Decoding Sentiment: Predicting Stock Returns from Japanese 10-K Reports with Advanced Large Language Models

Takahiro Yamasaki[1] Moe Nakasuji[2] Katsuhiko Okada[2] Yasutomo Tsukioka[3]
[1]Osaka Sangyo University, Faculty of Engineering [2]Kwansei Gakuin University, Institute of Business and Accounting
[3]Kwansei Gakuin University, School of Business Administration
yamasaki@eic.osaka-sandai.ac.jp, Nakasuji@kwansei.ac.jp, katsuokada@kwansei.ac.jp,
tsukioka@kwansei.ac.jp

Abstract

This study leverages advanced natural language processing techniques and large language models (LLMs)—including ChatGPT, Claude, and Gemini—to extract sentiment from Japanese 10-K reports and predict stock returns. Using a dataset of 11,135 firm-years from Tokyo Stock Exchange-listed companies (2014–2023), we compare LLMs with dictionary-based methods and a DeBERTaV2 model. While traditional approaches show no significant sentiment-return relationship, LLM-derived sentiment reveals a significant negative correlation with future stock performance, challenging the efficient market hypothesis. These findings underscore the transformative potential of LLMs in financial analysis, offering predictive insights undetected by traditional methods.

1. Introduction

The rapid evolution of artificial intelligence (AI) and natural language processing (NLP) is transforming the financial industry, particularly in the analysis of textual data such as annual financial reports. Japanese 10-K filings, rich in qualitative information, hold significant potential for influencing investor behavior and stock performance. However, traditional methods like dictionary-based sentiment analysis often fail to capture the nuanced language and contextual subtleties inherent in financial texts, especially in non-English markets like Japan. This gap is addressed by the advent of large language models (LLMs) such as ChatGPT, Claude, and Gemini, which bring unprecedented capabilities for understanding and generating human-like text.

This study employs LLMs alongside traditional methods to analyze the complete set of 10-K reports from all companies listed on the Tokyo Stock Exchange between 2014 and 2023, encompassing 11,135 firm-years and over 70 million words. By comparing the performance of LLMs, dictionary-based methods, and a DeBERTaV2-based model, we assess the ability of advanced NLP techniques to extract sentiment and predict stock returns. Our findings reveal a significant negative correlation between LLM-derived sentiment and future stock returns, suggesting that high sentiment scores may indicate overvaluation, leading to subsequent market corrections. This challenges the efficient market hypothesis and underscores that qualitative information in corporate disclosures may not be fully priced into the market.

Building on prior research that highlights the limitations of traditional sentiment analysis in finance, our study demonstrates the transformative potential of LLMs in extracting predictive insights from large-scale financial texts. By focusing on official corporate disclosures, we reduce noise and improve relevance, offering new evidence that markets may not immediately incorporate nuanced qualitative information. The remainder of the paper discusses prior literature, data and methods, empirical results, and the implications of our findings for financial theory and practice.

2. Data and Methodology

2.1 Data

The 10-K reports analyzed in this study were sourced from the Electronic Disclosure for Investors' NETwork

Table 1 : Summary of Dataset Characteristics: 10-K Reports from Firms Listed on the Tokyo Stock Exchange (2014–2023)

	Number of firms	Average length of MD&A (Number of characters)	Min	25%	Median	75%	Max
2014	1,012	6,159	1,910	4,409	5,339	6,654	43,996
2015	1,052	6,136	2,230	4,384	5,365	6,764	43,943
2016	1,069	6,095	1,956	4,414	5,412	6,821	40,281
2017	1,091	5,829	2,094	4,290	5,149	6,491	32,411
2018	1,165	6,294	2,291	4,720	5,751	7,106	29,553
2019	1,190	6,510	2,233	4,896	5,960	7,289	29,542
2020	1,188	7,501	2,966	5,633	6,834	8,536	29,622
2021	1,191	7,345	2,855	5,537	6,639	8,233	29,576
2022	1,176	7,278	2,663	5,506	6,630	8,251	29,614
2023	1,001	7,210	2,797	5,497	6,607	8,117	29,558
Total	11,135	6,636	1,910	4,899	5,052	7,575	43,996

Note: Table 1 presents breakdown of the dataset used in this study, summarizing the annual number of firms listed on the Tokyo Stock Exchange, their corresponding 10-K reports, and the total stocks analyzed. A particular focus is placed on the Management Discussion and Analysis (MD&A) sections. This section, equivalent to the MD&A in U.S. filings, provides in-depth qualitative narratives on financial conditions, operational results, and cash flow status.

(EDINET), operated by Japan's Financial Services Agency. EDINET ensures access to authentic and complete financial filings for companies listed on Japanese stock exchanges. We focused on firms whose fiscal year ends in March, the standard for most TSE-listed companies, resulting in 16,363 firm-year observations and over 90 million words of textual data. This extensive dataset enables a comprehensive analysis of management sentiment across the Japanese market.

Table 1 summarizes the dataset, including the number of firms (10-K reports) per year, the total stocks analyzed, and the average length of the Management Discussion and Analysis (MD&A) sections in characters. The MD&A section, equivalent to its U.S. counterpart, offers rich qualitative insights into financial condition, operational results, and management outlook, making it a critical focus for evaluating sentiment and investor perceptions.

2.2 Sentiment Extraction Method

To extract sentiment from the Management Discussion and Analysis (MD&A) sections, we utilize five approaches: a financial polarity dictionary for Japanese financial texts, DeBERTaV2, and three large language models (LLMs)—GPT-4, Claude, and Gemini.

Dictionary-Based Methods

Using the Financial Polarity Dictionary, we compute two metrics:

Tone Ratio: Measures the balance of positive and negative words:

$$Tone\ Ratio = \frac{N^+ - N^-}{N^+ + N^-}$$

where N+ and N- denote the number of positive and negative words, respectively. Tone Score: Captures sentiment intensity by summing scores assigned to each word:

$$Tone\ Score = \sum_{i=1}^{n} s_i \qquad s_i \in \mathbb{R}$$

where s_i is the sentiment score of the i-th word, and n is the total number of words in the MD&A section.

DeBERTaV2 Model

The DeBERTaV2 model uses a combined objective for masked language modeling (MLM) and sentiment classification:

$$L = L_{MLM} + \lambda L_{SC}$$

where

$$L_{MLM} = -\frac{1}{|M|}\sum_{i \in M} logP(x_i | X_{\setminus i})$$

$$L_{SC} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{c \in C} y_{j,c} logP(c|X_j)$$

Here, M is the set of masked tokens, N is the number of samples, C represents sentiment classes, $y_{j,c}$ is the ground-truth label for class c, and λ is a hyperparameter. The final sentiment score for a document D is:

$$S(D) = \sum_{j=1}^{m} w_j \cdot \sum_{c \in C} s(c) \cdot P(c|X_j)$$

Table 2 Distribution of Sentiment Scores Extracted from MD&A Sections Using Five Different Methods

	Mean	Min	25%	Median	75%	Max
Tone Ratio	-0.144	-0.393	-0.184	-0.142	-0.102	0.083
Tone Score	-3.265	-194.445	-11.188	-1.872	6.108	92.435
DeBERTaV2	0.689	0.004	0.182	0.998	0.999	0.999
GPT-4	0.575	0.100	0.400	0.600	0.700	0.900
Claude	0.616	0.000	0.500	0.600	0.700	0.800
Gemini	0.517	0.100	0.400	0.600	0.700	0.900

Note:Sentiment scores are calculated using six measures from five methodologies. The first two, Tone Ratio and Tone Score, are based on the Financial Polarity Dictionary by the University of Tokyo. Tone Ratio measures the balance of positive (N+) and negative (N−) words while Tone Score aggregates sentiment strength of each word. The third measure uses the DeBERTaV2 model, fine-tuned on the chABSA dataset, to compute sentiment probabilities and document-level scores. The final three measures employ GPT-4-mini, Claude 3 Haiku, and Gemini 1.5 Flash, which generate sentiment scores scaled to sum to 1 based on standardized prompts in Japanese.

where w_j is the weight for sentence j, and $s(c)$ is the numerical sentiment score for class c.

Large Language Models (LLMs)

For GPT-4, Claude, and Gemini, we develop standardized Japanese prompts instructing models to calculate positive and negative sentiment scores summing to 100. The prompts ensure consistent evaluation across MD&A sections. Each LLM processes the text using its unique architecture:

- GPT-4 emphasizes adaptability through task-specific prompts.
- Claude focuses on logical consistency and contextual depth.
- Gemini integrates semantic and thematic context dynamically.

These methodologies provide a comparative framework to evaluate the efficacy of traditional dictionary-based metrics and advanced LLMs in capturing management sentiment and its implications for future stock returns. **Table 2** summarizes the key characteristics of these measures, offering insights into their distribution, central tendencies, and variability.

3. Results

3.1 Sentiment based Long-Short Portfolio Return

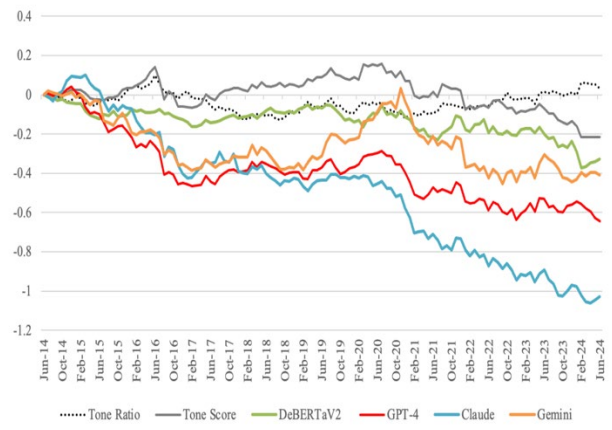
The primary aim of this study is to investigate whether large language models (LLMs) can identify information within Japanese 10-K reports that anticipates future stock returns. Under the Efficient Market Hypothesis (EMH) outlined by [1], all public information should already be embedded in asset prices, rendering attempts to achieve abnormal returns through conventional analysis futile. However, previous empirical work points to a more complex reality. Studies such as [5] and [6] reveal that textual characteristics of corporate disclosures can encode signals predictive of subsequent stock performance. This tension between theory and empirical evidence raises the question of how novel analytical tools—particularly LLMs—might bridge the gap.

To explore the predictive value of such sentiment measures, this research constructs value-weighted portfolios sorted on sentiment extracted from the Management Discussion and Analysis (MD&A) sections of 10-K reports. For each fiscal year from 2014 through 2023, all firms listed on the Tokyo Stock Exchange are ranked according to their sentiment scores as of June. These scores, derived from the previous fiscal year's filings ending in March, determine which firms enter the long and short portfolios. The top quintile of firms with the highest sentiment forms the long portfolio, and the bottom quintile with the lowest sentiment forms the short portfolio. Both portfolios are value-weighted by market capitalization at the formation date and are held for one year, spanning from July of year t to June of year t+1. This procedure repeats annually for each of the six sentiment extraction methods—Tone Ratio, Tone Score, DeBERTaV2, GPT-4, Claude, and Gemini—providing a comparative framework that evaluates conventional and advanced approaches in tandem.

Figure 1 presents the cumulative returns for the long-short portfolios derived from each method. The results reveal important distinctions in how these methods capture return-predictive sentiment. Contrary to conventional expectations, the long-short portfolios

constructed with GPT-4 and Claude sentiment measures exhibit sustained negative cumulative returns. This finding indicates that firms displaying high positive sentiment underperform those classified as having strong negative sentiment, suggesting that positive sentiment may be overvalued, negative sentiment may be overvalued, or both. The portfolios based on Tone Ratio and Tone Score, while grounded in dictionary-based approaches, show only modest or even negative results, indicating that these traditional metrics do not isolate the type of sentiment that correlates with future stock performance. DeBERTaV2 similarly struggles to produce strong predictive signals. The Gemini-based portfolios produce intermediate results, performing better than the traditional methods yet not reaching the levels of GPT-4 and Claude. Together, these findings imply that certain LLM architectures yield a more nuanced sentiment measure that aligns with future returns, challenging the assumption that all publicly available information is already reflected in share prices.

Figure 1: Cumulative Returns of Long-Short Portfolios Based on Sentiment Scores



3.2 Performance Evaluation of Sentiment-Based Portfolios

To evaluate the performance of sentiment-based portfolios, we use standard asset pricing models, including the Fama-French three-factor (FF3)[2], Carhart four-factor (FFC4)[3], and Fama-French five-factor (FF5) models[4]. These models help determine whether portfolio returns can be explained by systematic risk factors or exhibit abnormal performance (alpha). **Table 3** shows that sentiment-based portfolios using GPT-4 and Claude generate statistically significant negative alphas across all models. For GPT-4, annualized alphas are -5.95% (FF3), -6.49% (FFC4), and -6.29% (FF5), all significant at the 5% level. Claude-based portfolios exhibit even larger negative alphas of -9.15% (FF3), -9.90% (FFC4), and -9.46% (FF5), all significant at the 1% level. In contrast, other sentiment-based methods (Tone Ratio, Tone Score, DeBERTaV2, and Gemini) yield alphas that are not statistically significant (not shown in the table), indicating no abnormal returns after accounting for common risk factors.

Table 3 Asset Pricing Regression Results for Long-Short Portfolios Based on Sentiment Scores

	GPT-4			Claude			Gemini		
	FF3	FFC4	FF5	FF3	FFC4	FF5	FF3	FFC4	FF5
α	-0.005 **	-0.005 **	-0.005 **	-0.008 ***	-0.008 ***	-0.008 ***	-0.004	-0.005	-0.004
Rm-Rf	0.088	0.135 **	0.003	-0.0002	0.065	-0.077	0.167 **	0.218 ***	0.085
SMB	-0.094	-0.147	-0.080	-0.183	-0.257 *	-0.171	-0.256 *	-0.316 **	-0.219
HML	-0.523 ***	-0.461 ***	-0.185	-0.602 ***	-0.516 ***	-0.259 *	-0.821 ***	-0.753 ***	-0.559 ***
WML		0.192 *			0.267			0.212 *	
RMW			0.578 **			0.538 *			0.116
CMA			-0.186			-0.194			-0.538 **
Adj R-Squared	0.249	0.265	0.291	0.219	0.246	0.234	0.377	0.388	0.384

4. Conclusion

This study demonstrates that advanced large language models (LLMs), such as GPT-4 and Claude, can extract return-predictive information from Japanese 10-K reports, challenging the weak form of the Efficient Market Hypothesis (EMH). Portfolios based on these LLMs generated significant negative alphas, even after accounting for common risk factors, while traditional methods showed no such predictive power. These findings highlight the potential of LLMs to uncover subtle signals in corporate disclosures that traditional approaches miss, raising questions about market efficiency.

Future research should explore why certain LLMs outperform others, focusing on their architectures, training data, and ability to capture linguistic nuances. Understanding these differences is crucial for refining textual analysis in financial contexts and better integrating advanced NLP into investment strategies.

Acknowledgement

This research was supported by the Grant-in-Aid for Scientific Research (B), Project Number 24K00298.

Reference

1. Efficient Capital Markets: A Review of Theory and Empirical Work. Fama Eugene F. *Journal of Finance,* 1970, Vol. 25, 383-417.
2. Common Risk Factors in the Returns on Stocks and Bonds. Fama Eugene F., French Kenneth R. *Journal of Financial Economics*, 1993, Vol. 33, 3-56.
3. On Persistence in Mutual Fund Performance. Carhart Mark M. *Journal of Finance*, 1997, Vol. 52, 57-82.
4. A Five-Factor Asset Pricing Model. Fama Eugene F., French Kenneth R. *Journal of Financial Economics*, 2015, Vol. 116, 1-22.
5. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Tetlock Paul C. J*ournal of Finance,* 2007, Vol. 62, 1139-1168.
6. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10‐Ks. Loughran Tim, McDonald Bill. *Journal of Finance*, 2011, Vol. 66, 35-65.