Can ChatGPT Forecast Returns in the Korean Stock Market?

Multi-Dimensional Evidence from News-Based Long-Short Strategies

Sungwoo Park[1], **Donghyun Lim**[2], **Hyoung-Goo Kang**[3]

[1]AI Research Center, Shinhan Bank; Ph.D. Candidate, Hanyang University

[2]Undergraduate Student, Department of Computer Science, Hanyang University

[3]Professor, Department of Finance and Computational Engineering, Hanyang University

(Corresponding Author)

July 9, 2025

Abstract

This study evaluates the predictive power of news sentiment analysis using the GPT-4.1 Nano model in the Korean stock market. Approximately 500,000 news articles from *Maeil Business Newspaper*, collected between 2022 and March 2025, were combined with daily stock price data from FnGuide to assess the effectiveness of a long-short investment strategy across multiple dimensions. Sentiment scores were derived from news headlines and categorized by time of news release (pre-market, intra-day, post-market, and holidays) to analyze realistic trading performance. While the overall strategy demonstrated stable cumulative returns for long positions, short positions and combined long-short strategies experienced structural losses during specific periods. Detailed analyses revealed that large-cap stocks showed the most favorable performance by firm size, and growth stocks were particularly responsive to sentiment analysis strategies by investment style. Additionally, significant performance variations across industry sectors indicated the necessity of sector-specific strategic approaches. This research suggests that GPT-based sentiment analysis strategies are conditionally effective in the Korean stock market, emphasizing the importance of strategic filtering and position management.

Keywords: Sentiment Analysis, GPT-4.1 Nano, Korean Stock Market, Long-Short Strategy, Industry Style

1. Introduction

Investor sentiment plays a crucial role in financial markets, often influencing asset prices through the transmission of news and media information. Traditional approaches to quantifying sentiment have relied heavily on textual analysis. For instance, Tetlock (2007) measured the proportion of negative words in *Wall Street Journal* columns and found a significant inverse relationship between negative sentiment and next-day stock returns. Such findings suggest that sentiment-based news signals may contain predictive information, a phenomenon also explained by behavioral finance theories such as Barberis et al. (1998). Subsequent studies, such as Tetlock et al. (2008), further incorporated the tone and context of words to explain firm fundamentals, while Li (2010) demonstrated that forward-looking textual statements can signal future profitability.

The advancement of deep learning has significantly enhanced the capacity to interpret unstructured financial text data. By the mid-2010s, researchers began vectorizing news content using deep neural networks to forecast price movements Huang et al. (2016); Ding et al. (2015). The introduction of Transformer-based models in 2017 marked a turning point in natural language processing (NLP), dramatically improving the contextual understanding of text Devlin et al. (2019). The release of ChatGPT by OpenAI in 2022 accelerated the integration of large language models (LLMs) into finance. Trained on massive text corpora, LLMs can capture subtle nuances and contextual relationships that traditional models struggled to identify, potentially enabling more accurate forecasting of market reactions to news.

A seminal study by Lopez-Lira and Tang (2023) demonstrated the use of ChatGPT to classify individual stock news headlines as positive, neutral, or negative in the U.S. equity market. Their results revealed that the sentiment scores generated by GPT-3.5 and GPT-4 significantly outperformed traditional sentiment dictionaries and earlier NLP models (e.g., GPT-1, GPT-2, BERT). Stocks classified as "positive" experienced significantly higher returns on the following day, whereas "negative" stocks underperformed. A long-short portfolio constructed from these signals yielded cumulative returns exceeding 550% within a year, with a Sharpe ratio above 3, far surpassing the market index, which incurred losses over the same period. These results highlight the considerable potential of modern LLMs for financial forecasting using unstructured textual data.

Despite the promising evidence from U.S. markets, the application of GPT-based sentiment

analysis in the Korean equity market remains in its infancy. Structural differences—such as the linguistic characteristics of Korean financial news, high retail investor participation, and short-selling restrictions—make it uncertain whether findings from U.S. studies will generalize to Korea. Some recent domestic research has shown the value of Korean-language sentiment indices. For example, the Korean News Financial Sentiment Index (KNFSI), constructed using FinBERT, has been found to correlate significantly with the KOSPI 200 Index, particularly during downturns Lee and Kwon (2025). However, these studies focus on aggregate market indicators and do not directly explore firm-level returns or tradable strategies.

While some Korean studies have employed sentiment dictionaries or FinBERT models for price prediction Kang and Choi (2025), the direct application of GPT-based LLMs to stock-level news sentiment in Korea has not been widely explored. To address this gap, this study applies GPT-based sentiment classification to Korean-language financial news to evaluate its effectiveness in predicting returns through long-short strategies. Unlike prior research focusing solely on aggregate market behavior, we examine performance across multiple dimensions: firm size, investment style (value vs. growth), and industry sectors.

Inspired by Lopez-Lira and Tang (2023), who observed stronger predictive signals for small-cap stocks and negative news in the U.S., we assess whether similar patterns emerge in the Korean context or whether market-specific features lead to different outcomes. By incorporating style indices from KODEX and industry classifications from KRX, we provide a comprehensive assessment of the generalizability and conditional effectiveness of GPT-based sentiment strategies.

The remainder of the paper is organized as follows: Section 2 reviews the relevant literature. Section 3 details the data and methodology. Section 4 presents the empirical results across various dimensions, and Section 5 concludes with key findings and practical implications.

2. Related Literature

The relationship between media information and asset prices has long been a subject of academic inquiry. Tetlock (2007) was among the first to show that a quantified sentiment index derived from newspaper column tone could predict market returns. Subsequent studies reinforced this finding, showing that increased negative sentiment in financial news is associated with higher trading volumes and greater return volatility.

At the firm level, numerous studies have demonstrated that the tone of news articles—particularly

in terms of positive or negative sentiment—can influence short-term stock movements. Heston and Sinha (2017) proposed models that use news text to predict firm-specific returns, while Kim et al. (2012) applied sentiment analysis to Korean online news to construct intelligent trading systems. In a related stream, Hanley and Hoberg (2010) showed that textual disclosures in IPO filings contain material predictive information about firm performance. However, many of these earlier studies relied heavily on sentiment dictionaries (e.g., lists of positive and negative words) and basic bag-of-words methods, which lack contextual understanding. As Loughran and McDonald (2011) pointed out, general-purpose sentiment dictionaries may be ill-suited for financial contexts, highlighting the need for domain-specific resources.

The 2010s witnessed the integration of more advanced natural language processing (NLP) techniques into sentiment analysis. With the rise of machine learning, researchers began using supervised models and vectorized representations of text, such as Word2Vec and Doc2Vec, to analyze financial news. Huang et al. (2016), for example, utilized neural network-based models to examine the relationship between news sentiment and price direction, demonstrating that deep learning models could outperform traditional statistical methods. Ding et al. (2015) introduced event-driven architectures to extract meaningful signals from text, using deep learning to infer causal links between headlines and stock movements. While these models were powerful, they were often narrow in scope, lacking the broad knowledge and flexibility of today's large-scale language models.

The emergence of large language models (LLMs) marked a turning point in financial text analysis. With the release of GPT-3 (175 billion parameters) by OpenAI, finance researchers began exploring its utility in investment applications. Early attempts included prompt-based financial question answering and structured adaptations like TabGPT for time-series prediction. However, the launch of ChatGPT in 2022 sparked explosive interest in using pre-trained LLMs for specialized tasks through prompt engineering. Despite not being fine-tuned on financial data, ChatGPT demonstrated a remarkable ability to understand news context and assess its market implications, as highlighted by Lopez-Lira and Tang (2023).

Their study systematically compared the forecasting accuracy of various LLMs, including GPT-1, GPT-2, GPT-3.5, GPT-4, and BERT. Only the latest GPT-based models delivered statistically significant predictive power for stock returns. Using data from the U.S. stock market (October 2021 to December 2022), they showed that a GPT-3.5-powered long-short strategy achieved a Sharpe ratio of 4.4—substantially outperforming strategies based on traditional sentiment

dictionaries such as Loughran and McDonald (2011). Similarly, Kirtac and Germano (2024) analyzed 960,000 U.S. news headlines and reported that an OPT model based on GPT-3 outperformed FinBERT and other domain-specific models in both classification accuracy (74.4%) and portfolio performance (Sharpe ratio of 3.05 vs. 1.23).

In Korea, studies using sentiment-based strategies are still developing. The KNFSI index, which uses FinBERT to evaluate macro-level market sentiment, found significant correlations with the KOSPI 200 Index Lee and Kwon (2025), especially in down markets. At the micro-level, some research has explored headline-based trading strategies, but these efforts have often been limited by the relatively underdeveloped NLP resources for the Korean language. Recently, domain-specific models such as KR-FinBERT have emerged, fine-tuned on Korean financial texts using the BERT architecture Devlin et al. (2019). While these models have improved sentiment classification accuracy, direct application of GPT-class LLMs in Korean finance remains rare.

In summary, prior literature consistently suggests that news sentiment carries predictive signals for asset prices. Recent advances in GPT-based models further amplify this potential, offering enhanced interpretability and predictive power. Building on this foundation, our study applies GPT-based sentiment analysis to the Korean stock market and rigorously evaluates its forecasting performance through a multi-dimensional lens.

3. Data and Methodology

3.1. Data and Analytical Framework

This study combines Korean stock market data and news headlines from January 2022 to March 2025. Approximately 500,000 news articles were sourced from *Maeil Business Newspaper* via BigKinds, a public news archive. Stock data—including daily open and close prices—were obtained from FnGuide. The sample includes firms listed on KOSPI and KOSDAQ that are covered by FnGuide, excluding REITs, KONEX stocks, and non-traded holding companies.

Figure **??** illustrates the overall process. News articles were cleaned and classified, after which a sentiment score was assigned to each headline using a GPT model. These sentiment scores were then matched with the relevant company's stock data to determine trading positions. A long-short portfolio was constructed daily, and cumulative performance was calculated across the entire sample period. The strategy was further evaluated across subgroups by firm size,

investment style, and industry.

3.2. Data Preprocessing

News data included metadata such as publication date, title, content, and category. Because BigKinds does not provide time stamps, each article's URL was crawled to extract the exact publication time. Based on this, articles were classified into four groups: pre-market (PRE), intra-day (IN), post-market (AFTER), and holiday (DAY-OFF) news. For example, articles published before 9:00 a.m. on trading days were labeled as pre-market.

To improve financial relevance, only articles classified under the "Economy" category were retained. Headlines that referred to non-listed entities (e.g., government agencies, unlisted firms) were excluded. To ensure accurate entity recognition, we employed GPT-4's Named Entity Recognition (NER) capabilities to map company mentions in headlines to their corresponding stock tickers. Ambiguous cases were resolved by referencing firm descriptions in FnGuide, assisted by GPT-3.5 Turbo. This AI-assisted approach significantly reduced manual error and improved processing speed.

Finally, the sentiment score for each firm on each day was aggregated across multiple articles. If no article mentioned a firm on a given day, it was excluded from that day's portfolio. The matching of sentiment to stock price used appropriate open or close prices based on the news release timing, as discussed below.

3.3. Sentiment Analysis Using GPT

The core of the analysis involved sentiment classification of news headlines using OpenAI's GPT model. While GPT-4 API access was available, we opted for the lighter-weight GPT-4.1 Nano model (April 14, 2025 version) for efficiency. This model was found to perform well in understanding financial context in Korean.

The prompt posed to the model was structured as:

> "Does this news article have a positive or negative impact on [Company Name]'s stock price?"

Inspired by Lopez-Lira and Tang (2023), we instructed the model to respond with "Yes / No / Uncertain" followed by a one-sentence explanation. The responses were mapped to sentiment scores: "Yes" = +1, "No" = -1, "Uncertain or unclear" = 0.

The temperature was set to 0 to ensure consistent outputs. Although English translation was considered, the model was able to handle Korean headlines effectively, so the original text was used. To reduce ambiguity, we appended a brief context—e.g., industry and product information—for the named firm within the prompt.

The resulting sentiment scores ranged from -1 to +1. If multiple headlines for the same firm appeared on the same day, the average score was used. The overall distribution skewed slightly positive (mean $\approx$ +0.2), consistent with prior studies. Whether these scores have predictive power is tested in the following section.

3.4. Construction of News-Based Long-Short Strategies

To translate sentiment signals into investable strategies, we constructed daily rebalanced long-short portfolios, following frameworks established by Lopez-Lira and Tang (2023) and Kelly et al. (2021). The basic rule is to long (buy) stocks with positive news and short (sell) those with negative news. All positions are equally weighted and constructed to be self-financing (i.e., net capital exposure = 0), minimizing market directionality and emphasizing relative performance. Position timing depends on the news publication time, with the following rules:

- **PRE (pre-market)**: If a firm receives positive news before market open, the position is entered at the day's open and exited at the close. Negative news triggers a short position at open, closed at the same day's close.
- **IN (intra-day)**: For news released during market hours, we assume that traders cannot respond immediately. Positions are entered at the same day's close and exited at the following day's close.
- **AFTER (post-market)**: For news released after the market closes, positions are entered at the next day's open and exited at the same day's close.
- **DAY-OFF (holiday/weekend)**: Treated identically to post-market news.

This results in a short-horizon trading strategy that closes all positions within one or two trading days. Portfolios are rebalanced daily based on updated news sentiment. In cases where insufficient stocks are available on one side (e.g., few short candidates), the portfolio may temporarily deviate from full neutrality, though this was rare during the sample period.

We evaluated performance using cumulative return, annualized return, volatility, Sharpe and Sortino ratios, and maximum drawdown. Subgroup analyses by size, style, and sector provided additional insights into the robustness and contextual effectiveness of the strategy.

4. Empirical Results

4.1. Overall Performance of the GPT Long-Short Strategy

We begin by evaluating the overall performance of the GPT-based long-short strategy over the full sample period from January 2, 2022, to March 31, 2025. As shown in Table 1, the strategy generated a cumulative return of 11.91%, with an annualized return of 4.14% and a Sharpe ratio of 0.38, outperforming benchmark portfolios such as the equal-weighted portfolio (1.73%, Sharpe 0.19) and the KOSPI index (-14.43%, Sharpe -0.19). The GPT strategy also exhibited favorable downside risk metrics, including a Sortino ratio of 0.45 and a maximum drawdown of -20.74%, suggesting effective risk-adjusted performance.

Disaggregated analysis reveals that the short-only component outperformed the long-only component. The short-only strategy yielded a 6.17% cumulative return with a Sharpe ratio of 0.224, while the long-only strategy achieved a more modest 4.04% return and a Sharpe ratio of 0.186. This asymmetry indicates that negative news had a stronger impact on stock returns in the Korean market, mirroring the overreaction patterns observed in U.S. markets by Lopez-Lira and Tang (2023).

In contrast, the equal-weighted benchmark posted a marginal gain (1.73%) with weak risk-adjusted returns, while the KOSPI index suffered a substantial decline (-14.43%) with high volatility (annualized volatility of 17.92%). These findings underscore the alpha-generating potential of GPT-based sentiment strategies, particularly given their market-neutral construction.

In summary, the GPT-based long-short strategy delivered statistically and economically significant performance in the Korean equity market. The strategy's effectiveness appears to stem primarily from the predictive power of short positions, especially in response to negative sentiment, supporting the view that market inefficiencies persist in processing adverse news.

4.2. Performance by Firm Size

To examine the heterogeneity in performance across different firm sizes, we divide the sample into large-cap and small-to-mid-cap (SMID-cap) groups. Firms with a market capitalization

Table 1: Overall Performance of the GPT Long-Short Strategy

Metrics	GPT_LongShort	GPT_Long	GPT_Short	EqualWeighted	KOSPI
Cumulative Return	11.91%	4.04%	6.17%	1.73%	-14.43%
Annualized Return	4.14%	1.60%	2.50%	0.39%	-3.31%
Annualized Volatility	10.86%	8.58%	11.17%	2.09%	17.92%
Sharpe Ratio	0.38	0.19	0.22	0.19	-0.19
Sortino Ratio	0.45	0.27	0.27	0.23	-0.26
Maximum Drawdown	-20.74%	-14.84%	-24.01%	-5.06%	-27.49%
Skewness	-1.15	-0.08	-1.02	-0.90	-0.57
Kurtosis	9.80	4.91	7.75	9.34	5.17

above KRW 1 trillion are categorized as large-cap, while those below this threshold are classified as SMID-cap. Long-short portfolios are constructed separately for each group using GPT sentiment signals.

As shown in Table 2, the GPT Large-Cap Long-Short Strategy achieved a cumulative return of 26.13% and an annualized return of 8.03%, with a Sharpe ratio of 0.75 and Sortino ratio of 0.96—the strongest risk-adjusted performance among all strategy variants. The strategy significantly outperformed both the equal-weighted benchmark (6.04%, Sharpe 0.44) and the large-cap ETF benchmark (-8.13%, Sharpe -0.06), confirming its ability to generate substantial alpha.

Within the large-cap group, most of the profits came from short positions. The GPT Large-Cap Short Strategy delivered a 20.14% cumulative return (Sharpe 0.60), while the long-only strategy showed modest performance (3.58%, Sharpe 0.17). This suggests that negative sentiment had a particularly strong downward impact on prices of large-cap stocks, likely due to greater liquidity and the availability of short-selling.

Table 2: Large-cap Performance of the GPT Long-Short Strategy

Metrics	GPT_LongShort	GPT_Long	GPT_Short	EqualWeighted	KOSPI
Cumulative Return	26.13%	3.58%	20.14%	6.04%	-8.13%
Annualized Return	8.03%	1.51%	6.43%	1.30%	-0.81%
Annualized Volatility	10.70%	9.09%	10.70%	2.95%	14.33%
Sharpe Ratio	0.75	0.17	0.60	0.44	-0.06
Sortino Ratio	0.96	0.24	0.74	0.59	-0.07
Maximum Drawdown	-12.12%	-12.92%	-15.47%	-5.66%	-26.60%
Skewness	-1.19	0.38	-2.16	-0.29	-0.43
Kurtosis	28.46	8.31	36.81	4.76	3.70

In contrast, Table 3 reports that SMID-cap stocks showed more limited performance. The GPT SMID-Cap Long-Short Strategy yielded a cumulative return of 7.85% with a Sharpe ratio of 0.22. The short-only strategy generated 10.55% (Sharpe 0.27), while the long-only component

produced a loss of -4.11%. Volatility levels were higher for SMID caps, with annualized volatility reaching 17–19%.

Table 3: SMID-cap Performance of the GPT Long-Short Strategy

Metrics	**GPT_LongShort**	**GPT_Long**	**GPT_Short**	**EqualWeighted**	**ETF**
Cumulative Return	7.85%	-4.11%	10.55%	-12.28%	3.67%
Annualized Return	4.26%	-0.41%	4.68%	-2.61%	1.94%
Annualized Volatility	19.17%	13.35%	17.19%	5.37%	15.11%
Sharpe Ratio	0.22	-0.03	0.27	-0.49	0.13
Sortino Ratio	0.28	-0.05	0.30	-0.57	0.15
Maximum Drawdown	-18.90%	-23.81%	-28.13%	-19.53%	-28.19%
Skewness	-0.21	1.66	-1.18	0.10	-0.56
Kurtosis	10.65	20.27	12.41	36.80	3.96

Interestingly, these findings diverge from Lopez-Lira and Tang (2023), who observed stronger sentiment predictability for small-cap stocks in the U.S. market due to greater information asymmetry. In the Korean market, however, our results suggest that GPT-based sentiment analysis is more effective for large-cap stocks, potentially due to fewer short-selling restrictions, higher institutional participation, and more efficient news dissemination.

In conclusion, the GPT sentiment strategy produced more stable and profitable outcomes in the large-cap segment. This implies that practical implementation of such strategies should prioritize large-cap stocks, while treating SMID caps as supplementary components with additional volatility and execution risks.

4.3. Performance by Investment Style

This section investigates whether the GPT-based sentiment strategy performs differently depending on investment style—specifically between value and growth stocks. Stocks are categorized using the underlying indices of KODEX Value and KODEX Growth ETFs, and separate long-short portfolios are constructed for each group.

As shown in Table 4, the GPT Value Long-Short Strategy achieved a cumulative return of 23.03% and an annualized return of 7.08%, with a Sharpe ratio of 0.74 and a Sortino ratio of 1.13. These results indicate solid risk-adjusted returns compared to benchmarks such as the equal-weighted portfolio (1.86%) and the KODEX Value ETF (-11.22%).

However, a decomposition of strategy components reveals a stark asymmetry. The long-only strategy for value stocks generated a loss of -7.63% (Sharpe -0.24), while the short-only strategy achieved a return of 31.63% (Sharpe 0.98). This suggests that GPT sentiment scores were

particularly effective at identifying value stocks likely to underperform due to negative news, providing strong downside protection.

Table 4: Value Performance of the GPT Long-Short Strategy

Metrics	**GPT_LongShort**	**GPT_Long**	**GPT_Short**	**EqualWeighted**	**ETF**
Cumulative Return	23.03%	-7.63%	31.63%	1.86%	-11.22%
Annualized Return	7.08%	-2.06%	9.33%	0.43%	-1.71%
Annualized Volatility	9.57%	8.55%	9.56%	2.78%	13.06%
Sharpe Ratio	0.74	-0.24	0.98	0.16	-0.13
Sortino Ratio	1.13	-0.34	1.56	0.20	-0.14
Maximum Drawdown	-14.78%	-16.51%	-16.13%	-5.22%	-27.64%
Skewness	0.51	0.09	0.43	-0.47	-0.48
Kurtosis	6.48	5.78	3.12	5.69	5.52

The growth stock portfolio produced even more impressive results. The GPT Growth Long-Short Strategy recorded a cumulative return of 58.28%, an annualized return of 16.32%, and a Sharpe ratio of 1.17—the highest among all analyzed strategies. The short-only strategy contributed most of the profits with 53.31% cumulative return (Sharpe 1.13), while the long-only strategy delivered a modest gain of 1.43% (Sharpe 0.10), as shown in Table 5.

Table 5: Growth Performance of the GPT Long-Short Strategy

Metrics	**GPT_LongShort**	**GPT_Long**	**GPT_Short**	**EqualWeighted**	**ETF**
Cumulative Return	58.28%	1.43%	53.31%	12.38%	-16.03%
Annualized Return	16.32%	1.06%	15.10%	2.62%	-2.48%
Annualized Volatility	13.95%	11.14%	13.37%	4.02%	15.92%
Sharpe Ratio	1.17	0.10	1.13	0.65	-0.16
Sortino Ratio	1.59	0.14	1.43	0.94	-0.19
Maximum Drawdown	-19.24%	-16.27%	-23.36%	-7.82%	-33.81%
Skewness	-0.51	-0.03	-1.01	0.43	-0.21
Kurtosis	7.96	4.60	12.87	6.85	2.70

These findings align with behavioral finance theory, which holds that growth stocks are more sensitive to investor sentiment due to their reliance on future expectations and narratives Barberis et al. (1998). The GPT model was thus particularly effective at capturing both upside and downside opportunities in growth stocks. Conversely, value stocks tend to be anchored in fundamentals, limiting their short-term sensitivity to news—except in negative cases, where sentiment signals helped identify underperformance.

In summary, the sentiment strategy was significantly more effective in the growth segment, where both long and short positions benefitted from high price responsiveness. In contrast, value stocks were best approached through short-focused sentiment strategies, given their relatively muted reaction to positive news.

4.4. Performance by Industry Sector

This section analyzes the performance of the news-based GPT strategy across various industry sectors. For all 26 sectors classified by the Korea Exchange (KRX), we evaluate the GPT long-short strategy over the period from 2022 to 2025 using a range of performance metrics, including cumulative and annualized returns, Sharpe ratio, Sortino ratio, maximum drawdown, skewness, and kurtosis.

The GPT strategy recorded positive cumulative returns in 12 out of 26 sectors, supporting its broad applicability and overall effectiveness. Among them, the Chemicals, Textiles & Apparel, Insurance, and Construction sectors achieved cumulative returns exceeding 20%, demonstrating strong alpha-generating potential.

The most notable performance was observed in the Chemicals sector, where the GPT strategy delivered a cumulative return of 106.74%, with a Sharpe ratio of 1.26 and a Sortino ratio of 1.78—indicating excellent risk-adjusted performance. This suggests that the GPT model effectively captured dynamic information flows in the chemical industry.

The Textiles & Apparel sector also produced solid results, with a cumulative return of 69.94% and a Sharpe ratio of 1.13. Despite annualized volatility of 17.34%, the strategy showed consistent profitability. Furthermore, the sector's maximum drawdown remained relatively modest at -22.03%, indicating good downside risk management.

In the Insurance sector, the GPT strategy delivered a cumulative return of 62.34% with a Sortino ratio of 1.44. This implies that the strategy maintained strong performance despite high volatility, and that news-based sentiment signals can also be effective in traditionally conservative sectors.

The Construction sector posted a cumulative return of 58.05%. Although it experienced higher volatility (24.70% annualized), the strategy still achieved meaningful absolute returns, underscoring its robustness in cyclical industries.

Other notable sectors with positive returns included Retail Trade (32.48%), Other Manufacturing (23.42%), and Entertainment & Culture (19.99%). These results suggest that the GPT-based sentiment strategy can extract informational value across a wide range of industries beyond just a few sectors.

In contrast, the remaining 14 sectors exhibited negative cumulative returns under the GPT strategy. Notably, the Machinery & Equipment (-41.65%), Paper & Wood (-62.88%), Electricity & Gas (-67.25%), and General Services (-70.16%) sectors consistently showed negative alpha.

These sectors also displayed high volatility, negative skewness, and elevated kurtosis, suggesting structural challenges in translating news signals into profits, or a possible mismatch between GPT model capabilities and the information structure of those industries.

A detailed summary of the quantitative performance across all 26 sectors is provided in Table 9 in the Appendix. The next section offers a focused comparative analysis of key sectors—Chemicals, Insurance, and Construction—where benchmark ETFs are available.

4.4.1. Chemicals Sector: Outstanding Outperformance Driven by Short Positions

The GPT-based news sentiment strategy delivered the strongest performance in the Chemicals sector. As shown in Table 6, the cumulative return of the GPT long-short strategy over the sample period reached 106.74%, significantly outperforming both the equal-weighted portfolio benchmark (11.69%) and the sector ETF benchmark (-45.71%). This level of performance is not only impressive in absolute terms but also remarkable from a risk-adjusted perspective.

The strategy recorded a Sharpe ratio of 1.26 and a Sortino ratio of 1.78, indicating that sentiment signals extracted from news consistently produced statistically significant alpha. The maximum drawdown was -26.60%, demonstrating superior downside protection relative to the ETF benchmark, which experienced a drawdown of -55.93%. These results confirm the competitiveness of the GPT strategy in terms of risk management.

A decomposition of the strategy reveals that short positions were the primary driver of returns. The GPT Short strategy alone generated a cumulative return of 103.45% with a Sharpe ratio of 1.46, indicating that the model effectively captured negative sentiment and its impact on prices. In contrast, the long-only component produced a negligible return of -0.36%, suggesting limited contribution from positive news signals. This performance pattern implies that negative news was more frequent or impactful during the period, making short-selling the core mechanism of the strategy's alpha generation.

In summary, the GPT sentiment strategy in the Chemicals sector outperformed traditional benchmarks across all key dimensions—information efficiency, risk-adjusted return, and downside risk control. The high contribution of short signals highlights the model's sensitivity to negative news and its ability to implement an effective sell-side strategy based on sentiment analysis.

Table 6: Chemical Performance of the GPT Long-Short Strategy

Metrics	**GPT_LongShort**	**GPT_Long**	**GPT_Short**	**EqualWeighted**	**ETF**
Cumulative Return	106.74%	-0.36%	103.45%	11.69%	-45.71%
Annualized Return	28.30%	1.36%	26.58%	2.98%	-9.47%
Annualized Volatility	22.46%	17.09%	18.25%	8.18%	24.64%
Sharpe Ratio	1.26	0.08	1.46	0.37	-0.38
Sortino Ratio	1.78	0.11	1.64	0.45	-0.48
Maximum Drawdown	-26.60%	-34.02%	-27.86%	-17.58%	-55.93%
Skewness	0.23	-0.05	0.26	-0.58	-0.03
Kurtosis	6.13	6.48	11.89	5.63	4.52

4.4.2. Insurance Sector: Stable Long-Oriented Outperformance and Diversification Benefits

The GPT-based strategy also exhibited strong performance in the Insurance sector, characterized particularly by steady returns driven by long positions. As shown in Table 7, the cumulative return of the full GPT Long-Short strategy reached 62.34%, with an annualized return of 17.89%, outperforming both the ETF benchmark (52.12%) and the equal-weighted strategy (45.43%). The Sharpe ratio was 1.02, and the Sortino ratio was 1.44, indicating favorable risk-adjusted performance. The maximum drawdown stood at -15.58%, which lies between the ETF (-20.11%) and the equal-weighted benchmark (-7.96%).

A notable feature is that the GPT Long strategy alone recorded a cumulative return of 66.22%, significantly higher than the GPT Short strategy, which posted a loss of -2.73%. The long-only strategy also showed superior risk-adjusted returns, with a Sharpe ratio of 1.25, a Sortino ratio of 1.87, and a maximum drawdown of -12.20%, outperforming even the combined Long-Short portfolio. This suggests that positive sentiment was strongly reflected in the insurance sector and that the model effectively captured upward trends through long positions based on news signals.

In summary, the GPT strategy in the Insurance sector delivered stable and consistent excess returns primarily through long positions. This indicates that the sentiment-based signals were particularly responsive to positive news flows. In contrast, the contribution of short positions was minimal, possibly due to informational noise or excessive downside hedging.

4.4.3. Construction Sector: Short-Driven Alpha and Outperformance Amid High Volatility

The GPT-based news sentiment strategy also demonstrated significant outperformance in the Construction sector compared to traditional benchmark strategies. According to the results in Table 8, the GPT Long-Short strategy achieved a cumulative return of 58.05% and an annualized

Table 7: Insurance Performance of the GPT Long-Short Strategy

Metircs	**GPT_LongShort**	**GPT_Long**	**GPT_Short**	**EqualWeighted**	**ETF**
Cumulative Return	62.34%	66.22%	-2.73%	45.43%	52.12%
Annualized Return	17.89%	18.20%	-0.26%	11.06%	11.33%
Annualized Return	17.49%	14.58%	10.88%	9.26%	19.08%
Sharpe Ratio	1.02	1.25	-0.02	1.19	0.59
Sortino Ratio	1.44	1.87	-0.02	1.84	0.78
Maximum Drawdown	-15.58%	-12.20%	-23.47%	-7.96%	-20.11%
Skewness	0.72	1.28	0.55	1.70	0.37
Kurtosis	6.73	7.37	19.31	11.71	4.41

return of 18.82% from 2022 to 2025. This clearly surpassed the performance of the ETF strategy (-6.77%, 0.82%) and the equal-weighted strategy (25.62%, 6.14%) over the same period. A distinguishing feature of this sector is the dominant contribution from short positions.

The GPT Short strategy generated a cumulative return of 41.86%, accounting for the majority of total returns, whereas the Long strategy only achieved 8.37%. This suggests that the GPT model was particularly effective in identifying underperforming stocks and reacting sensitively to negative news. In terms of risk-adjusted performance, the Long-Short strategy recorded a Sharpe ratio of 0.76 and a Sortino ratio of 0.85—comparable to the equal-weighted strategy (Sharpe 0.72, Sortino 0.90), but far superior to the ETF benchmark (Sharpe 0.03, Sortino 0.05). On the risk side, significant volatility was observed. The annualized volatility of the Long-Short strategy was 24.70%, with a maximum drawdown of -34.45%, indicating a high level of exposure compared to other sectors. Furthermore, the skewness of -2.39 and kurtosis of 65.10 highlight extreme distribution characteristics, signaling the need for robust tail-risk management.

In summary, the GPT strategy in the Construction sector validated its alpha-generating ability through short positions, showing that sentiment analysis is particularly effective in capturing downside risks. However, elevated volatility and asymmetric distribution characteristics imply the necessity of complementary risk management tools when deploying the strategy in this sector.

5. Conclusion

This study empirically investigates the effectiveness of GPT-based sentiment analysis in the Korean stock market by constructing and evaluating long-short investment strategies derived from news headlines. Unlike traditional sentiment-based models that rely on static lexicons or

Table 8: Construction Performance of the GPT Long-Short Strategy

Metrics	GPT_LongShort	GPT_Long	GPT_Short	EqualWeighted	ETF
Cumulative Return	58.05%	8.37%	41.86%	25.62%	-6.77%
Annualized Return	18.82%	4.17%	14.06%	6.14%	0.82%
Annualized Volatility	24.70%	18.08%	21.51%	8.58%	21.52%
Sharpe Ratio	0.76	0.23	0.65	0.72	0.04
Sortino RAtio	0.85	0.33	0.46	0.90	0.05
Maximum Drawdown	-34.45%	-25.94%	-33.06%	-12.58%	-34.78%
Skewness	-2.39	2.11	-2.96	0.03	0.26
Kurtosis	65.10	20.84	76.17	9.89	4.31

narrow indicators, we apply a large language model (GPT-4.1 Nano) to interpret the tone of news articles and dynamically generate trading signals across firms.

Using over 500,000 Korean-language news headlines published between 2022 and 2025, we assign sentiment scores to individual companies based on headline analysis and construct daily rebalanced, self-financing long-short portfolios. The results show that GPT-based sentiment strategies can produce statistically and economically significant returns, especially in market-neutral settings. These findings are consistent with recent evidence from U.S. markets Lopez-Lira and Tang (2023), indicating that modern language models can uncover predictive signals from unstructured textual data—even in linguistically and structurally different environments like Korea.

We further explore the cross-sectional heterogeneity of strategy performance. The GPT model exhibited the strongest results among large-cap stocks, while SMID-cap performance was modest, potentially due to liquidity constraints and short-selling frictions. In terms of investment style, growth stocks were particularly responsive to sentiment signals, whereas value stocks exhibited limited upside but significant downside predictability, favoring short-only strategies. Sector analysis revealed highly divergent patterns, with standout performance in chemicals, insurance, and construction, driven by different combinations of long vs. short contributions.

Key Implications

- Sentiment signals derived from large language models can complement traditional financial analysis by capturing investor psychology and narrative dynamics.
- GPT models are effective even in Korean, despite not being explicitly trained on Korean financial text, suggesting applicability in non-English markets.
- Strategy customization by firm type and industry is essential, as the effectiveness of

sentiment-based trading varies substantially across market segments.

Limitations and Future Works

Several limitations warrant consideration. First, transaction costs and liquidity constraints, particularly for short-selling, are not explicitly incorporated and may reduce real-world profitability. Second, the sentiment model uses only headline-level text. Including full article content, additional sources such as social media, or multimodal inputs (e.g., earnings calls, charts) could further improve predictive power. Finally, the GPT model used here was not fine-tuned for Korean finance. Developing a domain-specific, Korean-language LLM could enhance performance even further.

Despite these limitations, this study provides one of the first comprehensive demonstrations of GPT-based sentiment trading in the Korean stock market. It lays the foundation for future work that combines AI and behavioral finance to better understand the mechanisms through which market sentiment drives asset prices. The insights gained from this research may benefit asset managers, hedge funds, and regulators interested in the growing intersection of language models and financial markets.

Appendix

Table 9: All Sector Performance of the GPT Long-Short Strategy

Rank	Sector	Cumulative Return	Annualized Return	Annualized Volatility	Sharpe Ratio	Sortino Ratio	Maximum Drawdown	Skewness	Kurtosis
1	Chemicals	106.74%	28.30%	22.47%	1.26	1.78	-26.80%	0.23	6.13
2	Textiles & Apparel	69.94%	19.86%	17.34%	1.13	0.96	-22.03%	-1.65	29.11
3	Insurance	62.34%	17.89%	17.49%	1.02	1.44	-15.58%	0.72	6.73
4	Construction	58.05%	18.82%	24.70%	0.76	0.85	-34.45%	-2.38	65.10
5	Retail	32.48%	10.42%	15.91%	0.66	1.06	-18.42%	1.36	9.80
6	Other Manufacturing	23.42%	6.92%	5.92%	1.17	2.07	-8.78%	1.74	6.96
7	Entertainment & Media	19.99%	5.47%	13.02%	0.44	0.58	-32.56%	1.07	18.49
8	Medical & Precision Instruments	16.77%	5.78%	13.02%	0.44	0.58	-17.15%	0.90	9.71
9	Real Estate	11.32%	3.28%	9.94%	0.26	0.26	-10.16%	0.58	73.17
10	Electrical & Electronics	9.74%	2.92%	8.99%	0.36	0.20	-14.97%	-0.07	12.38
11	Food, Beverage & Tobacco	6.24%	1.99%	10.36%	0.19	0.20	-44.54%	-2.57	33.10
12	Telecommunications	2.79%	1.18%	8.01%	0.15	0.18	-11.25%	0.53	9.62
13	Banks	-0.67%	1.13%	16.21%	0.07	0.07	-28.89%	-1.28	12.73
14	Pharmaceuticals	-2.96%	1.09%	7.96%	0.05	0.07	-36.42%	1.23	24.61
15	Agriculture, Forestry & Fisheries	-5.92%	-2.74%	3.25%	-0.84	-0.19	-7.74%	-1.67	3.45
16	Shipbuilding	-8.17%	-1.20%	10.63%	-0.12	-0.07	-19.77%	-0.34	17.79
17	Transportation & Logistics	-8.91%	-1.65%	12.66%	-0.15	-0.17	-27.77%	1.63	13.45
18	Auto Components & Equipment	-9.78%	-1.96%	10.63%	-0.16	-0.16	-28.61%	0.54	7.55
19	IT Services	-10.76%	-1.94%	20.40%	-0.11	-0.09	-32.38%	0.08	9.97
20	Other Financials	-15.58%	-3.43%	16.85%	-0.18	-0.25	-28.85%	0.85	14.31
21	Non-metallic Minerals	-16.52%	-2.75%	23.56%	-0.13	-0.13	-38.45%	-0.08	15.53
22	Metals & Mining	-37.07%	-5.92%	28.79%	-0.33	-0.39	-51.52%	1.66	29.45
23	Machinery & Equipment	-41.65%	-9.44%	29.74%	-0.30	-0.54	-71.93%	-4.75	32.66
24	Paper & Forest Products	-62.45%	-8.95%	25.89%	-0.24	-0.19	-69.57%	-6.91	88.38
25	Utilities (Electricity & Gas)	-67.25%	-26.80%	40.25%	-0.80	-0.82	-69.57%	1.67	89.07
26	Commercial Services	-70.16%	-26.81%	33.40%	-0.80	-0.65	-73.97%	-5.17	74.93

References

Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ding, X., Zhang, Y., Liu, T., and Duan, J. (2015). Deep learning for event-driven stock prediction. In *Proceedings of IJCAI 2015*, pages 2327–2333.

Hanley, K. W. and Hoberg, G. (2010). The information content of ipo prospectuses. *The Review of Financial Studies*, 23(7):2821–2864.

Heston, S. L. and Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73(3):67–83.

Huang, M., Liu, Y., and Chen, Y. (2016). Predicting stock market movement direction with deep learning. *Neurocomputing*, 205:227–233.

Kang, J.-W. and Choi, S.-Y. (2025). Comparative investigation of gpt and finbert's sentiment analysis performance in news across different sectors. *Electronics*, 14(6):1090.

Kelly, B., Pruitt, S., and Su, Y. (2021). Characterizing the investment opportunity set: A new dimension. *Journal of Financial Economics*, 142(2):697–718.

Kim, Y., Kim, N., and Jung, S. (2012). 뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자 의사결정모형. *지능정보연구*, 18(2):143–156.

Kirtac, K. and Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, 62:105227.

Lee, J. and Kwon, T. (2025). 한국 뉴스 금융 감성지수(knfsi) 제안 및 kospi 200과의 연관성 평가. *한국데이터분석학회 저널*, 27(2):521–534.

Li, F. (2010). The information content of forward-looking statements in corporate filings. *Journal of Accounting Research*, 48(5):1049–1102.

Lopez-Lira, A. and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *SSRN Working Paper*. Posted April 2023, Last rev. April 2024.

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

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

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