

Japanese Term Selection for Stock Price Fluctuation by Large Language Models

Takehito Utsuro and Shunsuke Nishida

Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba
{utsuro.takehito.ge, s2320778}@u.tsukuba.ac.jp

Abstract

In Japanese articles on stock price fluctuations, technical terms in the stock domain are frequently used to precisely describe stock price fluctuations. We proposed the methods for the selection of such terms that appropriately represent the characteristics of stock price fluctuations and conducted evaluation by feeding closing prices to large language models and a chart of stock price fluctuations over several days to large multimodal models. The results showed that, with high accuracy, all the models were able to select terms that are manually assigned by human writers in stock price fluctuation articles or those with similar meanings to them. It suggests the potential to generate stock price fluctuation articles containing appropriate terms from time series stock price data and text of articles in which stock price fluctuations are not directly mentioned but are related to them. The results also showed that the method of conducting few-shot learning with GPT-4o exhibited the highest accuracy in term selection among other approaches.

1 Introduction

News articles reporting stock price fluctuations are useful in providing information not only about how much stock prices have risen or declined but also in understanding the factors influencing price fluctuations, such as announcements of new products and social conditions. Although such articles are usually written manually, it is desirable that they are generated automatically in large quantities. This can be realized when information regarding any cause closely related stock price fluctuations as well as events on relevant companies are automatically collected through the Internet and aggregated, from which stock price fluctuation articles texts are automatically generated. Once those technologies are broadly available, it is then ensured that we can avoid spending any manual effort writing those stock price fluctuation articles, allowing us

to redirect the effort spent on article generation towards investment decisions, fluctuation predictions, and actions that maximize economic profits. With such an environment, economic activities can be significantly accelerated.

In Japanese articles on stock price fluctuations, technical terms related to stocks are often used. In particular, stock terms describing stock price fluctuations (henceforth, stock price fluctuation terms) are frequently used, and they are used differently depending on the magnitude and continuity of stock price fluctuations. For example, in the case of a rise in stock price, there is a distinction between “急伸 (sharp rise)” when the stock price rises sharply and suddenly and “続伸 (continuous rise)” when the stock price rises continuously. In the process of automatically generating stock price fluctuation articles, it is crucial to analyze time series data of stock prices automatically and to use the terms correctly based on subtle nuances of their meanings.

In this paper, we addressed this issue by using large language models (LLMs) (GPT-4o (OpenAI, 2024), Claude 3.5 Sonnet and Gemini 1.0 Pro) and a large multimodal model (LMM) (GPT-4V (OpenAI, 2023; Yang et al., 2023)). We proposed how to design the procedures of selecting technical terms that appropriately represent the characteristics of stock price fluctuations and conducted evaluation. The results showed that, with high accuracy, all the models were able to select terms that were manually assigned by human writers in stock price fluctuation articles or terms with similar meanings. Furthermore, the method of conducting fine-tuning with GPT-4o exhibited the highest accuracy in term selection among other approaches.

The following briefly summarizes the contributions of this paper.

1. It was revealed that the accuracies of converting stock price fluctuation data into corresponding stock terms using LLMs and an

LMM were relatively high around 90%.

- It was demonstrated that the method of conducting fine-tuning with GPT-4o exhibited the highest accuracy in term selection among other approaches.

2 Related Work

Various studies were conducted on “data to text” tasks that interpret data and generate text describing the contents. Among them, several approaches were made for the task of generating text from time series stock price data, as in this study.

Murakami et al. (2017) proposed an encoder-decoder model as a method for automatically generating market comments from short-term and long-term *Nikkei Stock Average* data. They compare the performance when CNN, MLP and RNN are used as the encoder. Aoki et al. (2021) addressed the issue of controlling text generation by inputting topic labels that represent the content of the generated sentences in addition to stock price data. While these studies aim to generate sentences, this study focuses specifically on the generation (selection) of stock price fluctuation terms.

Zhang et al. (2018) proposed methods that utilize probability models to select verbs representing stock price fluctuations from the volatility. Sekino and Sasaki (2022) also proposed to use an MLP encoder model to choose words describing stock price movements and volatility based on the closing price trends of the *Nikkei Stock Average* and *Dow Jones Industrial Average*.

Unlike the aforementioned related studies, all of which take numerical data as input, this paper differs in that we further study incorporating a multimodal model. In this approach, the model is designed to generate (select) stock price fluctuation terms based on stock price chart images.

In the context of studies on news article headlines and stock prices, Nishida et al. (2023) studied the task of headlines generation of stock price fluctuation articles, derived from the articles’ content, where they solve three distinct tasks of generating article headlines, extracting the stock names, and ascertaining the trajectory of stock prices, whether they are rising or declining. Tsutsumi and Utsuro (2022) studied the issue of detecting causes of stock price rise and decline from the stock price fluctuation articles by machine reading comprehension models. In the context of stock price prediction using news headlines, Kalshani et al. (2020) studied

sharp rise	continuous rise	rebound	continuous sharp rise	sharp rebound
43	82	99	25	31
sharp decline	continuous decline	pullback	continuous sharp decline	sharp pullback
93	55	53	28	59

Table 1: Number of articles for each stock price fluctuation term (568 articles in total)

to combine news headlines with technical indicators to predict stock prices. Chen (2021) studied to predict the short-term movement of stock prices after financial news events using only the headlines of the news. Kalyani et al. (2016) proposed a method for stock trend prediction using news. Two other approaches evaluate different machine learning and deep learning methods, such as Support Vector Machines (SVM) and Long Short-term Memory (LSTM), to predict stock price movement using financial news (Liu et al., 2018; Gong et al., 2021).

3 Stock Price Fluctuation Terms

Stock price fluctuation terms in this paper are intended to be regarded as the terminology in the stock domain that are used to describe stock price fluctuations. “急伸 (sharp rise)” and “続伸 (continuous rise)” given as examples in section 1 are also included in the stock price fluctuation terms.

It is expected to maximize the advantages of using LLMs / an LMM by freely generating stock price fluctuation terms from stock price fluctuations over several days. However, to facilitate the evaluation of performance, it is necessary to select candidate terms and let models select terms among them. Based on this discussion, we made a list of 28 phrases that are commonly used in stock price fluctuation articles as candidates of stock price fluctuation terms. Out of those 28 phrases, based on the criteria we introduce below, we adopted the following 10 terms for the study in this paper, which can be determined from short-term stock price fluctuations and have a high frequency of occurrences in stock price fluctuation articles.

“急伸 (sharp rise)”, “続伸 (continuous rise)”, “反発 (rebound)”, “急落 (sharp decline)”, “続落 (continuous decline)”, “反落 (pullback)”, “続急伸 (continuous sharp rise)”, “急反発 (sharp rebound)”, “続急落 (continuous sharp decline)”, and “急反落 (continuous sharp decline)”

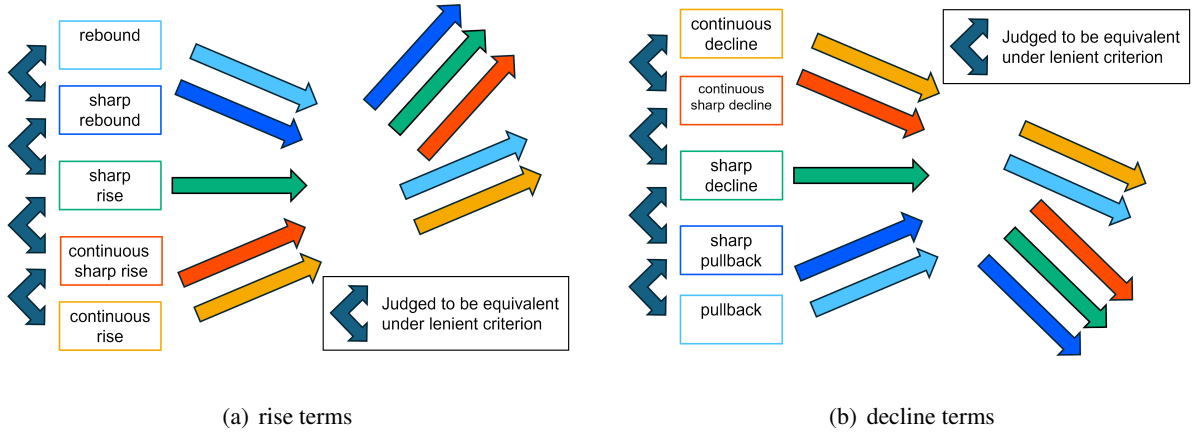


Figure 1: Illustrating the differences of 10 stock price fluctuation terms

Definitions of stock price fluctuations corresponding to those 10 terms are illustrated in Figure 1, where Figure 1(a) shows 5 terms representing stock price rise, while Figure 1(b) shows the other 5 terms representing stock price decline. The illustration of each term consists of left hand side and right hand side fluctuations, where some of them share half of those fluctuations with other term(s)¹. In Figure 1, each of the eight pairs connected with the “Judged to be equivalent under lenient criterion” arrows share one of their left or right fluctuation, which corresponds to satisfying the *lenient criterion* defined in section 5.5.

The followings give the detailed discussion on the criteria on selecting those 10 terms out of the overall 28 phrases. We first investigate 8,024 articles collected from Web media that deliver news about stock price fluctuations². Out of the overall 8,024 articles, 980 articles, which accounts for 12.2%, contained at least one of those selected 10 terms, while 776 articles (9.67%) contained at least one of the remaining 18 terms that were not adopted due to several reasons³. We also analyzed 100 articles randomly sampled from 6,268

(= 8,024 – 980 – 776) articles that include none of the overall 28 phrases representing stock price fluctuations. The majority of those remaining articles correspond to articles on whole market trends and promotional articles for companies.

4 Dataset

“Yahoo! Finance”⁴ and “MINKABU”⁵, two Web media that distribute news articles on finance, were used for collecting news articles on stock price fluctuations. We focus on the headlines of news articles taken from the “Japanese stocks” tab of “Yahoo! Finance” and the “individual stocks” tab of “MINKABU”⁶, from which we collected 568 articles. Each of those collected 568 articles satisfies the requirement that its headline contains only one of the 10 stock price fluctuation terms selected in the previous section.

From the collected articles, the article headlines and distribution dates were extracted. In addition, the stock price time series data for the relevant stocks linked from the article pages were referred to, where the closing prices were obtained from “Yahoo! Finance” and “MINKABU”, for up to one week prior to the distribution dates of the articles. Based on the information obtained, a dataset was created consisting of pairs of stock price fluctuation terms appearing in the articles (e.g. “sharp rise”) and closing prices up to one week backward from the distribution date. Table 1 summarizes the number of articles for each stock price fluctuation

¹For example, in Figure 1(a), “rebound” (light blue) and “sharp rebound” (dark blue) share left hand side fluctuation, while their right hand side fluctuations differ. .

²Articles distributed from “MINKABU” (<https://minkabu.jp/>) between February 26, 2024 and March 26, 2024.

³For example, requiring long-term stock price fluctuation data (e.g., “堅調 (rising in the long term)” and “軟調 (declining in the long term)”), having difficulty in differentiation from other terms due to representing rather general features such as generally rising and declining (e.g., “上昇 (rise)” and “下落 (decline)”) and representing accidental fluctuations within a day (e.g., “ストップ高 (hit limit-up, i.e., stop further selling/buying in the market due to relatively large rise)” and e.g., “ストップ安 (hit limit-down, i.e., stop further selling/buying in the market due to relatively large decline)”).

⁴<https://finance.yahoo.co.jp/>

⁵<https://minkabu.jp/>

⁶Articles distributed between November 8, 2023 and January 5, 2024, where those 568 articles are collected independently of the 8,024 articles collected in the previous section, but only for the purpose of evaluation.

y [%] \ x [%]	$x > 5$	$0 \leq x \leq 5$	$-5 \leq x < 0$	$x < -5$
$y > 1$	continuous sharp rise	continuous rise	pullback	sharp pullback
$0 < y \leq 1$	sharp rise			rebound
$-1 \leq y < 0$		sharp rebound	continuous sharp decline	
$y < -1$				

Table 2: Rule-based term selection (x stands for percentage change in stock price from 1 day ago to article distribution date and y stands for percentage change in stock price from 2 days ago to 1 day ago.)

term.

5 Experiment

5.1 Rule-based Term Selection

Stock price fluctuation terms are selected based on a simple rule derived from the rate of change in the stock’s closing price. Specifically, the rate of change is calculated from the closing price of one day prior to the article publication date to the closing price on the publication date itself (x in Table 2), as well as from the closing price two days prior to one day prior to the publication date (y in Table 2). As shown in Table 2, for both x and y , these rates of change are divided into four ranges using three thresholds. The combinations of these rate of change ranges are then mapped to one of those 10 stock price fluctuation terms as shown in Table 2. The rule was created by the second author, referencing the rate of change in closing prices within the training data used in the experiment.

5.2 GPT-4o (Large Language Models)

The task involves providing closing prices for several consecutive days to LLMs and prompting it to select, from the 10 stock price fluctuation terms defined in section 3, the term that best describes fluctuation of stock terms. Based on the results of the preliminary experiment to be conducted in section 5.4, we decide to reference closing prices over three days. We employed GPT-4o (*gpt-4o-2024-05-13*) as the LLM and conducted zero-shot learning, few-shot learning, and fine-tuning to examine the most appropriate method.

5.2.1 Zero-shot Learning without Giving Definitions of Terms

Only the following information is given to the prompt and GPT-4o is asked to select a stock price fluctuation term based on zero-shot learning.

- 10 candidate stock price fluctuation terms
- closing stock prices over three days

This allows us to investigate the extent to which GPT-4o can discriminate terms using only the generic linguistic knowledge it has acquired during pre-training. The actual prompt is shown below. The actual prompt is written in Japanese, and the following is its translation into English.

```
messages=[
  {"role": "system", "content":
    "You are an AI who looks at closing
    stock prices of the day before yester-
    day, yesterday and today and selects
    the term that best fits the characteris-
    tics of the price fluctuation from the fol-
    lowing terms: "sharp rise", "continuous
    rise", "rebound", "sharp decline", "con-
    tinuous decline", "pullback", "continu-
    ous sharp rise", "sharp rebound", "con-
    tinuous sharp decline" and "sharp pull-
    back"."},
  {"role": "user", "content": "(928.0,
    926.0, 1030.0). . ."}]
```

5.2.2 Zero-shot Learning with Giving Definitions of Terms

The following information is given to the prompt:

- 10 candidate stock price fluctuation terms
- definition of each term
- closing stock prices over three days

and GPT-4o is asked to select a stock price fluctuation term based on zero-shot learning. The actual prompt is shown below.

```
messages=[
  {"role": "system", "content":
    "You are an AI who looks at closing
    stock prices of the day before yester-
    day, yesterday and today and selects
```

the term that best fits the characteristics of the price fluctuation from the following terms: “sharp rise”, “continuous rise”, “rebound”, “sharp decline”, “continuous decline”, “pullback”, “continuous sharp rise”, “sharp rebound”, “continuous sharp decline” and “sharp pullback”.”

“Sharp rise: a significant rise in the stock price from yesterday to today.”

...

“Sharp pullback”: the transition of the stock price from a rise to a significant decline.”},

{“role”: “user”, “content”: “(928.0, 926.0, 1030.0). . . ”}]

5.2.3 Few-shot Learning

As a few-shot, a total of 10 examples, one for each term, are collected from the candidate set of training examples in the dataset prepared in section 4. The prompt therefore contains the following information.

- 10 candidate stock price fluctuation terms
- as a few-shot, each stock price fluctuation term and the corresponding closing prices over three days
- closing stock prices over three days

GPT-4o is used as the model. The actual prompt is shown below.

messages=[

{“role”: “system”, “content”:

“You are an AI who looks at closing stock prices of the day before yesterday, yesterday and today and selects the term that best fits the characteristics of the price fluctuation from the following terms: “sharp rise”, “continuous rise”, “rebound”, “sharp decline”, “continuous decline”, “pullback”, “continuous sharp rise”, “sharp rebound”, “continuous sharp decline” and “sharp pullback”.”}

{“role”: “user”, “content”: “(102.0, 100.0, 118.0)”},

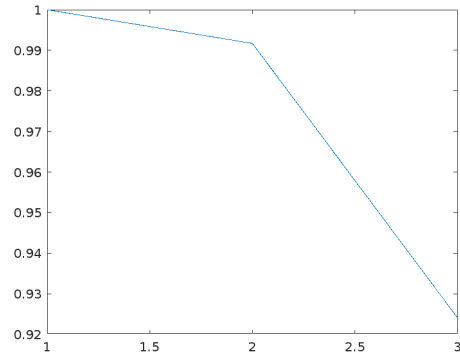


Figure 2: An example of graph images input to GPT-4V

{“role”: “assistant”, “content”: “sharp rise”},

...

{“role”: “user”, “content”: “(1808.0, 2087.0, 1818.0)”},

{“role”: “assistant”, “content”: “sharp pullback”},

{“role”: “user”, “content”:“(928.0, 926.0, 1030.0). . . ”}]

5.2.4 Fine-tuning

Using the OpenAI API, we fine-tuned *gpt-4o-2024-08-06*⁷. As training examples, a total of 100 examples are collected, 10 for each term, from the candidate set of training examples in the dataset prepared in section 4⁸⁹.

The fine-tuned *gpt-4o-2024-08-06* is used to select stock price fluctuation terms. The prompts during evaluation are the same as “zero-shot learning without giving definitions of terms”.

5.3 GPT-4V (Large Multimodal Models)

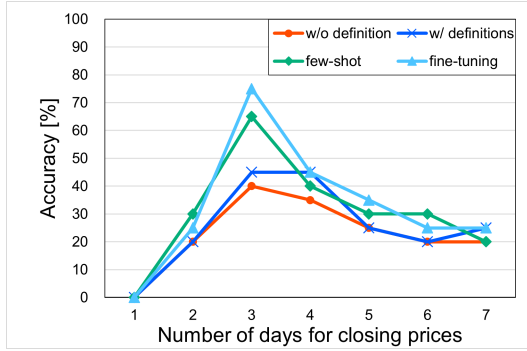
The task was to provide GPT-4V¹⁰ with an image of a stock chart represented by a line graph and have it

⁷At the time of writing this paper, GPT-4o points to *gpt-4o-2024-05-13* at the OpenAI API site, while *gpt-4o-2024-08-06* is the first version of GPT-4o that supports fine-tuning.

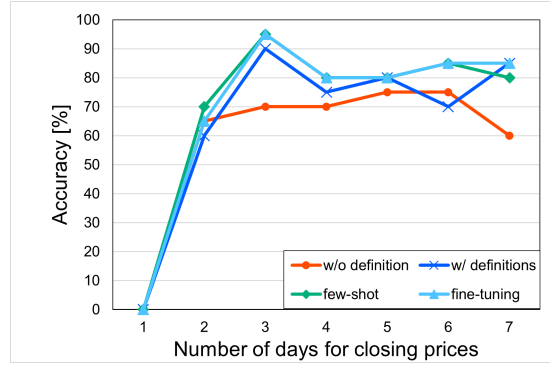
⁸As to be presented in section 5.5, in this paper, we introduce two types of evaluation criteria, namely, *strict* and *lenient*, where for the *lenient* evaluation criteria, we augment examples according to the “ \Leftrightarrow ” relation. For example, examples of “続急伸 (continuous sharp rise)” are augmented as those of “続伸 (continuous rise)” and “急伸 (sharp rise)”. After this augmentation, 10 training examples each for a stock price fluctuation term are randomly selected.

⁹The number of training examples is optimized through evaluation against a held-out 100 development examples, where the optimal number of examples was 100 when examining with the number of candidate training examples as 150 or less.

¹⁰GPT-4V of <https://chat.openai.com/>.



(a) under strict criterion



(b) under lenient criterion

Figure 3: Results of optimizing the number of days of closing stock prices by GPT-4o

automatically select the best term to describe how the stock price fluctuates from among the 10 stock price fluctuation terms defined in section 3.

The stock price charts input to GPT-4V were created from the stock price fluctuation information in the dataset described in section 4. Based on the results of the preliminary experiment conducted in section 5.4, we decided to reference closing prices over three days. The closing stock price two days ago was set as 1.0, and the closing prices of the previous day and the current day were expressed as ratios to it. The vertical axis of the chart was taken to be the minimum to maximum values of the closing prices over the three days. An example of a stock chart is shown in Figure 2.

The GPT-4V was given the following prompt.

“You are an AI who looks at closing stock prices of the day before yesterday, yesterday and today and selects the term that best fits the characteristics of the price fluctuation from the following terms: “sharp rise”, “continuous rise”, “rebound”, “sharp decline”, “continuous decline”, “pullback”, “continuous sharp rise”, “sharp rebound”, “continuous sharp decline” and “sharp pullback”.”.

An image of stock price charts was then input. The following is an example of the GPT-4V output.

The graph shows that prices have consistently fallen. In particular, it is noticeable that after a slight decline at the beginning, there is a transition to a more rapid decline. Based on this, the most appropriate term to choose would be “continuous sharp decline”. This refers to a series of

Model	Criteria	
	strict	lenient
GPT-4o	63.0	97.0
Gemini 1.0 Pro	42.0	82.0
Claude 3.5 sonnet	45.0	80.5
rule-based	58.5	95.0
xlm_roberta_base	21.5	44.0
GPT-4V	33.0	87.0
CNN	53.5	78.0

Table 3: Overall evaluation results (%) (accuracies in 200 evaluation examples)

Model	Criteria	
	strict	lenient
w/o definition	41.5	81.5
w/ definitions	46.5	89.0
few-shot	58.5	91.0
fine-tuning	63.0	97.0

Table 4: Accuracies in 200 evaluation examples by GPT-4o (%) (w/o definition: zero-shot learning without giving definitions of terms, w/ definitions: zero-shot learning with giving definitions of terms)

significant decline over a short period of time and aptly describes the price movements shown in this graph.

5.4 Optimizing the Number of Days of Closing Stock Prices

Before the evaluation of selecting stock price fluctuation terms, we conducted a preliminary experiment to determine the optimal number of days to be referenced out of the seven days of closing stock prices when the model selects stock price fluctuation terms. For a total of 20 examples where

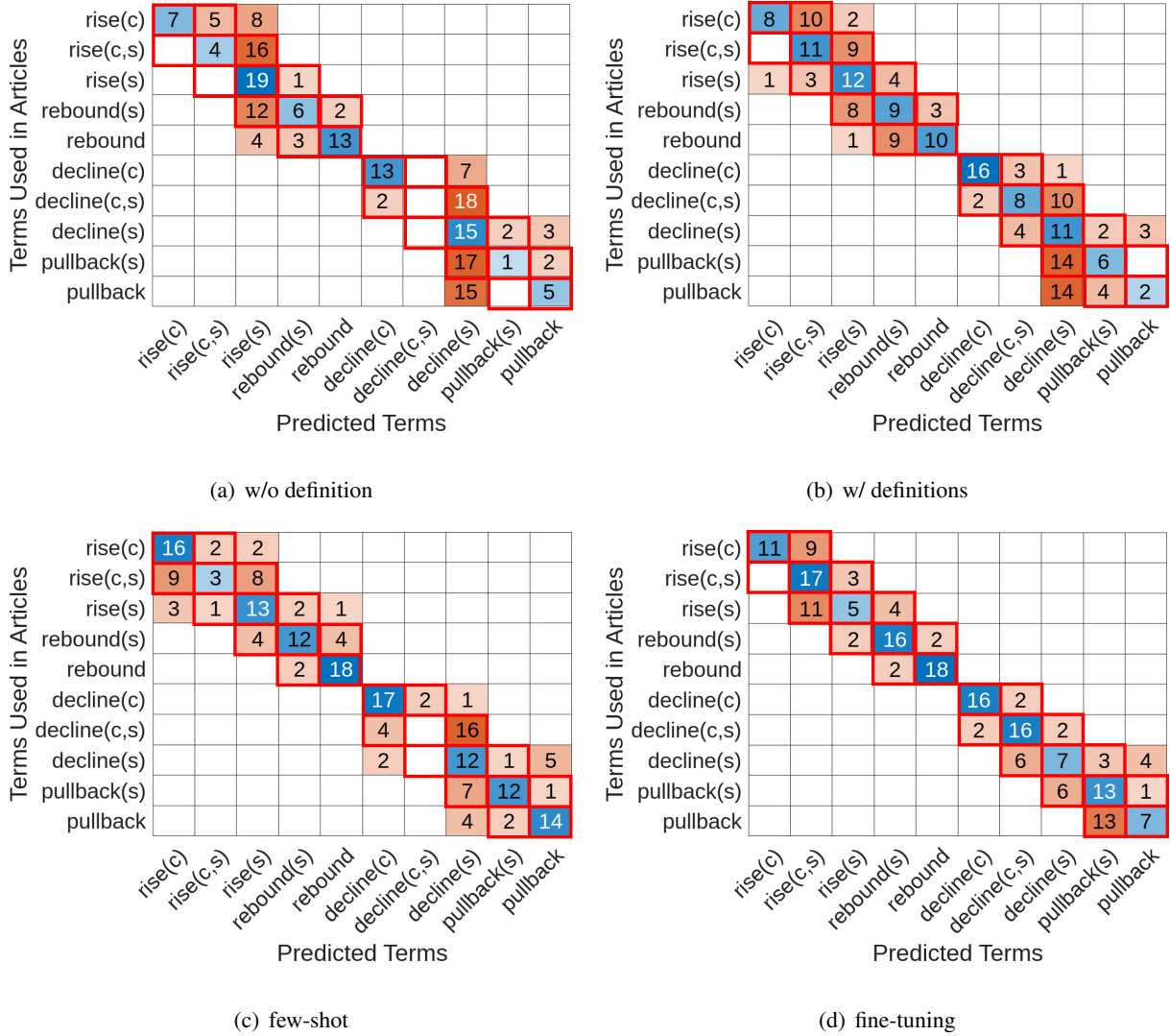


Figure 4: Confusion matrices in the evaluation results by GPT-4 (The letter “c” in brackets stands for “continuous rise / decline”, the letter “s” for “sharp rise / decline / rebound / pullback”, and the letters “c, s” for “continuous sharp rise / decline”).

two examples for each term as the held-out development dataset, we varied the number of days of closing price data given to the models from one to seven in one-day increments, and let the models select a stock price fluctuation term. Here, as shown in Figure 3, the optimal number of days of closing stock prices is three, where it is used throughout the evaluation in this paper.

5.5 Evaluation Procedure

As the evaluation experiment, the models selected stock price fluctuation terms for a total of 200 examples (i.e., 200 evaluation data), 20 for each term, which do not have any overlap with examples used for few-shot learning, the training data of fine-tuning, the development data for optimizing the number of the training data for fine-tuning, and the dataset used for optimizing the number of days of

closing stock prices in the previous section. The LLMs we used are GPT-4o, Claude 3.5 Sonnet and Gemini 1.0 Pro. For GPT-4o, we conducted two types of zero-shot learning (with / without giving definitions of terms), few-shot learning, and fine-tuning. For Claude 3.5 Sonnet and Gemini 1.0 Pro, we only performed few-shot learning following the procedure of section 5.2.3 as GPT-4o. For comparison with the LLMs, we evaluated XLM-RoBERTa (xlm-roberta-base), another language model with the same text input format, as well as a rule-based approach. XLM-RoBERTa was fine-tuned using the same data as when GPT-4o was fine-tuned. The LMM we used is GPT-4V. We also conducted a comparison with CNN, which are similar to LMMs in that they take images as input.

Two types of criteria are examined in the evaluation, i.e., *strict*, where errors between these 10

terms are not tolerated, and *lenient*, where errors between terms that are difficult to distinguish even manually are tolerated¹¹. Inter-annotator agreement rate is also measured between the terms found in the headlines of the articles and those annotated by the second author of this paper. For 100 articles that are randomly selected from the overall 568 articles, the second author selects one of the 10 candidate terms by referring to closing stock prices over three days for the stock that is relevant to each article. With the *strict* criterion, inter-annotator agreement rate between the writer of each article and the second author of this paper is 57% and Cohen’s kappa coefficient is 0.5222, while with *lenient* criterion, inter-annotator agreement rate is 93% and Cohen’s kappa coefficient is 0.9033, thus indicating sufficiently high degree of agreement.

6 Results and Discussion

6.1 Evaluation Results

Table 3 shows the overall evaluation results, while Table 4 shows those when applying GPT-4o as the model. Figure 4 also shows the confusion matrices in the evaluation results by GPT-4. These results indicate that the best performance is achieved when fine-tuning is conducted with GPT-4o.

In the strict evaluation criterion, the accuracy would be expected to be around 10% if all the terms were selected at random. For all the models evaluated in this paper, the accuracy was above 10%. In the lenient evaluation criterion, if all the terms were selected at random, the accuracy would be around 26%¹². For all the models evaluated in this paper, the accuracy was significantly higher than 26%.

The model based on stock chart images underperformed models based on numerical stock price information in terms of the strict evaluation criteria. On the other hand, for the lenient evaluation criteria, the accuracy was comparable to that of each model based on numerical stock price information.

¹¹Errors between the eight pairs directly connected with “ \Leftrightarrow ” below are allowed in the *lenient* criterion:

- continuous rise \Leftrightarrow continuous sharp rise \Leftrightarrow sharp rise \Leftrightarrow sharp rebound \Leftrightarrow rebound,
- and continuous decline \Leftrightarrow continuous sharp decline \Leftrightarrow sharp decline \Leftrightarrow sharp pullback \Leftrightarrow pullback.

¹²Out of the total 200 evaluation examples, the expected numbers of correct terms are $20 \times 3 = 60$ for 6 out of the 10 terms, while they are $20 \times 2 = 40$ for the remaining 4 terms, where their average is $((6/10) \times 60 + (4/10) \times 40) / 200 = 26\%$.

6.2 Analysis on Rule-based Term Selection

The strict accuracy of rule-based term selection is 58.5%, where we revealed that, for about half of those incorrect term selection cases, the reason can be explained by referring to stock price fluctuation for periods around one week or much longer as 25 days. The details of the analysis are described in section A of Appendix.

6.3 Analysis on Term Selection based on Stock Price Fluctuation for Periods Longer than Three Days

As a further analysis, out of the overall 200 articles of the evaluation data, we examined the 105 examples where the selected terms differed between “terms by the article writers” and “terms predicted by GPT-4o (few-shot)” in the strict criterion. For those 105 examples, we provided GPT-4o with the closing stock prices for a period longer than three days and made GPT-4o to select the terms by few-shot. The details of the analysis are described in section B of Appendix.

7 Conclusion

This paper proposed models for automatically generating stock price fluctuation terms used in stock price fluctuation articles from time series data of stock prices by LLMs. Experimental evaluation results indicated that the best performance is achieved when fine-tuning is conducted with GPT-4. It was also revealed that, under the lenient criterion, the accuracies of converting stock price fluctuation data into corresponding stock terms using LLMs were relatively high about 80% ~ 90%.

Among the future work of this paper, regarding the analyses in section 6.2 and in section 6.3, it is definitely necessary to incorporate stock price fluctuation for periods around one week or much longer as 25 days. However, overall, optimal number of days for stock price fluctuation data is three days. This indicates that whether stock price fluctuation for longer periods such as 25 days is required or not totally depends on each example. Thus, it is required to devise a framework of selecting the optimal number of days of stock price fluctuation depending on each test example. Another future work includes studying the relationship between the task of selecting stock price fluctuation terms and that of predicting future stock prices, and then integrating those two related tasks into the framework of multitask learning.

References

- K. Aoki, A. Miyazawa, et al. 2021. Controlling contents in data-to-document generation with human-designed topic labels. *Computer Speech & Language*, 66, Article 101154.
- Q. Chen. 2021. [Stock movement prediction with financial news using contextualized embedding from BERT](http://arxiv.org/abs/2107.08721). <http://arxiv.org/abs/2107.08721>. *Preprint*, arXiv:2107.08721.
- J. Gong, B. Paye, G. Kadlec, and H. Eldardiry. 2021. Predicting stock price movement using financial news sentiment. In *Proc. 22nd EANN*, pages 503–517.
- A. H. Kalshani, A. Razavi, and R. Asadi. 2020. [Stock market prediction using daily news headlines](https://ssrn.com/abstract=3685530). <https://ssrn.com/abstract=3685530>.
- J. Kalyani, H. N. Bharathi, and R. Jyothi. 2016. [Stock trend prediction using news sentiment analysis](http://arxiv.org/abs/1607.01958). <http://arxiv.org/abs/1607.01958>. *Preprint*, arXiv:1607.01958.
- Y. Liu, Q. Zeng, H. Yang, and A. Carrio. 2018. Stock price movement prediction from financial news with deep learning and knowledge graph embedding. In *Proc. 15th PKAW*, pages 102–113.
- S. Murakami, A. Watanabe, et al. 2017. Learning to generate market comments from stock prices. In *Proc. 55th ACL*, page 1374–1384.
- S. Nishida, Y. Zenimoto, X. Wang, T. Tamura, and T. Utsuro. 2023. [Headline generation for stock price fluctuation articles](https://arxiv.org/abs/2309.17421). In *Proc. 6th FinNLP*, pages 22–30.
- OpenAI. 2023. [GPT-4V\(ision\) system card](https://cdn.openai.com/papers/GPTV_System_Card.pdf). https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- OpenAI. 2024. [GPT-4o system card](https://openai.com/index/gpt-4o-system-card/). <https://openai.com/index/gpt-4o-system-card/>.
- I. Sekino and M. Sasaki. 2022. Generating market comments on stock price fluctuations using technical analysis features. *International Journal on Advances in Intelligent Systems*, 15(3,4):83–92.
- G. Tsutsumi and T. Utsuro. 2022. Detecting causes of stock price rise and decline by machine reading comprehension with BERT. In *Proc. 4th FNP*, pages 22–35.
- Z. Yang, L. Li, et al. 2023. [The dawn of LMMs: Preliminary explorations with GPT-4V \(ision\)](https://arxiv.org/abs/2309.17421). ArXiv.org, arXiv:2309.17421v2.
- D. Zhang, J. Yuan, X. Wang, and A. Foster. 2018. Probabilistic verb selection for data-to-text generation. *Transactions of the Association for Computational Linguistics*, 6:511–527.

A Analysis on Rule-based Term Selection

The strict accuracy of rule-based term selection is 58.5%, where Figure 5 shows the confusion matrix in the evaluation result by rule-based term selection. When we focus on the incorrect cases under the *strict* criterion, 58.5% accuracy corresponds to 41.5% difference, which is quite large. In order to identify the major causes of this large difference, we examine those four non-diagonal cells in the confusion matrix of Figure 5 each of which has the number of counts above or equal to four. The consequence of our analysis can be summarized as below. First, one of the most important facts here is that the rule selects terms based on the three days stock price fluctuations but not referring to stock price fluctuation for 25 days. Second, on the contrary, we found that, for about half of the articles where “the term selected by the article writers” and “rule-based term selection” differ, the reason why the article writers selected different terms can be explained by referring to stock price fluctuation for 25 days. Figure 6 ~ Figure 8 present examples of those differences between “the term selected by the article writers” and “rule-based term selection”.

Figure 6 represents fluctuation for 25 days for the case of difference between “pull back” as “the term selected by the article writers” as opposed to “sharp pullback” as “rule-based term selection” (corresponding to the cell with the count as 8 in the confusion matrix). In this figure, for the black thick plot, both “the term selected by the article writers” and “rule-based term selection” are “sharp pullback”. Here, stock price fluctuation for 25 days is without very sharp change, which makes the article writer judge its “pullback” at the end of the period as “sharp”. For the red dashed line, on the other hand, “the term selected by the article writers” is “pull back”, while “rule-based term selection” is “sharp pullback”. Stock price fluctuation for 25 days is with relatively sharper change, which makes the article writer judge its “pullback” at the end of the period as relatively “not sharp” compared with the relatively sharper fluctuation for 25 days.

Figure 7 represents fluctuation for 25 days for the case of difference between “sharp rise” as “the term by article writers” as opposed to “continuous sharp rise” as “rule-based term selection” (corresponding to the cell with the count as 4 in the confusion matrix). In this figure, for the black thick plot, both “the term selected by the article writers” and “rule-based term selection” are “continuous sharp rise”.

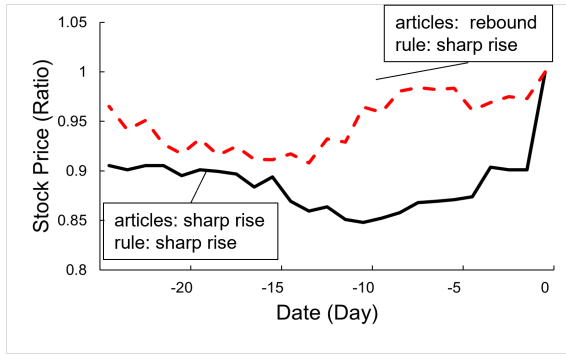


Figure 8: Analyzing the differences of “terms by the article writers” and “rule-based term selection” based on stock price fluctuation for 25 days (3) (“rebound” (by the article writer) v.s. “sharp rise” (by the rule))

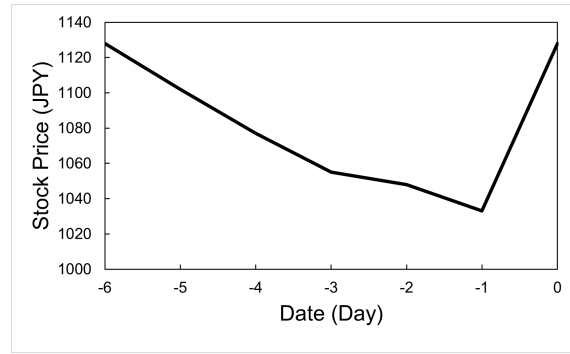
Number of Days of closing prices given to the model	4	5	6	7
Rates [%]	18.1	18.1	24.8	21.0

Table 5: Rates of examples for evaluation where “terms by the article writers” and “terms predicted by GPT-4o (few-shot)” are identical when 4~7 days of closing prices are given to the GPT-4 (out of the 105 examples where “terms by the article writers” and “terms predicted by GPT-4o (few-shot)” differ when 3 days of closing prices are given)

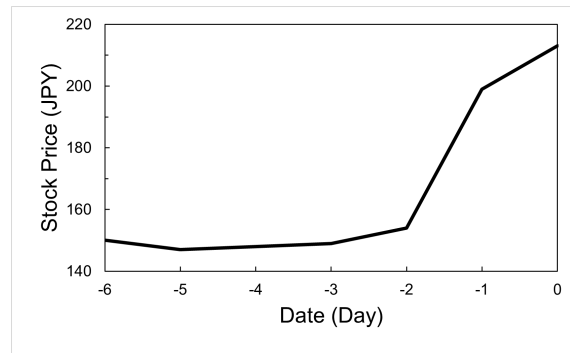
by GPT-4o (few-shot)” in the strict criterion¹⁴. For those 105 examples, we provided GPT-4o with the closing stock prices for a period longer than three days and made GPT-4o to select the terms by few-shot. Table 5 shows the rates of examples for evaluation where “terms by the article writers” and “terms predicted by GPT-4o (few-shot)” are identical when 4~7 days of closing prices are given to the GPT-4 out of the 105 examples where “terms by the article writers” and “terms predicted by GPT-4o (few-shot)” differ when 3 days of closing prices are given. As a result, for 41 out of the 105 examples, GPT-4o selected the same term as selected by the article writers for at least one of 4~7 days of closing prices.

Figure 9 presents charts of 7-day stock prices for examples in which the article writer and GPT-4o (few-shot) did not select the same term based on the closing prices for three days, while GPT-4o (few-shot) selected the term same as the article writer when based on at least one of 4~7 days of closing prices.

¹⁴The results of analysis when providing GPT-4o (fine-tuning) with the closing stock prices for a period longer than three days will be included in the camera-ready version of this paper.



(a) “sharp rebound” by the article writers v.s. “sharp rise” by GPT-4o (based on closing prices for 3 days) and “sharp rebound” by GPT-4o (based on closing prices for 4~7 days)



(b) “sharp rise” by the article writers v.s. “continuous sharp rise” by GPT-4o (based on closing prices for 3~5 days) and “sharp rise” by GPT-4o (based on closing prices for 6~7 days)

Figure 9: Charts of 7-day stock prices for examples in which the article writer and GPT-4o (few-shot) did not select the same term based on the closing prices for three days, while GPT-4o (few-shot) selected the term same as the article writer when based on at least one of 4~7 days of closing prices.

For Figure 9(a), when referring to the closing stock prices for three days, GPT-4o selected the term “continuous sharp rise” because of the small drop in the closing price of the stock from two days to one day before. On the other hand, when referring to closing prices for 4~7 days, GPT-4o selected “sharp rise”, the same term selected by the article writer, because of the continuous drop in stock prices up to 1 day before.

For Figure 9(b), when referring to the closing stock prices for 3~5 days, GPT-4o selected the term “continuous sharp rise” because of the continuous sharp rise in the closing price from 2 days before to the current day. On the other hand, when looking at longer-term fluctuations, GPT-4o selected “sharp rise”, the same term as selected by the article writer, when referring to closing prices

for 6~7 days, as there were no significant price fluctuations between 6 and 2 days prior.