Bank of Japan Review 2025-E-5

A New Approach to Textual Analysis using Large Language Models: Application to the Analysis of Recent Wage and Price Developments in Japan

Research and Statistics Department
IZAWA Kimihiko*, KAMEI Ikuo, SHIBATA Nao
TAKAHASHI Yusuke, YONEYAMA Shunichi

March 2025

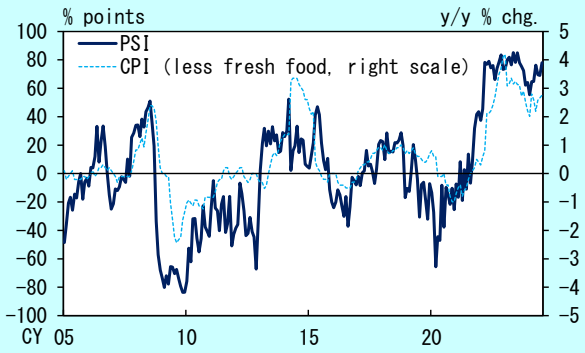
This paper examines whether textual data analysis using Large Language Models (LLMs) can be applied to assessing economic activity and prices in light of the rapid development of LLMs in recent years. LLMs have advantages in that there are a wide range of models available for use without large initial costs and that these models, which have already acquired basic knowledge of language, can analyze any topic or text and are beginning to be used in economic analysis more widely, including those of central banks. This paper, as an example, attempts to use LLMs to analyze recent wage and price developments in Japan using comments from the Cabinet Office's Economy Watchers Survey. The results suggest that the cause of increasing selling prices is gradually shifting from raw material costs to labor costs.

Introduction

The Bank of Japan has been increasingly using textual data to assess economic activity and prices for monetary policy making. Textual data, like high frequency data, high granularity data, etc., is one type of alternative data with characteristics different from traditional statistical data, and is one useful piece of information which can be used to assess economic activity and prices.[1] For example, the Cabinet Office's Economy Watchers Survey publishes an index showing business confidence along with comments on the background context, and the Bank of Japan has published several pieces of research on textual analysis using these comments.

Chart 1 shows the Price Sentiment Index (PSI) created using the textual data.[2] The PSI is calculated by extracting price-related comments using the machine learning method,[3] classifying them into the categories of "inflation," "deflation," and "zero inflation," and then dividing the number of "inflation" comments minus the number of "deflation" comments by the number of all price-related comments. The Economy Watchers Survey is released earlier than the Consumer Price Index (CPI) for Japan, and the PSI tends to correlate with the CPI several months in advance, making it a useful leading indicator.

[Chart 1] Price Sentiment Index (PSI)

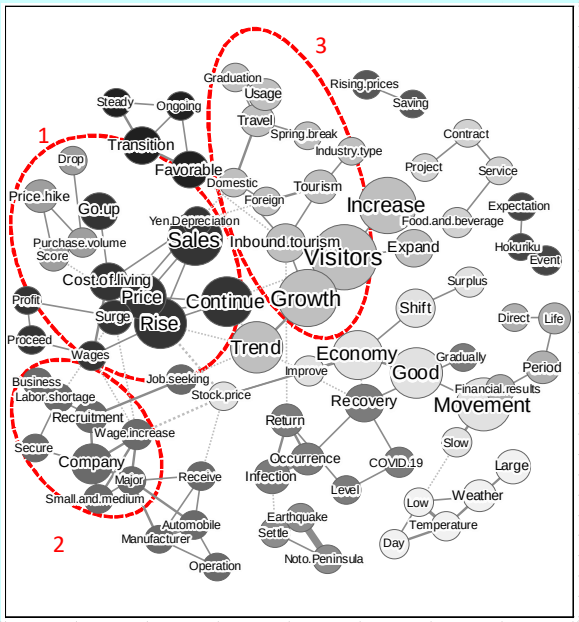


Source: Cabinet Office, Ministry of Internal Affairs and Communications

Chart 2 is a co-occurrence network diagram, which quantitatively visualizes the qualitative information behind the business confidence index that cannot be understood by the index itself. Specifically, we extract "characteristic words," i.e. words that are used more frequently compared with previous years, from the comments in the particular month and then measure and visualize the frequency with which those "characteristic words" appear in the same comments (the "co-occurrence relationships"). For example, using the co-occurrence relationships, we can confirm three topics from the March 2024 survey: 1) one related to price increases and consumption, 2) one related to wage increases and labor shortages at small and medium-sized firms, and 3) one related to inbound and spring tourism demand.

However, these analyses use a relatively simple method,[4] simply measuring word frequencies and co-occurrence relationships. On the other hand, in recent years, Large Language Models (LLMs)—machine learning methods that can even take into account the context of a sentence—have rapidly developed and are becoming more widely used. LLMs can accurately and efficiently convert unstructured data[5] such as textual and image data into structured data that is easy to analyze and interpret.[6] Therefore, in this paper, we employ this cutting-edge technology and attempt to use them in assessing economic activity and prices.

[Chart 2] Co-occurrence Network Diagram (March 2024)



Source: Cabinet Office

Note: Characteristic words were extracted from comments on current economic conditions in the March 2024 survey, and their co-occurrence relationships were plotted. The size of the words indicates the frequency of occurrence, and the thickness of the lines between the words indicates the strength of the co-occurrence relationship. This chart shows an English translation of a co-occurrence network diagram created in Japanese.

Overview of LLM

LLMs are natural language processing models built using deep learning techniques with a large amount of textual data as input data. It is characterized by the enormous amount of computation, the amount of input text for training, and the number of parameters for describing the model compared to conventional natural language processing models.[7]

The model that triggered the rapid spread of LLMs was BERT (Bidirectional Encoder Representations from Transformers), released by Google in 2018. BERT showed a significant improvement in performance compared to earlier natural language processing models and became the baseline for subsequent LLMs. Since then, models have continued to evolve in both size and performance, with Open-AI releasing GPT-3 (Generative Pretrained Transformer) in 2020 and GPT-4 in 2023.[8, 9] GPT is the LLM used in Chat-GPT, the use of which has expanded rapidly in recent years.

The process of language learning in LLMs can be broadly divided into two stages. The first stage is called "pre-training," in which unsupervised learning[10] is performed using a large amount of textual data generated from Wikipedia or other sources to acquire basic knowledge of the language. This stage is computationally very expensive, and is a difficult process without extremely large-scale computing resources. The second stage is called "fine-tuning," in which the model is adjusted with additional learning specific to the purpose. This stage has relatively small computational cost and can be performed without large-scale computing resources.

When using LLMs for analysis, it is possible to obtain a model that has already undergone the abovementioned "pre-training" and conduct only the second stage of "fine-tuning" according to the purpose of analysis. Furthermore, if a "fine-tuned" model is already publicly available for the intended use, it is also possible to use such a model directly for the analysis. This is possible because LLMs have already acquired the ability to read the context of language at the "pre-training" stage.

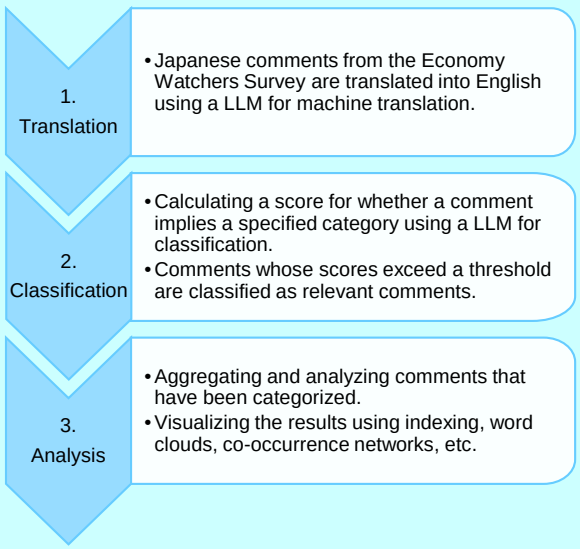
Thus, LLMs have the advantage that highly accurate models that have an understanding of the context of the text can be used relatively easily. In light of these advantages, some central banks are beginning to utilize LLMs in assessing economic conditions and research. For example, the Reserve Bank of Australia (RBA) is using LLMs to analyze corporate pricing behavior using text data such as transcripts of corporate earnings press conferences or records of interviews with companies.[11] In addition, the Federal Reserve Bank of New York is also using a LLM to analyze textual data such as Federal Open Market Committee (FOMC) minutes and the Tealbooks.[12] Given these trends, this paper examines whether textual data analysis using LLMs is useful in assessing economic activity and prices in Japan.

Analysis Process

In this section, we will explain the process of our analysis and introduce the LLMs used. Then, in the next section we will analyze recent wage and price developments in Japan. We used comments from the

Economy Watchers Survey as textual data. The process is as shown in Chart 3. First, 1) use a LLM for machine translation to translate the comments from Japanese into English. Next, 2) use a LLM for classification to determine the category of each translated comment. Finally, 3) perform analysis, making index, and visualization based on the determined categories.

[Chart 3] Analysis Process



Mechanical Japanese-English Translation

Our analysis is unique in that the comments are translated into English at the preprocessing stage. The classification model used in the analysis is superior in that it has been "fine-tuned" specifically for the task of classification, but the model only supports English. Therefore, by utilizing a machine translation model, we efficiently converted comments in the Economy Watchers Survey written in Japanese into English. Here, we employed a model called NLLB (No Language Left Behind) released by Meta.[13]

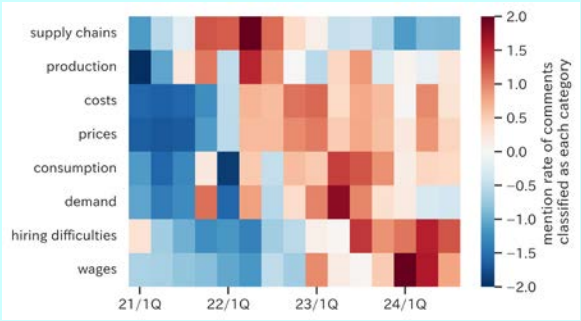
While many LLMs have been developed based on English texts, non-English speaking countries do not necessarily have LLMs that are compatible with their native languages and suitable for their purpose of analysis. In contrast, the machine translation model utilized in this paper supports as many as 200 languages, and once translated into English, it has the potential to open the door to textual analysis using LLMs for those in many non-English-speaking countries.

Classification

Next, we use a classification model to determine the topic that each translated comment implies. The model used here takes a comment and shows the probability that the comment belongs to the specified category as a score between 0 and 1 (the closer to 1, the higher the probability that it belongs to the category in question).[14] A similar classification model was used in the aforementioned RBA study, and we followed the study and determined that a comment belongs to the category when its score exceeded a certain threshold (0.7).

As a simple example of analysis using the classification model, we visualize the transition of topics using a heat map. Chart 4 shows the developments of the proportion of comments in the Economy Watchers Survey that were classified to each specified category, such as "production," "prices," and "wages" (the proportions are standardized as deviations from the historical average). Positive (red) means that the proportion of comments classified to a specific category is higher than the historical average, and negative (blue) means that the proportion is lower than the historical average. The results show that from the second half of 2021 to the second half of 2022, when supply constraints such as semiconductors was a hot topic, the proportion of comments on "supply chains" and "production" was higher than the historical average. On the other hand, since the second half of 2022, comments related to "costs" and "prices," have increased compared to the historical average, and looking more recently, comments on employment-related topics, such as "wages" and "hiring difficulties," have increased compared to the past. By using the classification model, it is possible to understand what the respondents were thinking in answering their business sentiment.

[Chart 4] Heat Map by Category



Source: Cabinet Office

Note: The mention rate for each category indicates the standardized ratio of the number of comments classified as each category to the total number of comments.

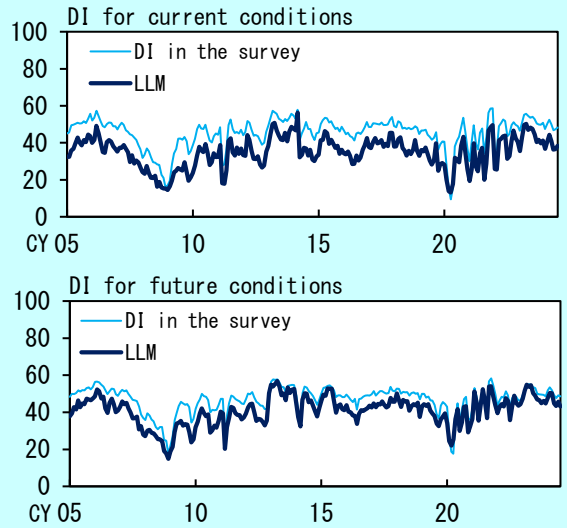
An analyst can set the categories as they wish. For example, they can specify the words themselves, such as "production," "prices," "wages," or can include changes, such as "production increase," "price increase," "wage increase." The aforementioned RBA study also used categories that included words implying changes in economic variables in their analysis. Furthermore, since the method computes a

score for each category, some comments may be classified as belonging to more than one category. Therefore, by calculating the percentage of comments classified as category A that are also classified as category B, it is possible to verify the linkage between categories. In the next section, we will take advantage of these features and deepen our analysis.

Sentiment Analysis

To check the performance of LLMs on textual data related to the Japanese economy, we will use an LLM that can classify the sentiment of text. In the Economy Watchers Survey, a five-level assessment of the current and future economic conditions and comments on the reasons for the assessment are simultaneously published for each respondent. Therefore, if the DI generated from the respondent's comments using the sentiment classification model can reproduce the DI in the survey, the performance of the LLM can be evaluated as high. Here, we use a model called FinBERT, which is fine-tuned for measuring financial and economic sentiment and published by Prosus AI. Since the model can be used only for English, as mentioned above, we measure sentiment after translating comments from Japanese into English using a machine translation model.

[Chart 5] DI



Source: Cabinet Office

Chart 5 shows the DI created from the results of measuring the sentiment (positive, negative, or neutral) of each comment using the sentiment classification model. The DI was created as (Positive comment ratio *100 + Neutral comment ratio*50 + Negative comment ratio*0)/100. When comparing the DI created using the LLMs with the DI in the survey for current and future economic conditions, they show very similar behaviors. Thus, the sentiment measured by the LLMs can be evaluated as successfully capturing the context and sentiment of respondents' comments. In other words, the LLMs used in this study are considered to have a reasonable level of accuracy in both machine translation and sentiment classification.[15]

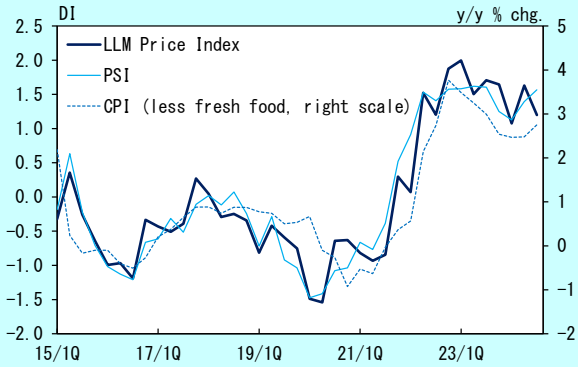
Analysis of Recent Wage and Price Developments

Using the LLMs outlined in the previous section, this section presents analysis of recent wage and price developments using comments from the Economy Watchers Survey.[16]

First, we created a price index using LLMs based on the same idea as the PSI mentioned in the Introduction. The PSI is created by using machine learning techniques to determine whether each comment refers to "inflation" or "deflation" and indexing it. Here, on the other hand, the index was created by using the LLM to classify each comment as "inflation" or "deflation" in the context.[17]

Chart 6 shows that the index calculated using the LLM (the LLM price index) shows the same trends as the PSI, whose model is trained to specialize in classifying inflation or deflation. The LLM price index is generally successful in capturing the recent increase in prices, and like the PSI, is considered to be useful in measuring price trends. In addition, compared to the PSI, which is calculated based on the occurrence of specific words selected by an analyst, the LLM price index has the advantage of being less arbitrary, as the analyst does not select words but the LLM instead reads the context of the comments.

[Chart 6] LLM Price Index



Source: Cabinet Office, Ministry of Internal Affairs and Communications

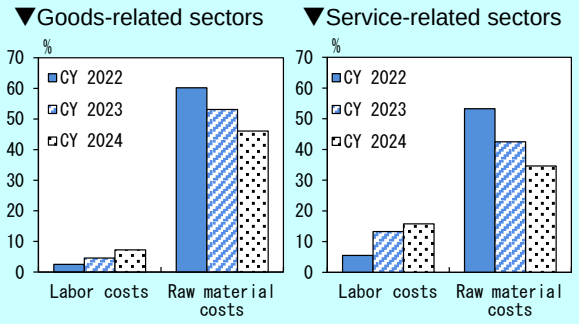
Note: The LLM price index and the PSI are standardized using data within the period shown in the graph.

Next, we analyze the background of recent price hikes in Japan. Using the LLM, we calculate the percentage of comments that are classified as "labor

costs" or "raw material costs" among the comments that are classified as "final prices increasing." We calculate the same percentages for both goods-related and service-related sectors and check the changes in the factors behind price hikes.

Chart 7 shows that in both sectors, the percentage of comments that belong to raw material costs has decreased since 2022, while the percentage of comments that are classified as labor costs has increased, albeit at a low level. This indicates that the respondents are gradually paying more attention to labor costs. When comparing sectors, service related sectors have more comments classified as labor costs and less comments classified as raw material costs than good-related sectors.

[Chart 7] Background of Price Hikes



Source: Cabinet Office

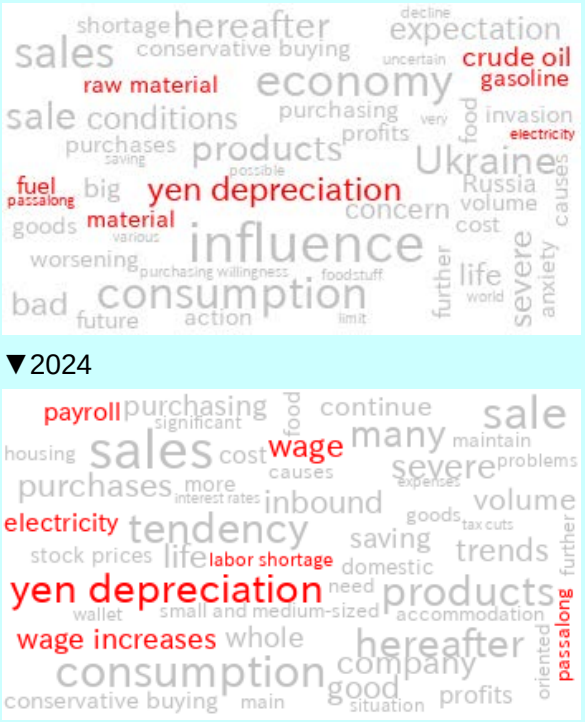
Note: The chart shows the percentage of comments that are classified as "labor costs" or "raw material costs" among those that are classified as "final prices increasing" using the classification model. The comments are on current and future economic conditions from January to September of each year. Goods-related sectors are manufacturing and retail-related. Service-related sectors are non-manufacturing (excluding construction), food and beverage-related, service-related, housing-related, and employment-related.

Furthermore, we also checked the factors behind changes in price increases by creating word clouds of comments that are classified by the LLM to imply "final prices increasing."[18] Here, we show the results on comments in 2022 and 2024.[19]

Chart 8 shows that while words related to raw material costs such as "crude oil" and "material" were prominent in 2022, words related to labor costs such as "wage increases" and "payroll" were prominent in 2024, suggesting that the cause of price increases for respondents of the Economy Watchers Survey has been shifting from raw material costs to labor costs. Meanwhile, the "yen depreciation," which is related to raw material costs, was referred frequently in both 2022 and 2024, indicating that it has continued to be recognized as a factor pushing up prices.

Using LLMs can reduce the burden of preparation when conducting this kind of textual analysis. For example, the PSI requires the creation of training data and estimation of the model. On the other hand, LLMs have already learned general linguistic knowledge, so with the LLM for classification, for example, users can classify categories simply by specifying any word (such as "price increase/decrease") without the additional effort of training models on their own. In addition, since many fine-tuned LLMs are publicly available, including the machine translation or classification models used in this study, users can immediately begin analysis according to their purpose.

[Chart 8] Word Cloud



Source: Cabinet Office

Note: Comments that are classified as "final prices increasing" were extracted using a classification model from comments on the current and future economic conditions for January to September of each year. A word cloud was created to show words that appeared relatively frequently in comments for the year analyzed, compared to the entire extracted comments. The size of the words indicates the frequency of appearance, and red text indicates words that are considered to be the background of the price increases (judged by the authors). Note that only nouns, adjectives, and adverbs were included here. However, words related to price increases themselves (prices, high prices, rise, price hike, price increase, go up, unit price, fee, high, expensive, inflation, price, price soaring) were excluded. This chart shows an English translation of word clouds created in Japanese.

Conclusion

This paper examines whether textual analysis using LLMs, which have been rapidly developing in recent years, can be applied to assessing economic activity and prices. LLMs have advantages that a variety of models created for different purposes are available and that they are highly versatile as LLMs have already

acquired general knowledge of the language. The results of our analysis using LLMs on the developments of wages and prices suggest that labor costs are becoming a major factor in the recent rise in sales prices.

Finally, we will comment on some points to keep in mind when conducting textual analysis using LLMs. The first point is the bias in the textual data to be analyzed, which applies not only to analysis using LLMs but also to textual data analysis in general. It is practically difficult to ensure that textual data is unbiased. There may be bias due to the attribute distribution of respondents, bias due to the same respondents repeatedly answering, bias due to the questions themselves being guiding, or bias in the answers themselves, for example, the comments from dissatisfied respondents being exaggerated. When conducting textual analysis, it is necessary to constantly check the bias in the textual data and interpret the results while recognizing the bias.

The second point is that there tend to be some noise in textual analysis using LLMs. For example, in the classification model, some of the comments that were classified as "final prices increasing" also included comments on price increases of intermediate goods. At this point in time, it is difficult to completely eliminate such noise, and it should be noted that not all classification results are necessarily as intended by the analyst. In addition, the classification process is a black box, making it difficult to understand the reasons behind the results. Therefore, it would be necessary to check the contents of the classified comments and to confirm whether the results are valid.

The third point is the computational cost. The burden of preparing models for analysis using LLMs is small because pre-trained models are easily available. On the other hand, analysis using LLMs can impose a large computational load compared to conventional textual analysis methods though it will not require the extremely large computational resources which is essential for "pre-training." Even if the processing time for a few lines of textual data may be negligible, depending on the amount of data and the scale of models considerable computational resources and time might be required compared to conventional methods. In addition, technological progress in the field of LLMs is extremely rapid, and new high-performance models are being released continuously. Thus, it is also important to follow the latest trends and select a model suitable for the purpose of analysis.

In this paper, we have examined the possibility of using LLMs to assess economic conditions, while discussing the advantages and disadvantages of LLMs. Although statistical data is the basis of assessing economic conditions, it is also useful to utilize alternative data, including unstructured data, such as through textual analysis using LLMs, as shown in the analysis in this paper.

---

* Currently at the Osaka Branch

1 The research by the Bank of Japan using alternative data is summarized in the website below.

https://www.boj.or.jp/en/research/bigdata/index.htm

For an overview of such research, see the following paper.

Kameda, S. (2022), "Use of Alternative Data in the Bank of Japan's Research Activities," Bank of Japan Review Series, No. 2022-E-1.

2 For details on the PSI, see the following paper.

Nakajima, J., H. Yamagata, T. Okuda, S. Katsuki, and T. Shinohara (2021), "Extracting Firms' Short-Term Inflation Expectations from the Economy Watchers Survey Using Text Analysis," Bank of Japan Working Paper Series, No. 21-E-12.

3 The PSI is created using a machine learning technique called the Naive Bayes Classifier.

4 For example, the PSI is calculated by focusing on the "words" mentioned, and does not take into account the "context" of the comments.

5 Unstructured data is the data that has not been neatly formatted into Excel files, CSV files, etc. Therefore, it is difficult to directly use unstructured data for analysis.

6 For the application of deep learning technology to economics, including LLMs, see the following paper.

Dell, M. (2025), "Deep Learning for Economists," *Journal of Economic Literature*, Vol. 63(1), pp.5-58.

7 Conventional methods used for natural language processing include the Naive Bayes Classifier used to create PSI or methods using deep learning such as models that apply RNN (Recurrent Neural Network) and its derivative LSTM (Long Short-Term Memory).

8 While BERT has 340 million parameters, GPT-3 has 175 billion, and GPT-4, although not disclosed, is considered to have a much larger number of parameters than GPT-3.

9 The General Language Understanding Evaluation (GLUE) Benchmark is a measure of natural language processing performance. It scores the performance of natural language processing models based on an overall evaluation of tasks such as judging grammatical correctness, sentiment, and sentence implication, and LLMs that outperform human scores have appeared. Note that GLUE is a benchmark for evaluating English expressions, but there is also JGLUE as a benchmark for evaluating Japanese expressions.

10 Unsupervised learning is a method that learns patterns from the data itself. In contrast, supervised learning is a method that enables the model to predict the label to which a sentence belongs based on data that pairs sentences with labels. For example, the aforementioned PSI was created using a model estimated using labeled text such as price increase/decrease, in other words, a model that uses supervised learning.

11 Windsor, C., and M. Zang (2023), "Firms' Price-setting Behaviour: Insights from Earnings Calls," Reserve Bank of Australia Research Discussion Paper, 2023-06.

12 Fischer, E., R. McCaughrin, S. Prazad, and M. Vandergon (2023), "Fed Transparency and Policy Expectation Errors: A Text Analysis Approach," Federal Reserve Bank of New York

Staff Reports, No.1081.

[13] Specifically, we used "facebook/nllb-200-distilled-600M" published on the Hugging Face website.

[14] Specifically, we used "MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli" published on the Hugging Face website. The model is based on the DeBERTa-v3-base model developed by Microsoft. In this paper, we used a simple general-purpose model to analyze a wide range of topics related to the economy. However, if one focuses on more specialized topics, it is important to carefully check the training data and select a model according to the purpose. In addition, using a larger, cutting-edge models may further improve the accuracy of the analysis.

[15] In general, the results of sentiment measurements can be difficult to interpret because different economic agents perceive the same event differently. For example, wage increases can be a positive event for consumers but possibly a negative event (i.e., rising costs) for companies. Thus, sentiments scores can be reversed depending on the position of the respondents. The results of the sentiment measurements should be used with a careful assessment of the attributes of the respondents.

[16] For an analysis of recent wage and price developments in Japan, see, for example, the following paper.

Ozaki, T., M. Jimbo, T. Yagi, and A. Yoshii (2024), "Recent Developments in the Linkage between Wages and Prices," Bank of Japan Review Series, No. 2024-E-2.

[17] The model determines whether a sentence implies "price increase/decrease," respectively, and if the score is greater than 0.7, the sentence is classified as "price increase/decrease." The index is calculated as (number of price increase comments - number of price decrease comments) / total number of comments.

[18] Compared to machine learning methods based on the frequency of word occurrence, classification using LLMs has the disadvantage that the classification process is a black box and hard to understand. To address this issue, it would be effective to examine the frequency of word occurrences in the classified text using word clouds or co-occurrence network diagrams, etc., and verify whether the classification is accurate.

[19] When drawing the word cloud, words related to "final prices increasing" are excluded. This is because, since the word cloud is drawn for comments classified as "final prices increasing," the related words are displayed large in the word cloud, making it difficult to see changes in words other than those related to final prices among the classified comments.

The Bank of Japan Review Series is published by the Bank to explain recent economic and financial topics for a wide range of readers. This report, 2025-E-5, is a translation of the Japanese original, 2024-J-14, published in December 2024. Views expressed are those of the authors and do not necessarily reflect those of the Bank. If you have any comments or questions, please contact Research and Statistics Department (E-mail:post.rsd22@boj.or.jp). The Bank of Japan Review Series and the Bank of Japan Working Paper Series are available at https://www.boj.or.jp/en/index.htm.