. RoBERTa-Finance excels in understanding and analyzing subtle sentiments in financial news and corporate reports, demonstrating its strong effectiveness in domain-specific financial tasks, which makes it a valuable tool for sentiment analysis in the financial sector [36].

Llama 3.1 and 3.2 are the latest AI language models developed by Meta, and feature hundreds of millions of diverse parameters. These models exhibit outstanding performance across various tasks, including text generation, code generation, and natural language understanding. They utilize an efficient tokenizer that accommodates 128,000 tokens and are trained on a massive dataset comprising over 15 trillion tokens. By employing the latest Transformer architecture and GQA, the Llama models enhance the inference efficiency and achieve superior performance in handling long contexts and accurate information retrieval. Notably, Llama 3.1 and 3.2 are open-source and freely available, facilitating their widespread application in both research and commercial settings [21]. Llama 3.1 includes parameter configurations such as 8 B and 405 B, excelling in general knowledge processing, mathematics, tool usage, and multilingual translation tasks. Llama 3.2 is offered in multiple sizes, including 1 B, 3 B, 11 B, and 90 B, with the 11 B and 90 B models incorporating image

recognition capabilities for multimodal processing. The Llama 3 series has demonstrated efficiency in managing long contexts and diverse natural language tasks, making it suitable for a wide range of applications.

Mistral is a lightweight language model launched in September 2023. It is equipped with hundreds of millions of parameters and delivers high performance across various NLP tasks. It leverages GQA and SWA to optimize the inference speed and memory efficiency, achieving excellent performance even when processing long sequences. Mistral 7 B has outperformed models such as Llama 2 13 B and Llama 1 34 B, excelling in complex tasks such as inference, mathematics, and code generation. Additionally, it has demonstrated performance comparable to that of Code-Llama 7 B on code datasets such as Humaneval and MBPP, maintaining high accuracy without the need for fine-tuning for specific tasks. This design achieves both efficient learning and inference, providing high performance despite being a lightweight model, and has garnered significant attention for its capabilities [26].

Gemma 2 is the latest LLM developed by Google, and is available in versions with 700 million (7 B), 900 million (9 B), and 2.7 billion (27 B) parameters. It significantly improves the performance and efficiency compared with previous versions. This model surpasses similarly sized models through enhancements in the Transformer architecture and knowledge distillation, achieving high scores in general language understanding (MMLU) with 75.2 points, reasoning ability (BBH) with 74.9 points, and HellaSwag with 86.4 points in the 27 B model. Specifically designed for efficient inference, Gemma 2 can be executed on a single device using NVIDIA A100 and H100 Tensor Core GPUs, thereby reducing the deployment costs while maintaining high performance. Furthermore, Gemma 2 is compatible with various AI frameworks and contributes to the advancement of the AI community through its open license, making it suitable for both research and commercial applications [27].

### *4.3. Prompt Configuration*

BERT-based models are equipped with built-in functions specifically designed for sentiment analysis tasks that ensure that the results are produced in a consistent and stable format. This inherent structure allows BERT-based models to deliver reliable outputs for sentiment analysis tasks without requiring additional configurations. In contrast, generative LLMs are primarily optimized for generative tasks, which necessitate the use of distinct prompt configurations to perform sentiment analysis effectively. Consequently, prompts specifically tailored for sentiment analysis were developed for generative LLMs to ensure consistency and reliability of the results.

According to Lopez-Lira and Tang [9], assigning a financial expert to a model significantly enhances the reliability and accuracy of its responses. Building on prior research, sentiment analysis was conducted by designating the role of the model as a "financial expert". This approach was informed by the findings of both studies, ensuring that the model responses aligned with the expectations of financial expertise. Figure 1 presents the basic prompt developed for this study, and Figure 2 illustrates the application of the CoT technique. In addition, Figure 3 shows a prompt that incorporates the SuperICL method, utilizing the results from the BERT model. The analysis environment for this study is summarized in Appendix A.

**System Message**: You are a stock analyst specializing in assessing sentiment in financial news articles.

**Prompt**: Based on the article titled {Title}, determine whether the tone is positive, negative, or neutral towards {Symbol}. Please respond in the following format, and omit any reasoning:

Sentiment: [positive/negative/neutral].

**Figure 1.** Basic prompt configuration.

**System Message**: You are a stock analyst specializing in assessing sentiment in financial news articles.

**Prompt**: Based on the article titled {Title}, determine whether the tone is positive, negative, or neutral towards {Symbol}. Please respond in the following format:

Sentiment: [positive/negative/neutral]

Reason: [Brief explanation based on the article]

**Figure 2.** CoT prompt configuration.

**System Message**: You are a stock analyst specializing in assessing sentiment in financial news articles.

**Prompt**: Based on the article titled {Title}, determine whether the tone is positive, negative, or neutral towards {Symbol}.

The sentiment prediction from BERT is {BERT Answer} Please respond in the following format, and omit any reasoning If your sentiment differs from the reference prediction, provide a brief reason why

Sentiment: [positive/negative/neutral]

Reason: [Provide reason only if your sentiment differs from the reference]

**Figure 3.** SuperICL prompt configuration.

## 5. Experimental Results

Table 3 presents the portfolio performance metrics derived from the various models used in the sentiment analysis, highlighting their daily average returns, standard deviations, Sharpe ratios, MDDs, and final returns based on different portfolio strategies (long, short, and long–short). Overall, long strategies generally yield better outcomes than short and long–short strategies.

**Table 3.** Portfolio performance metrics for basic models and basic prompting.

| Model Name | Portfolio Strategy | Daily Average Return (%) | Standard Deviation (%) | Sharpe Ratio | MDD (%) | Final Return (%) |
|---|---|---|---|---|---|---|
| FinBERT | Long | 0.0423 | 1.5575 | 0.4316 | −45.84 | 26.60 |
| | Short | −0.0759 | 1.5989 | −0.7538 | −59.93 | −50.03 |
| | Long–Short | −0.0168 | 0.8203 | −0.3250 | −34.12 | −14.57 |
| BERT | Long | 0.0428 | 1.2055 | 0.5637 | −31.01 | 31.98 |
| | Short | −0.0444 | 1.3478 | −0.5225 | −46.23 | −34.14 |
| | Long–Short | −0.0008 | **0.4075** | −0.0302 | −16.47 | −1.25 |
| RoBERTa Finance | Long | 0.0459 | 1.2625 | 0.5772 | −26.39 | 34.48 |
| | Short | −0.0328 | 1.6084 | −0.3235 | −42.60 | −30.08 |
| | Long-Short | 0.0066 | 0.6252 | 0.1667 | −17.17 | 3.67 |
| Llama 3.1 | Long | 0.0526 | 1.2975 | **0.6433** | −25.48 | 41.17 |
| | Short | −0.0320 | 1.5579 | −0.3264 | −39.54 | −29.21 |
| | Long–Short | 0.0103 | 0.5515 | 0.2957 | −15.80 | 7.07 |
| Llama 3.2 | Long | 0.0104 | 1.3729 | 0.1208 | −39.65 | 0.79 |
| | Short | −0.0586 | 1.4729 | −0.6311 | −50.25 | −41.89 |
| | Long–Short | −0.0241 | 0.5859 | −0.6519 | −32.05 | −18.24 |
| Mistral | Long | 0.0122 | 1.4916 | 0.1300 | −49.76 | 0.83 |
| | Short | −0.0275 | 1.5135 | −0.2883 | −43.20 | −26.24 |
| | Long–Short | −0.0076 | 0.6752 | −0.1795 | −24.31 | −7.45 |
| Gemma 2 | Long | **0.0546** | 1.3706 | 0.6329 | −26.05 | **42.39** |
| | Short | −0.0430 | 1.5237 | −0.4478 | −45.01 | −34.74 |
| | Long–Short | 0.0058 | 0.6654 | 0.1391 | **−12.37** | 2.86 |
| Nasdaq 30 | Nasdaq 30 | 0.0406 | 1.1772 | 0.5476 | −21.19 | 30.10 |

(Bold and underline indicate the best performance in each column.)

Among the discriminative models, RoBERTa-Finance demonstrated strong performance in the long strategy, with an average return of 0.0459% and a Sharpe ratio of 0.5772. FinBERT exhibited stable performance in the long strategy, achieving a daily average return of 0.0423%, Sharpe ratio of 0.4316, and final return of 26.60%. BERT showed better profitability than FinBERT in the long strategy, with an average return of 0.0428%, a Sharpe ratio of 0.5637, and a final return of 31.98%.

Interestingly, the general BERT model outperformed the finance-specialized FinBERT model in terms of profitability, which is a noteworthy finding. The data collected from the GNews API, which aggregates news from a wide range of English media sources, including general news and not exclusively finance-related news, contributed to the effectiveness of the models in a global investment context such as Nasdaq.

Among all models, Llama 3.1, which is recognized for its recent advancements in performance across various tasks, achieved the highest metrics in the long strategy, with a daily average return of 0.0526%, Sharpe ratio of 0.6433, and final return of 41.17%. However, its performance in the short strategy was limited, yielding an average return of –0.0320%

and a Sharpe ratio of −0.3264. Similarly, Gemma 2 demonstrated stable and favorable results in the long strategy, with a daily average return of 0.0546%, Sharpe ratio of 0.6329, and final return of 42.39%.

Moving forward, it is essential to explore methodologies for integrating discriminative and generative models, such as SuperICL, to enhance the overall performance further and capitalize on the strengths of both approaches.

Figure 4 visualizes the cumulative return trends for the Llama 3.1 model applied to financial sentiment analysis, showing the performance of various portfolio strategies (long, short, and long–short) over time.

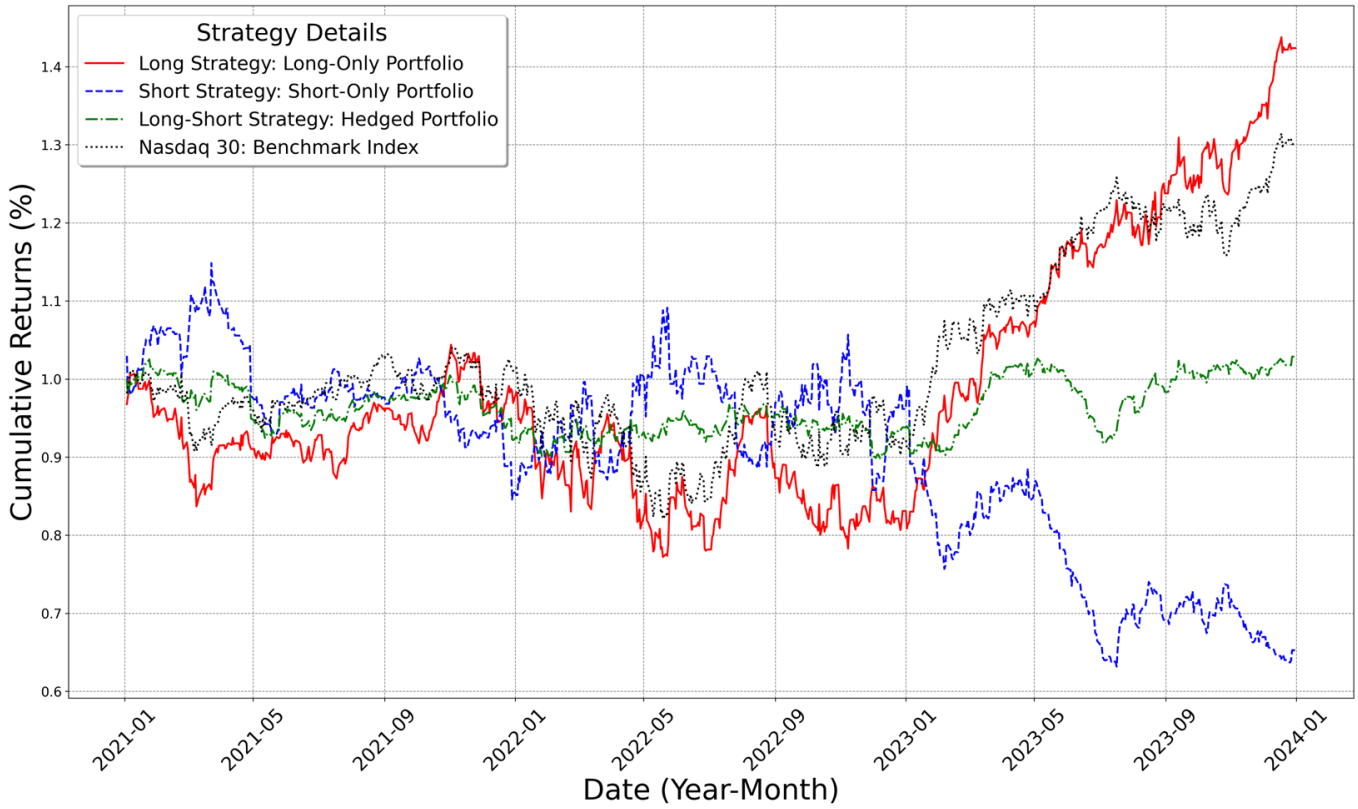


**Figure 4.** Llama 3.1 cumulative return trend: basic prompt.

The long strategy (red) exhibited a consistent upward trajectory, indicating a robust cumulative return that outperformed the benchmark index throughout the observed period. This suggests that the long strategy effectively capitalizes on positive market sentiment, leading to favorable investment outcomes.

In contrast, the short strategy (blue) exhibited a downward trend, reflecting the challenges associated with short-selling in a generally bullish market environment. The cumulative returns of this strategy declined significantly, highlighting the inherent risks and potential losses when market conditions are not conducive to short positions.

The long–short strategy (green) showed moderate performance relative to the long strategy but remained resilient compared with the short strategy. Its cumulative returns fluctuated within a narrower range, suggesting that the hedging mechanism provides protection against market volatility.

Overall, the cumulative return analysis indicates that the long strategy employed by the Llama 3.1 model outperformed both the short and long–short strategies, thereby reinforcing the efficacy of the model in sentiment analysis and portfolio construction.

The stark contrast in performance among these strategies underscores the importance of selecting appropriate investment approaches based on prevailing market conditions.

Notably, the short strategy showed considerable profit improvements during the initial bear market and periods of mid-term stagnation, suggesting that it can effectively capitalize on downturns. However, this strategy consistently underperformed in the subsequent bull markets, highlighting its vulnerability in rising market conditions. Conversely, while the long strategy struggled to outperform the Nasdaq benchmark in the early to mid-term phases, it eventually surpassed it during the final bull market, yielding relatively superior results.

Despite this success, the long strategy could not maintain a continuous advantage throughout the entire analysis period, indicating fluctuations in performance that warrant further investigation. Therefore, to address these limitations and enhance the overall performance, it is essential to explore additional prompting techniques and approaches for combining discriminative and generative models, such as SuperICL, to provide more consistent results across varying market environments.

Figure 5 presents the distribution of sentiments analyzed by Llama 3.1 with basic prompt. This reveals that a substantial proportion of the evaluations were classified as neutral, accounting for 49.8% of all sentiment assessments. This high percentage of neutral sentiments may be attributed to the nature of the news sources aggregated from the GNews API, which often includes a wide range of content that may not convey strong sentiment.

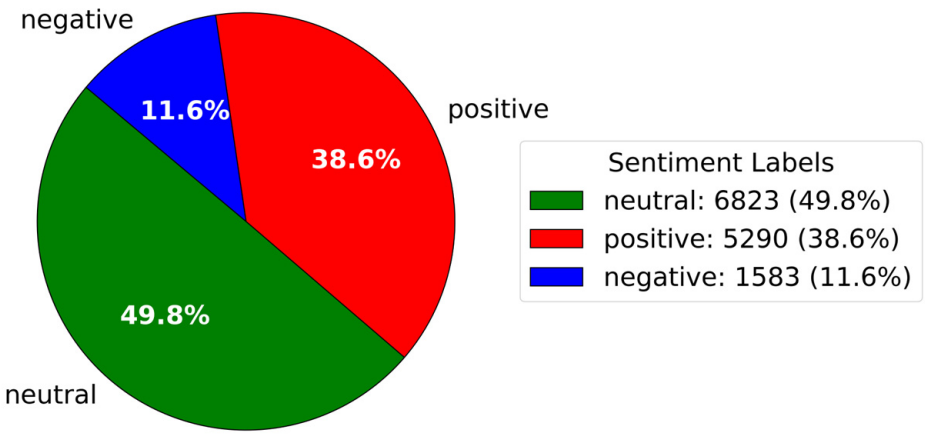


**Figure 5.** Sentiment distribution of the Llama 3.1 model: basic prompt.

Furthermore, the analysis indicates limitations in accurately capturing negative financial news, with only 11.6% of the sentiments classified as negative. This underrepresentation of negative sentiments suggests that critical adverse events may not have been adequately reflected, thereby skewing the overall sentiment analysis. These factors contributed to the ability of the long strategy to outperform the market during the bull phases that are characteristic of Nasdaq. However, the lack of an active sentiment reflection for specific companies limited the potential of the model to achieve maximum profitability.

To address these limitations, it is essential to implement various prompting techniques to improve the sentiment analysis accuracy. Figure 6 presents a comprehensive performance analysis of various prompting techniques applied to generative LLM models in the context of sentiment analysis of financial news. The boxplots illustrate the mean return, Sharpe ratio, final return (cumulative), and MDD metrics across different methodologies, including basic prompting (NOCoT), CoT (CoT), SuperICL (BERT-ICL, FinBERT-ICL, and RoBERTaFinance-ICL), and Bootstrapping (Bootstrap).

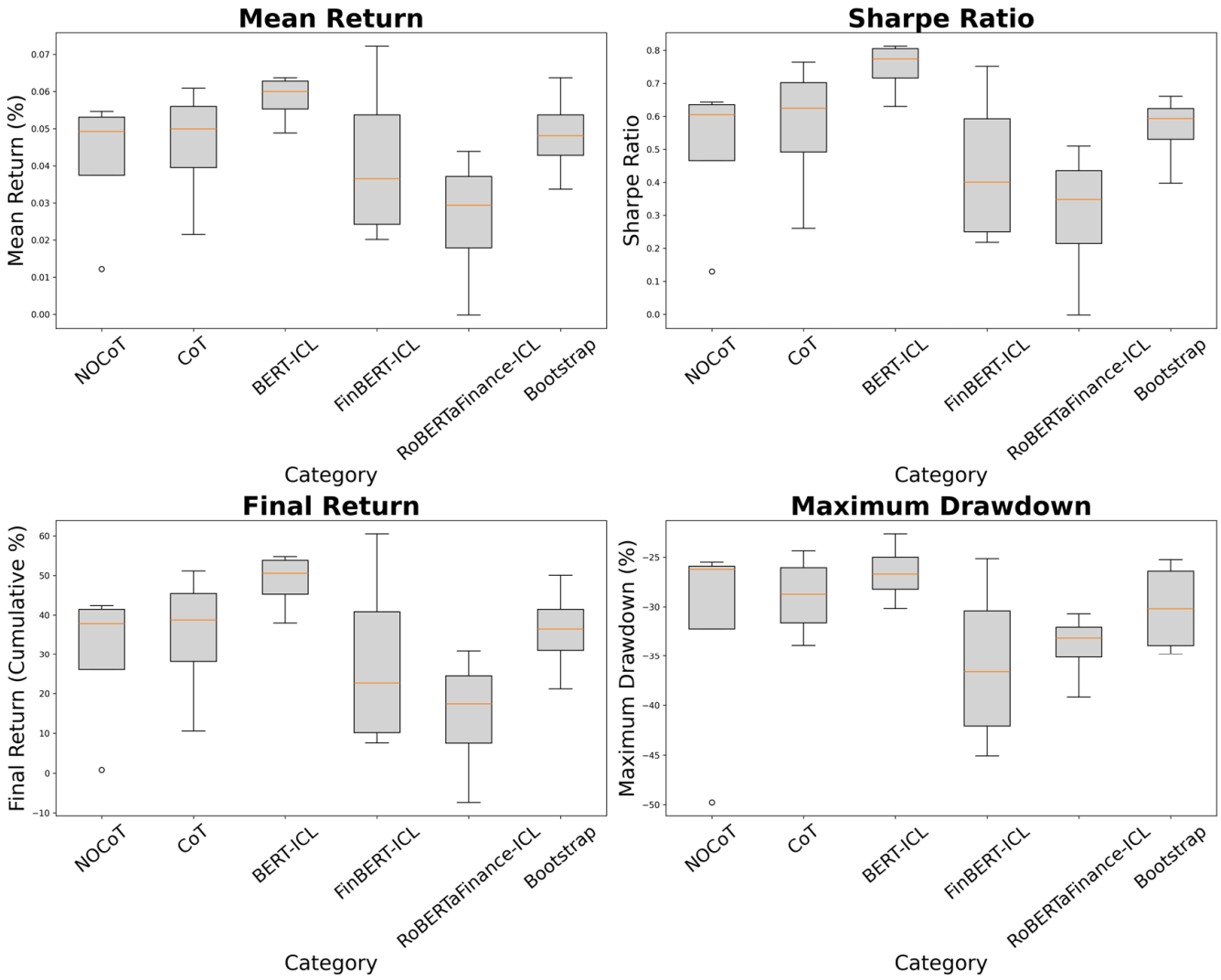


**Figure 6.** Performance metrics by prompting technique.

The results demonstrate that the BERT-based SuperICL approach achieved the highest levels of profitability and stability, as indicated by the superior daily average returns and Sharpe ratios compared with the other techniques. Furthermore, this method attained the highest final return among all evaluated models while exhibiting the lowest MDD, underscoring its consistent performance.

Following BERT-based SuperICL, LLM Bootstrapping and CoT exhibited favorable outcomes. LLM Bootstrapping ranked directly below BERT-based SuperICL in terms of the average return, Sharpe ratio, and final return, demonstrating its effectiveness as a prompting technique. CoT exhibited comparable returns and Sharpe ratios to Bootstrapping but exhibited greater variance, suggesting a degree of instability relative to the latter.

The NOCoT baseline strategy, while conservative, recorded intermediate average return and Sharpe ratio levels, with its final return falling short of those achieved by Bootstrapping and CoT. In addition, it experienced relatively larger MDDs, indicating less stability compared with the more effective strategies.

Interestingly, the finance-specialized models FinBERT-ICL and RoBERTaFinance-ICL performed below expectations. FinBERT-ICL ranked in the lower-middle range for the return, Sharpe ratio, and final return, with the largest MDD among all models. In contrast, RoBERTaFinance-ICL exhibited the lowest performance metrics overall, including the average return, Sharpe ratio, and final return.

Figure 7 presents the cumulative return trends of the top five portfolios ranked by the Sharpe ratio, constructed using a combination of various prompting techniques and LLM models for financial sentiment analysis, alongside the Nasdaq 30 benchmark. The results indicate that the combination of Llama 3.2 and BERT-ICL(llama3.2-BERT-ICL-Long) achieved impressive results in the long strategy, recording a daily average return of 0.0637%, Sharpe ratio of 0.8126, and final return of 54.78%.

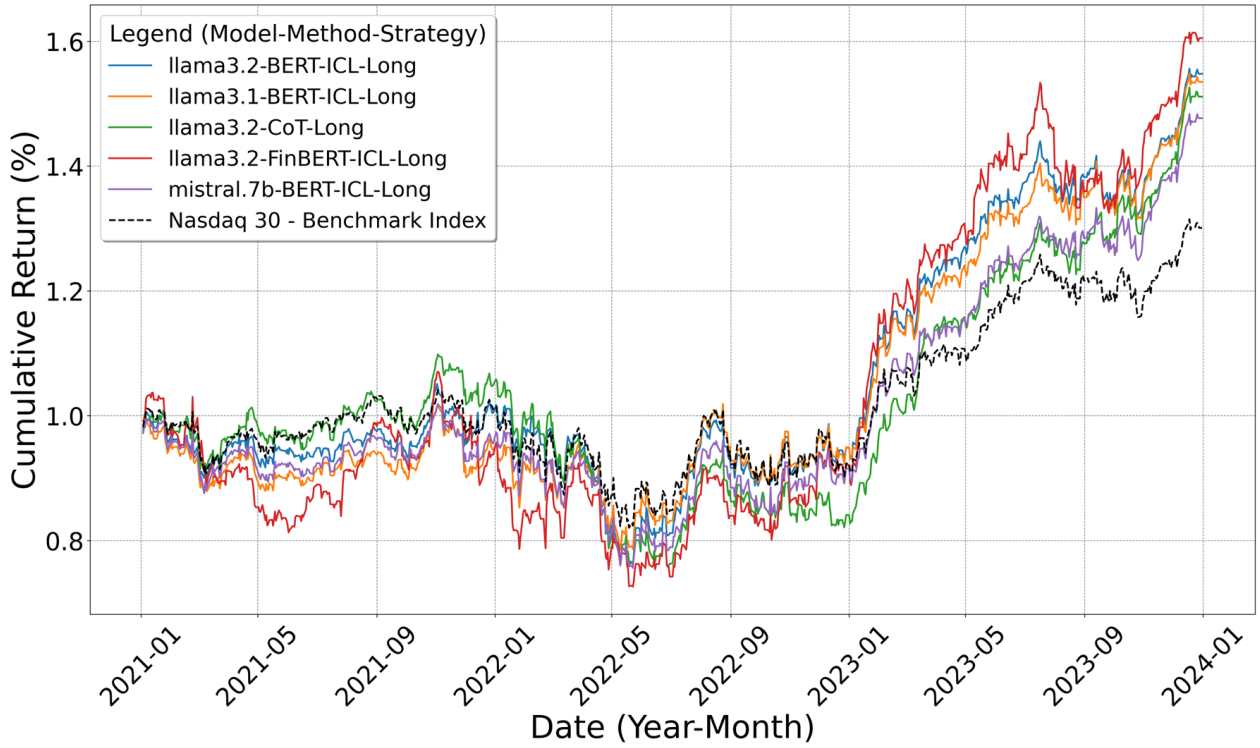


**Figure 7.** Cumulative trend of the top five Sharpe ratios and Nasdaq 30.

This performance represents a significant enhancement compared with the results of Llama 3.2 when utilizing only the basic prompt, where it failed to outperform the Nasdaq benchmark. These findings suggest that the integration of BERT-ICL has a positive influence on the model performance, enabling it to better capitalize on favorable market conditions. The analysis underscores the importance of employing advanced prompting techniques to refine sentiment analysis and optimize portfolio returns, highlighting the potential for improved risk-adjusted performance through strategic model enhancements.

Table 4 summarizes the performance metrics for the top five portfolios with the highest Sharpe ratios, along with the Nasdaq 30 benchmark. Notably, Llama 3.2 FinBERT-ICL achieved the highest returns among all evaluated models. However, as illustrated in Figure 6, the performance of FinBERT-ICL exhibited considerable variability depending on the generative model employed, indicating potential instability under certain market conditions. In contrast, BERT-ICL demonstrated more stable and consistent performance across different scenarios. Specifically, Llama 3.1 combined with BERT-ICL yielded strong results in the long strategy, recording an average daily return of 0.0625%, a Sharpe ratio of 0.8025, and a final return of 53.49%. This underscores the robust contribution of BERT-ICL in enhancing overall model stability and performance. These findings highlight the importance of carefully selecting both generative and retrieval-based models to optimize long-short investment strategies and maintain consistent profitability.

**Table 4.** Performance metrics of the top five Sharpe ratio portfolios and Nasdaq 30 benchmark.

| Model Name | Portfolio Strategy | Daily Average Return (%) | Standard Deviation (%) | Sharpe Ratio | MDD (%) | Final Return (%) |
|---|---|---|---|---|---|---|
| Llama 3.2 BERT-ICL | Long | 0.0637 | 1.2440 | **0.8126** | −27.57 | 54.78 |
| Llama 3.1 BERT-ICL | Long | 0.0625 | 1.2366 | 0.8025 | −22.63 | 53.49 |
| Llama 3.2 CoT | Long | 0.0609 | 1.2655 | 0.7638 | −30.84 | 51.12 |
| Llama 3.2 FinBERT-ICL | Long | **0.0722** | 1.5267 | 0.7511 | −32.18 | **60.50** |
| Mistral BERT-ICL | Long | 0.0574 | 1.2241 | 0.7447 | −25.79 | 47.68 |
| Nasdaq 30 | Nasdaq 30 | 0.0406 | **1.1772** | 0.5476 | **−21.19** | 30.10 |

(Bold and underline indicate the best performance in each column.).

In contrast, the CoT approach of Llama 3.2 exhibited slightly lower performance, with a Sharpe ratio of 0.7638, yet it still ranked among the top portfolios. BERT-ICL of Mistral also secured a position among the top five, with a Sharpe ratio of 0.7447, reflecting consistent performance across the board.

However, the Nasdaq market index exhibited relatively lower performance metrics, with an average daily return of 0.0406%, a Sharpe ratio of 0.5476, and a final return of 30.10%. This suggests that newer models, particularly Llama 3.2, which leverage an advanced understanding of example data, demonstrate competitive performance against the market benchmarks. The integration of BERT-ICL is crucial for maximizing model performance, emphasizing the significance of prompt design and utilization strategies in future research.

As illustrated in Figure 7, the top five strategies recorded returns that exceeded the Nasdaq market index post-2022. Specifically, the combination of BERT-ICL and Llama 3.2 and Llama 3.1 maintained stable and steep upward trajectories during the bull market, reflecting excellent performance with substantial improvements after 2022.

In addition, while the Llama 3.2 CoT approach exhibited somewhat limited performance in terms of the Sharpe ratio and final return, it nonetheless demonstrated excess performance relative to the market, affirming that the CoT-based strategy remains a viable option.

Overall, although the performance was somewhat constrained in bear markets, the long strategy achieved outstanding results in bull markets, particularly in upward-trending environments such as Nasdaq. This observation aligns with the characteristics of the US market, suggesting that long position-based strategies can achieve superior performance by leveraging bull markets.

In conclusion, these results confirm that BERT-ICL maximizes performance, especially when integrated with contemporary models such as Llama 3.2, and establish prompt-based learning approaches as vital tools for enhancing model understanding and data utilization capabilities. Collectively, these findings further emphasize the importance of model selection and prompt design for future research and practical implementations.

Figure 8 presents a comparative analysis of the prediction results from the BERT model and Llama 3.2 BERT-ICL, revealing a significant level of agreement between the two models. The graph shows that the predictions made by both Llama 3.2 BERT-ICL and BERT were identical in more than 10,000 cases, whereas 3077 instances exhibited differing prediction values. This marked difference suggests that both models generally arrived at similar conclusions in their sentiment judgments, lending credence to the ability of Llama 3.2 BERT-ICL to maintain the predictive performance characteristics of the BERT model. These findings ultimately indicate that Llama 3.2 BERT-ICL effectively replicates the sentiment analysis capabilities of BERT, confirming its alignment in evaluating financial news.

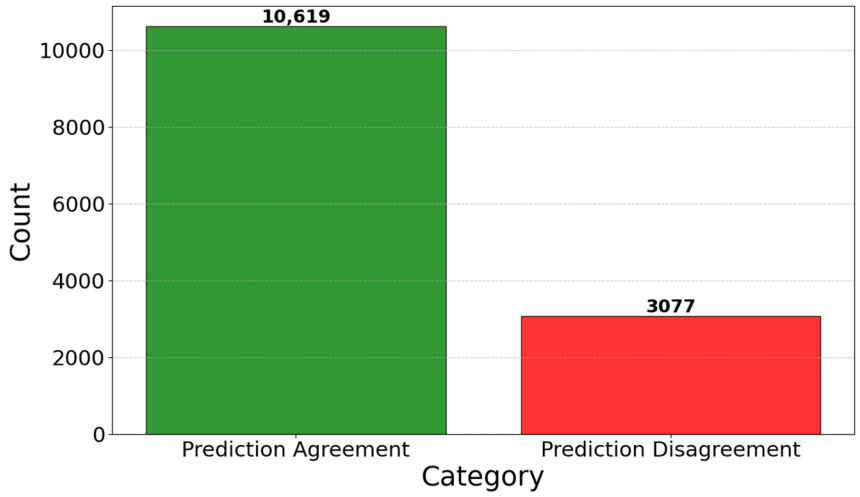


**Figure 8.** Prediction consistency between BERT and Llama 3.2 BERT-ICL in sentiment analysis.

In Figure 9, the sentiment distributions of the BERT model (a) and Llama 3.2 BERT-ICL (b) are compared to highlight the differences in their prediction outputs. The analysis reveals that BERT exhibited a high proportion of positive judgments, accounting for 61.4% of the total predictions, whereas the proportions of neutral (6.0%) and negative (32.6%) sentiments were relatively low.

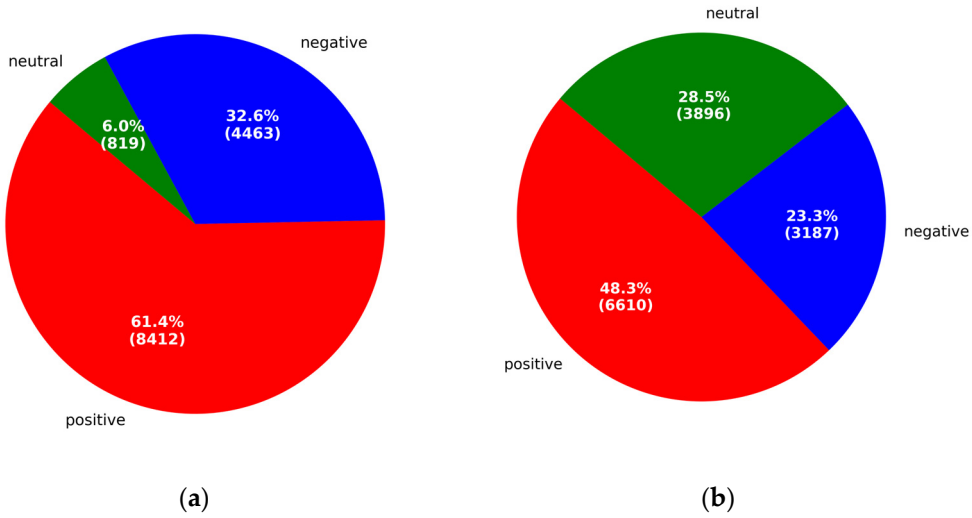


**Figure 9.** Sentiment distribution of (**a**) BERT and (**b**) Llama 3.2 BERT-ICL.

In contrast, Llama 3.2 BERT-ICL showed a more balanced distribution among the sentiment categories, with 48.3% positive, 28.5% neutral, and 23.3% negative sentiments. This indicates that Llama 3.2 BERT-ICL adopted a more cautious approach, reflecting a tendency to issue more conservative judgments and maintain a neutral stance more frequently than the BERT model. Overall, these findings suggest that while BERT is inclined towards more optimistic sentiment assessments, Llama 3.2 BERT-ICL provides a more nuanced analysis, demonstrating greater sensitivity to neutral and negative sentiments in financial news. This difference in sentiment distribution highlights the varying analytical approaches of the two models for sentiment analysis tasks.

Figure 10 presents an analysis of the changes in sentiment predictions from the BERT model to Llama 3.2 BERT-ICL. The data indicate that a substantial number of cases that were originally classified as positive or negative by BERT were revised to neutral by Llama

3.2 BERT-ICL. Notably, the most frequent adjustments were from positive to neutral, with 1802 cases, followed by changes from negative to neutral, totaling 1275 cases.

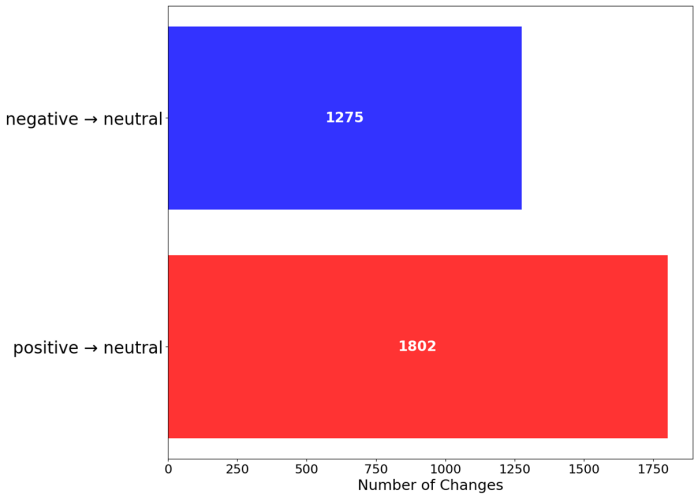


**Figure 10.** Changes in sentiment predictions from BERT to Llama 3.2 BERT-ICL.

Interestingly, the predictions did not change from neutral to positive or negative in any instances. This pattern suggests that Llama 3.2 BERT-ICL employs a more conservative judgment strategy, particularly in its assessment of positive and negative sentiments. The tendency to shift positive and negative predictions to neutral indicates enhanced prudence in the sentiment analysis of Llama 3.2 BERT-ICL, reflecting a cautious approach to sentiment categorization. This finding highlights the inclination of the model to mitigate overconfidence in positive or negative assessments, favoring a more balanced and measured interpretation of sentiment in financial news.

Figure 11 shows the cumulative return trends of the top four portfolios exhibiting the highest Sharpe ratios, achieved through the application of the Bootstrapping method combined with the BERT-ICL method for the sentiment analysis of financial news. The graph illustrates a pronounced upward trajectory for these portfolios, particularly compared with the Nasdaq 30 benchmark index, thereby demonstrating the effectiveness of the applied methodologies in enhancing portfolio performance.

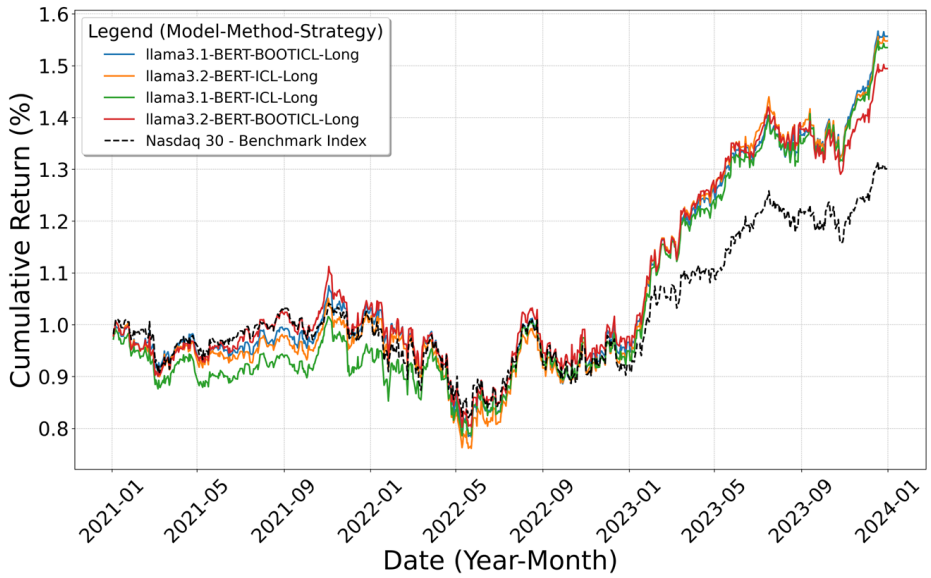


**Figure 11.** Top four Sharpe ratio portfolios and market cumulative trends after applying the Bootstrapping and SuperICL hybrid method.

Table 5 provides a comprehensive overview of the performance metrics for these portfolios, revealing that the hybrid method (BERT-BOOTICL) of Llama 3.1 achieved exceptional results in the long strategy, with an average return of 0.0641%, a Sharpe ratio of 0.8347, and a cumulative return of 55.68%. This represents a significant improvement in both the Sharpe ratio and cumulative return compared with the Llama 3.1 BERT-ICL model, which utilized only the BERT-ICL method without the Bootstrapping approach (Sharpe ratio of 0.8025, final return of 53.49%).

**Table 5.** Performance metrics of the top four Sharpe ratio portfolios.

| Model Name | Portfolio Strategy | Daily Average Return (%) | Standard Deviation (%) | Sharpe Ratio | MDD (%) | Final Return (%) |
|---|---|---|---|---|---|---|
| Llama 3.1 BERT-BOOTICL | Long | 0.0641 | 1.2194 | **0.8346** | −27.14 | **55.67** |
| Llama 3.2 BERT-ICL | Long | **0.0636** | 1.2439 | 0.8125 | −27.56 | 54.77 |
| Llama 3.1 BERT-ICL | Long | 0.0625 | 1.2366 | 0.8025 | −22.63 | 53.49 |
| Llama 3.2 BERT-BOOTICL | Long | 0.0588 | 1.2169 | 0.7677 | −27.88 | 49.44 |
| Nasdaq 30 | Nasdaq 30 | 0.0406 | **1.1772** | 0.5476 | **−21.19** | 30.10 |

(Bold and underline indicate the best performance in each column.).

Conversely, the hybrid method (BERT-BOOTICL) of Llama 3.2 exhibited an average return of 0.0589%, a Sharpe ratio of 0.7677, and a cumulative return of 49.44% in the long strategy, showing a slight decrease in the Sharpe ratio compared with the Llama 3.2 BERT-ICL model, which achieved a Sharpe ratio of 0.8126 and final return of 54.78% without the Bootstrapping method. However, the performance difference between the two models was relatively modest, indicating that the impact of the Bootstrapping method on the performance may vary depending on the model architecture or data processing techniques employed.

The Bootstrapping method demonstrated the capacity to provide more stable performance amid data volatility and uncertainty during the training process, indicating its potential to enhance investment outcomes when integrated with the BERT-ICL method. Portfolios that used both methods exhibited high average Sharpe ratios and resilience against market volatility, reinforcing the effectiveness of these techniques in optimizing investment strategies.

Overall, these results underscore the potential of combining advanced prompting techniques, such as Bootstrapping with generative LLM models to improve performance metrics in sentiment analysis-driven portfolios. These findings suggest that such hybrid approaches can foster more robust investment strategies that are adaptable to varying market conditions, paving the way for future research on financial forecasting and portfolio management.

## 6. Explainability Analysis of LLMs

Local Interpretable Model-agnostic Explanations (LIME) is a powerful technique for elucidating the predictions of complex machine learning models in a manner that is comprehensible to humans. The core of LIME lies in constructing a surrogate model that simplifies the original complex model, allowing for the reconstruction of the prediction process. By modifying the input data of the original model and generating multiple predictions, LIME trains the surrogate model to represent significant features that contribute to specific prediction outcomes. This approach enables a clearer understanding of the functioning of complex models through the lens of simpler models, such as linear regression or decision trees.

To evaluate the interpretability of the BERT-based sentiment analysis model, LIME was utilized to investigate a specific case in which BERT and Llama 3.2 CoT produced conflicting sentiment classifications. In this instance, the text "India sees Apple nearly tripling investment, exports in coming years—Reuters India" was classified as negative by BERT, whereas Llama 3.2 CoT assigned a positive sentiment. The LIME analysis, illustrated in Figure 12, highlights the words that significantly contributed to BERT's negative classification. As shown, the words "Reuters", "nearly", and "years" had the most substantial impact on the model's decision. Specifically, "Reuters", a well-known news agency that frequently reports crises, contributed the highest weight of +0.20 to the negative sentiment classification. Similarly, "nearly" was misinterpreted as carrying a negative connotation, potentially due to its frequent association with uncertainty in financial news. Overall, BERT classified the sentence as 70% negative, with only a 28% probability of being positive.

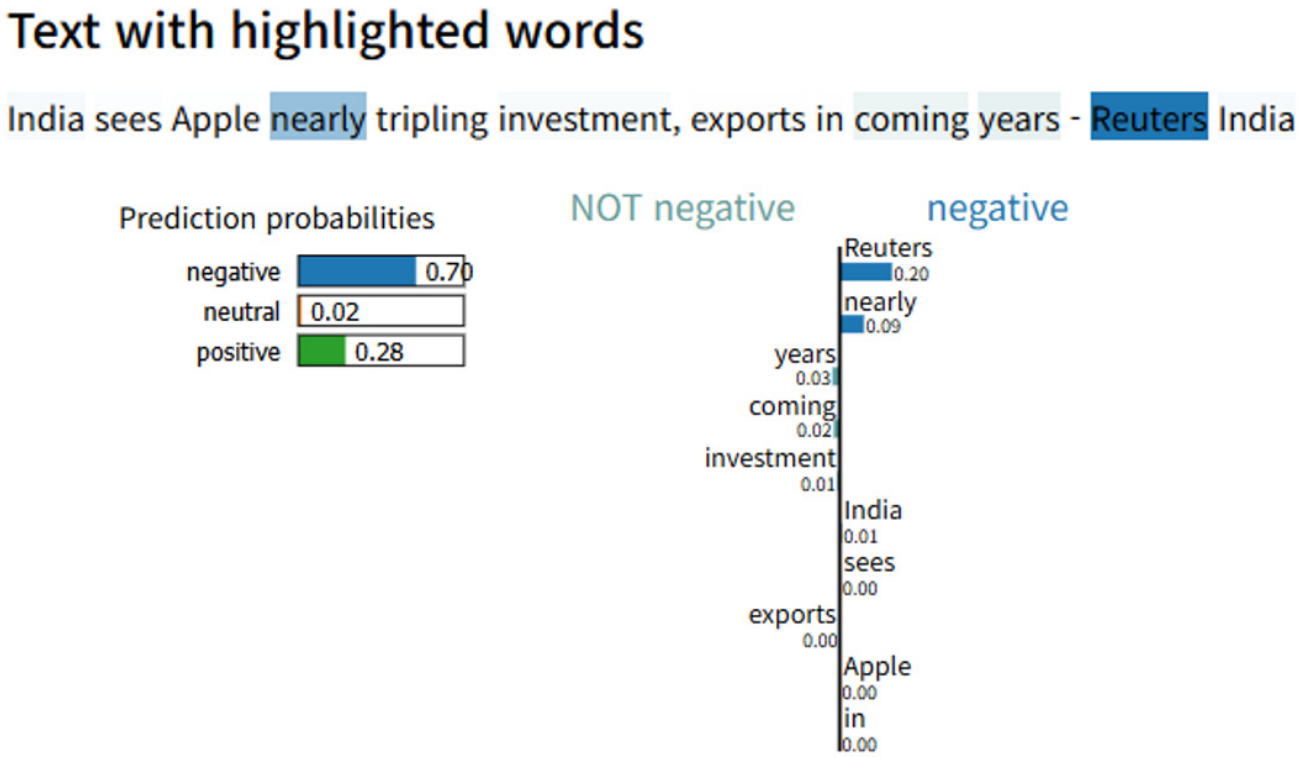


**Figure 12.** BERT LIME analysis.

In contrast, Llama 3.2 CoT demonstrated superior contextual understanding by recognizing the economic implications of Apple's increased investment in India. Figure 13 shows that the Llama 3.2 CoT model transcends mere word-level sentiment evaluation. Rather than being misled by isolated keywords, this model incorporated logical reasoning and identified the article's core message: "The article highlights a significant increase in Apple's investment and exports to India, indicating a growing partnership between the two countries. This suggests that the tone of the article is positive as it presents a beneficial development for both parties involved". This result underscores Llama 3.2 CoT's ability to integrate contextual meaning and structured reasoning into sentiment classification, thereby enhancing the reliability of financial sentiment analysis.

**Title:** India sees Apple nearly tripling investment, exports in coming years - Reuters India

**Symbol**: AAPL

**Sentiment:** Positive

**Reason**: The article highlights a significant increase in Apple's investment and exports to India, indicating a growing partnership between the two countries. This suggests that the tone of the article is positive as it presents a beneficial development for both parties involved.

**Figure 13.** Llama 3.2 CoT analysis.

These findings highlight the fundamental differences between traditional discriminative models, such as BERT, and generative models incorporating CoT reasoning, such as Llama 3.2 CoT. While BERT relies on predefined token-level sentiment patterns, it often fails to capture the broader contextual meaning of financial news articles. In contrast, Llama 3.2 CoT effectively leverages logical reasoning and contextual awareness to derive sentiment assessments that align more closely with human interpretations. This study demonstrates that explainability techniques such as LIME can reveal underlying biases in sentiment models and provide deeper insights into their decision-making processes. By integrating models capable of context-aware reasoning, sentiment analysis systems can enhance their reliability and interpretability, which is particularly crucial in financial applications where sentiment-driven decisions impact investment strategies and risk assessments.

Furthermore, regulatory frameworks such as the EU AI Act emphasize core requirements for AI products, including transparency, safety, ethical compliance, and the prevention of fundamental rights violations. In this context, explainability techniques like LIME can play a crucial role in ensuring that AI-driven sentiment analysis models operate with greater clarity and transparency, particularly when their classification decisions are challenging to express through human language and reasoning.

Beyond enhancing interpretability, the ability of models to analyze contextual meaning can be leveraged in concrete applications, such as data filtering processes or reinforcement learning mechanisms where rewards and penalties are weighted accordingly. By integrating these approaches, AI-driven sentiment analysis models can not only improve reliability of their decision-making but also contribute to their future development, refining their ability to process financial news and market dynamics with greater accuracy and depth.

## 7. Conclusions

We established and compared portfolios in the US stock market based on sentiment analysis derived from various LLM models. The results demonstrate that sentiment analysis-driven investment strategies significantly outperformed the market index, highlighting their efficacy as robust tools for formulating investment strategies. Notably,

specialized models such as Llama 3.1 showed exceptional performance, underscoring the importance of using models that capture the linguistic and cultural nuances of the data to develop market-optimized investment strategies.

In the context of the US market, the long strategy exhibited outstanding performance during bullish market conditions, with advanced prompting techniques such as SuperICL playing a crucial role in enhancing the model efficacy. Our findings confirm that portfolios constructed using LLM-based news sentiment analysis not only showed higher profitability than the market index but also adeptly captured the fluctuating impacts of positive and negative news sentiment in relation to market movements. This indicates that advanced prompting techniques contribute substantially to improving investment performance in the US market.

Furthermore, we anticipate that the integration of additional multimodal data, encompassing macroeconomic indicators, industry news, and visual data such as chart images, could further enhance investment performance. This suggests that LLM-based investment strategies can effectively address market complexities by leveraging a broader array of data sources beyond mere text sentiment analysis. Incorporating these diverse data modalities could pave the way for more sophisticated and responsive investment strategies that are better equipped to navigate the dynamic financial market landscape. Additionally, future research could explore the dynamic adjustment of long-short portfolio allocations based on market conditions, utilizing indicators such as RSI, %B, and VIX to modify the weighting of long and short positions accordingly. By adapting portfolio allocations to bull markets—where long positions tend to perform better—or to bearish conditions that favor short positions, such strategies could further optimize returns. Future research could also consider factors related to trading costs, such as turnover rates. While this study assumes a daily trading framework, trading costs are likely to be a significant factor in the implications of LLM-based sentiment analysis for long-term portfolio management. Therefore, future work should explicitly model and analyze the impact of these costs on portfolio performance.

**Author Contributions:** Conceptualization, Y.M. and N.K.; methodology, Y.M. and N.K.; software, Y.M.; validation, Y.M. and N.K.; formal analysis, Y.M.; investigation, Y.M. and N.K.; resources, N.K.; data curation, Y.M.; writing—original draft preparation, Y.M. and N.K.; writing—review and editing, Y.M. and N.K.; visualization, Y.M.; supervision, N.K.; project administration, N.K.; funding acquisition, N.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT), grant number No. 2021R1F1A1050602.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data and source code are available at https://github.com/examplemoon/LLM_Portfolio (accessed on 27 February 2025).

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Appendix A

All analysis results were stored in clearly defined formats to ensure reproducibility and were actively used in the portfolio design and performance evaluation processes employed in the study. The code developed during the research was made publicly

available at https://github.com/examplemoon/LLM_Portfolio (accessed on 20 January 2025) to facilitate future research endeavors.

The analysis environment was built using various Python 3.11.11 libraries to support the entire workflow, including data processing, sentiment analysis, utilization of generative LLMs, and result visualization. Key libraries included pandas = 2.2.3 and numpy = 1.26.4 for data manipulation and analysis, and gnews = 0.3.8 was utilized to collect financial news data. For sentiment analysis, transformers = 4.45.1 was employed to integratively apply financial specialized models like BERT, FinBERT, and RoBERTa-Finance, as well as the latest generative LLMs based on OpenAI. Stock data collection was performed using yfinance = 0.2.43, and the visualization of research results was carried out with matplotlib = 3.9.2. Additionally, interaction with LLMs was integrated into the research using the ollama = 0.3.3 library to incorporate the functionalities of various generative LLMs.

## References


1. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *J. Financ.* **2007**, *62*, 1139–1168. [CrossRef]
2. Alanyali, M.; Moat, H.S.; Preis, T. Quantifying the relationship between financial news and the stock market. *Sci. Rep.* **2013**, *3*, 3578. [CrossRef] [PubMed]
3. Bollen, J.; Mao, H.; Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of the International AAAI Conference on Web and Social Media, Barcelona, Spain, 17–21 July 2011; Volume 5, pp. 450–453.
4. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
5. Chan, W.S. Stock price reaction to news and no-news: Drift and reversal after headlines. *J. Financ. Econ.* **2003**, *70*, 223–260. [CrossRef]
6. Malo, P.; Sinha, A.; Takala, P.; Ahlgren, O.; Lappalainen, I. Learning the roles of directional expressions and domain concepts in financial news analysis. In Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX, USA, 7–10 December 2013; pp. 945–954.
7. Brown, T.B. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
8. Deng, X.; Bashlovkina, V.; Han, F.; Baumgartner, S.; Bendersky, M. LLMs to the Moon? Reddit Market Sentiment Analysis with Large Language Models. In Proceedings of the Companion Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May 2023; pp. 1014–1019.
9. Lopez-Lira, A.; Tang, Y. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv* **2023**, arXiv:2304.07619. [CrossRef]
10. Jang, E.; Choi, H.; Lee, H. Stock prediction using combination of BERT sentiment analysis and macro economy index. *J. Korea Soc. Comput. Inf.* **2020**, *25*, 47–56.
11. Bendi-Ouis, Y.; Dutarte, D.; Hinaut, X. Deploying Open-Source Large Language Models: A Performance Analysis. *arXiv* **2024**, arXiv:2409.14887.
12. Konstantinidis, T.; Iacovides, G.; Xu, M.; Constantinides, T.G.; Mandic, D. FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications. *arXiv* **2024**, arXiv:2403.12285.
13. Mai, Z.; Zhang, J.; Xu, Z.; Xiao, Z. Financial sentiment analysis meets Llama 3: A comprehensive analysis. In Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI), Osaka, Japan, 2–4 August 2024; pp. 171–175.
14. Brown, P.F.; Della Pietra, V.J.; Desouza, P.V.; Lai, J.C.; Mercer, R.L. Class-based n-gram models of natural language. *Comput. Linguist.* **1992**, *18*, 467–480.
15. Cavnar, W.B.; Trenkle, J.M. N-gram-based text categorization. In Proceedings of the SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, USA, 11–13 April 1994; pp. 161–175.
16. Kondrak, G. N-gram similarity and distance. In *International Symposium on String Processing and Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 115–126.
17. Li, Y.H.; Jain, A.K. Classification of text documents. *Comput. J.* **1998**, *41*, 537–546. [CrossRef]
18. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998; Springer: Berlin/Heidelberg, Germany, 1998; pp. 137–142.

19. Lee, J.Y.; Dernoncourt, F. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv* **2016**, arXiv:1603.03827.
20. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
22. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *OpenAI* **2018**. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 18 April 2025).
23. Ghatoura, P.S.; Hosseini, S.E.; Pervez, S.; Iqbal, M.J.; Shaukat, N. Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM. *Big Data Cogn. Comput.* **2024**, *8*, 199. [CrossRef]
24. Wawer, M.; Chudziak, J.A.; Niewiadomska-Szynkiewicz, E. Large Language Models and the Elliott Wave Principle: A Multi-Agent Deep Learning Approach to Big Data Analysis in Financial Markets. *Appl. Sci.* **2024**, *14*, 11897. [CrossRef]
25. Delgadillo, J.; Kinyua, J.; Mutigwe, C. FinSoSent: Advancing Financial Market Sentiment Analysis through Pretrained Large Language Models. *Big Data Cogn. Comput.* **2024**, *8*, 87. [CrossRef]
26. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.D.L.; Sayed, W.E. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
27. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Kenealy, K. Gemma: Open models based on gemini research and technology. *arXiv* **2024**, arXiv:2403.08295.
28. Onwuegbuche, F.C.; Wafula, J.M.; Mung'atu, J.K. Support Vector Machine for Sentiment Analysis of Nigerian Banks' Financial Tweets. *J. Data Anal. Inf. Process.* **2019**, *7*, 153–170. [CrossRef]
29. Antweiler, W.; Frank, M.Z. Is all that talk just noise? The information content of internet stock message boards. *J. Financ.* **2004**, *59*, 1259–1294. [CrossRef]
30. Sun, Y.; Liu, X.; Chen, G.; Hao, Y.; Zhang, Z.J. How mood affects the stock market: Empirical evidence from microblogs. *Inf. Manag.* **2020**, *57*, 103181. [CrossRef]
31. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213.
32. Xu, C.; Xu, Y.; Wang, S.; Liu, Y.; Zhu, C.; McAuley, J. Small models are valuable plug-ins for large language models. *arXiv* **2023**, arXiv:2305.08848.
33. Wu, R. Portfolio Performance Based on LLM News Scores and Related Economical Analysis. *SSRN*: 4709617, **2024**. Available online: http://dx.doi.org/10.2139/ssrn.4709617 (accessed on 18 April 2025).
34. Sahoo, A.; Chanda, R.; Das, N.; Sadhukhan, B. Comparative Analysis of BERT Models for Sentiment Analysis on Twitter Data. In Proceedings of the 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, India, 17–19 August 2023; pp. 658–663.
35. Liu, Z.; Huang, D.; Huang, K.; Li, Z.; Zhao, J. FinBERT: A pre-trained financial language representation model for financial text mining. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 4513–4519.
36. Soleimanian, M. Board Environmental, Social, and Governance (ESG) Expertise and the Usefulness of ESG Reports. McGill University, **2024**. Available online: https://www.mcgill.ca/desautels/files/desautels/board_environmental_social_and_governance.pdf (accessed on 18 April 2025).