Attempt to effectively utilize text data by applying AI technologies

(Summary)

This paper applies AI technologies, including large language models (LLMs), to analyze textual data from financial institutions' disclosure publications. By extracting texts related to designated themes from large volumes of text, the analysis identifies institution-type-specific characteristics and time-series trends. Focusing on real estate and housing loans, the results reveal that key terms and descriptive patterns differ across institution types. The proposed approach is expected to offer more efficient information gathering and the early detection of emerging signals. Future efforts will continue to incorporate text data analysis into monitoring frameworks, while carefully considering the specific risks associated with AI.

I. Introduction

In recent years, the rapid advancement of AI technologies—particularly in the field of text analysis—has accelerated the development and widespread application of natural language processing tools. The FSA has been working to enhance data utilization to better understand the management conditions of individual financial institutions, vulnerabilities and resilience of the financial system, and broader market trends. Such insights can be drawn not only from quantitative data—such as loan-level data and equity transaction data—but also from unstructured text data, including various reports and publications related to financial services. Leveraging AI for text analysis is expected to enhance the effectiveness of information gathering and facilitate the identification of new analytical perspectives, thereby strengthening FSA's monitoring capabilities.

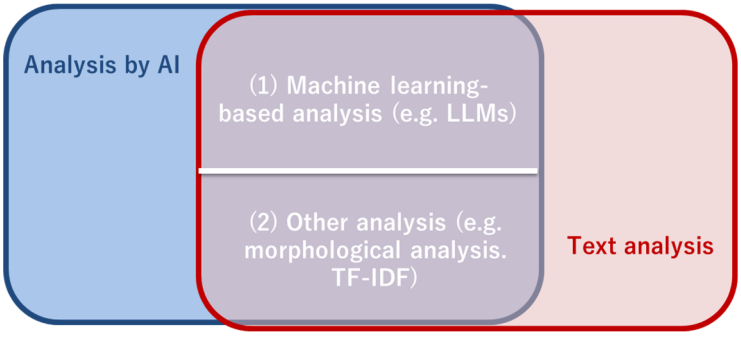
This paper attempts to apply text analysis techniques, including large language models (LLMs[1]; see Box 1), to extract and analyze targeted information from financial institutions' disclosure publications[2]. Specifically, the analysis focuses on identifying institution-type-specific characteristics and tracking time-series trends in descriptions related to selected themes using LLM-based methods.

[1] LLM is natural language processing models built using deep learning techniques trained on large volumes of text data.
[2] In this paper, "disclosure publications" refer to the sections of annural reports and integrated reports that describe business activities and related information, excluding detailed financial data (e.g., "Data Section") and accompanying technical notes. When both types of documents are published, the integrated report is used in principle.

BOX 1: Classification of text analysis

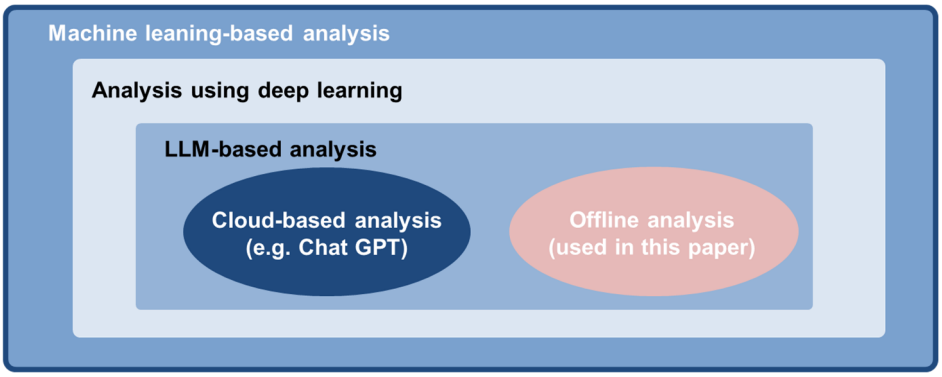
This paper categorizes AI-based text analysis into two approaches: (1) machine learning-based analysis and (2) rule-based analysis without machine learning (Figure 1). The former involves training data to build models, while the latter relies primarily on predefined rules and formulas. Both approaches are utilized in this paper. For (1), large language models (LLMs) are applied; for (2), morphological analysis—breaking down text into individual words—and TF-IDF[3] (Term Frequency–Inverse Document Frequency) are used to evaluate the importance of terms.

Figure 1: Classification of text analysis



Within machine learning-based text analysis, approaches using deep learning are often distinguished from other machine learning methods. Among deep learning techniques, analyses using LLMs—which are trained on vast amounts of text data—can be considered a separate category. LLM-based analysis has recently evolved into two main approaches: cloud-based analysis using services such as ChatGPT, and offline analysis using publicly available AI models (Figure 2). This paper employs the latter, conducting LLM-based analysis in an offline environment.

Figure 2: Analysis by machine lerning



[3] "TF-IDF" is a commonly used metric for evaluating the importance of individual words within a document (see Box 2 for details).

II. Text analysis on disclosure publications

Although disclosure publications are required to include certain items under the legislation, their formats are not standardized like securities reports, and vary across financial institutions. As a result, rule-based analysis alone is insufficient to comprehensively extract information on specific themes or to conduct cross-institutional comparisons. In contrast, advanced AI models such as LLMs are capable of interpreting unstructured and non-standardized texts, making it possible to perform meaningful analysis even in the absence of uniform formatting.

Sub-section 1 outlines the process used in this study to analyze disclosure publications using AI. Sub-section 2 presents the results of an analysis—focused on the theme of "real estate and housing loans"—conducted on five years of disclosure publications[4] (fiscal years 2019 to 2023, covering publication years 2020 to 2024) from major banks, regional banks, and others (internet only banks)[5].

1. Process of text analysis using AI

Figure 3 illustrates the text analysis process employed in this paper. First, text data from disclosure publications are collected. Then, using LLMs, each sentence is classified based on whether it is relevant to a selected theme (Figure 4). Sentences identified as relevant are extracted and then term frequency analysis is conducted using TF-IDF (see Box 2), ranking the most salient keywords. In addition, LLMs are used to evaluate changes in content, such as year-over-year comparisons. This process enables identification of institution-specific keywords associated with the selected theme, tracking their evolution over time, and assessing whether textual descriptions have become more detailed or comprehensive compared to the previous year.

To mitigate risks such as hallucinations—i.e., the generation of inaccurate or unfounded information—the LLMs' classification decisions and content evaluations are accompanied by explanation outputs, which are then reviewed and corrected by human analysts as necessary.

[4] Disclosure publications available on each financial institution's website as of January 2025 were used as the analysis target.
[5] The analysis covers banks and banking groups that publish disclosure publications. In this paper, "major banks" refer to Mizuho Financial Group, Mitsubishi UFJ Financial Group, Sumitomo Mitsui Financial Group, Resona Holdings, Sumitomo Mitsui Trust Group, SBI Shinsei Bank, and Aozora Bank. "Regional banks" refer to banks and groups that are members of the Regional Banks Association of Japan and the Second Association of Regional Banks. "Others (internet only banks)" include PayPay Bank, Seven Bank, Sony Bank, Rakuten Bank, SBI Sumishin Net Bank, au Jibun Bank, AEON Bank, Daiwa Next Bank, Lawson Bank, and GMO Aozora Net Bank.

Figure 3: Test analysis process using LLMs developed in this paper

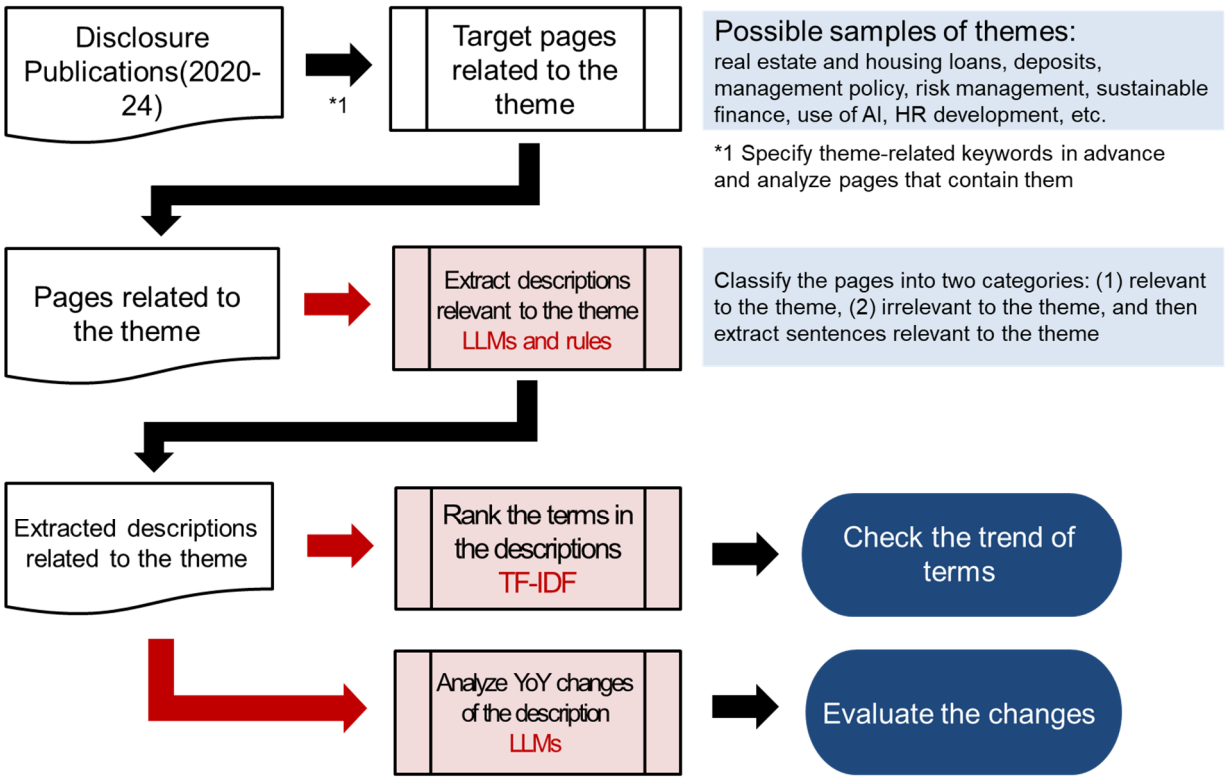


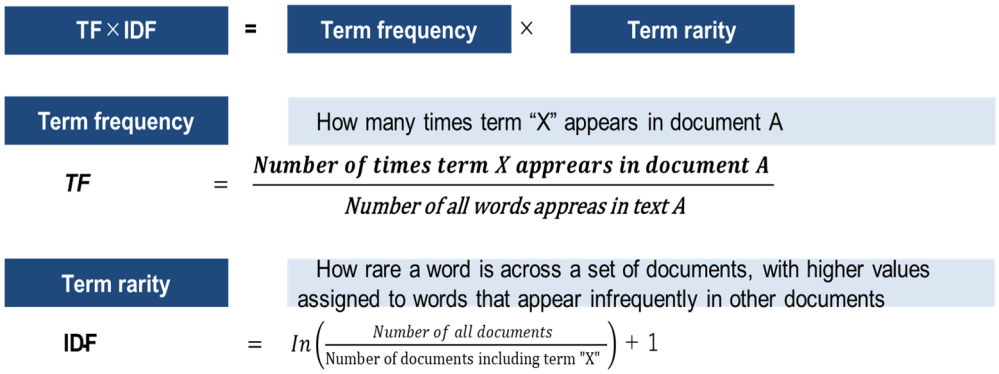
Figure 4: Category classification by LLMs

Categories	Rationale	Samples of the rationale provided by LLMs (cases where “real estate and housing loans” is selected for the theme)
Relevant to the theme	Description directly related to the theme or sentences which include words relevant to the theme	An apartment loan is a type of real estate financing that can be used when purchasing or constructing an apartment or condominium for investment purposes or for porposes other than personal residence.
Irrelevant to the Theme	Description Irrelevant or only have indirect relation with the theme	The description does not directly refer to real estate financing or housing loans, but rather focuses on XX.

BOX 2: TF-IDF

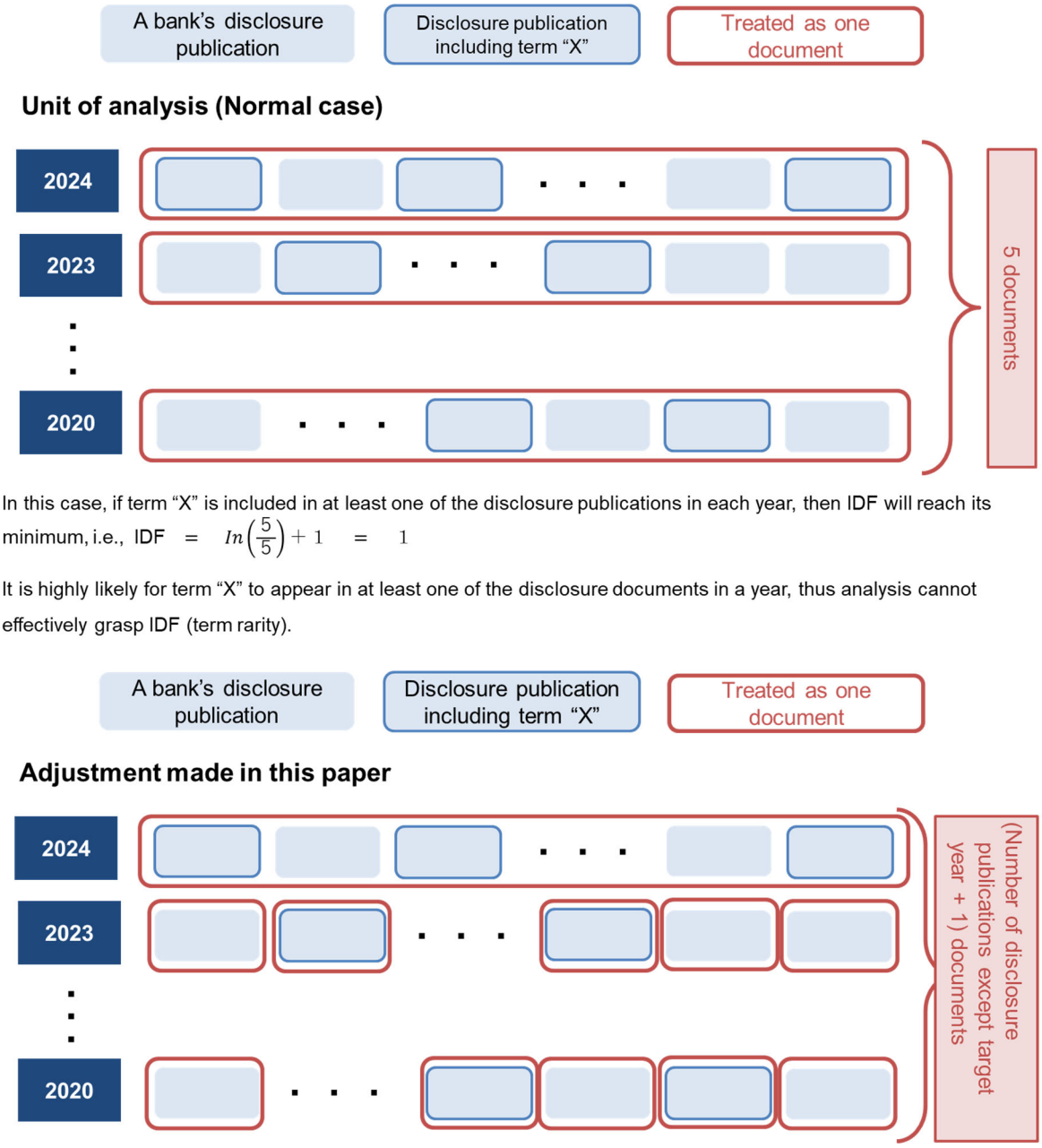
TF-IDF (Term Frequency – Inverse Document Frequency) is a commonly used metric for assessing the importance of individual words within a document. It is calculated by multiplying the term frequency (TF) and the term rarity (IDF). Term frequency is computed by dividing the number of times a specific word appears by the total number of words in the document. Term rarity reflects how rare a word is across a set of documents, with higher values assigned to words that appear infrequently in other documents. Words with high TF-IDF scores are considered to be key terms that characterize the document (Figure 5).

Figure 5: Calculation method of TF-IDF



For the TF-IDF calculation in this paper, specific adjustments were made to the unit of analysis, as shown in Figure 6, to enable year-by-year comparisons (2020–2024) of key terms by bank or banking group. This approach avoids aggregating disclosure content at the annual level, which would otherwise cause the IDF values of many terms to reach their minimum across all years—thus limiting the influence of IDF in the evaluation. Additionally, only nouns were included in the TF-IDF calculation, with other types of words excluded.

Figure 6: Adjustment made in this paper on calculation of IDF



2. Analysis on the theme of “real estate and housing loans”

This chapter presents an example analysis focused on the theme of “real estate and housing loans” by using disclosure publications of major banks, regional banks, and others (internet only bank). The theme was selected as it is the topic addressed across institution types and with consistent coverage in disclosure publications.

Figure 7 summarizes the results of a TF-IDF-based analysis conducted on text extracted from disclosure publications related to this theme, aggregated by institution type. Among major banks, the term “real estate” consistently ranks high, and terms such as “finance” also appear prominently, reflecting a notable volume of text related to structured and real estate finance. In addition, The term “corporate” appears within the top 10 rankings in all years except 2021. In regional banks, “housing loan” ranks higher than in major banks, and “risk” has been the top-ranked term over the past three years. For others (internet only banks), “housing loan” was the highest-ranked term through 2023, with other frequently appearing terms including “service” and “launch.” In 2024, however, terms such as “exposure” and “credit risk” emerged in the rankings.

Additionally, in 2024, environment-related terms such as “emission,” “environment,” and “carbon” appeared in the top 10 rankings for both major and regional banks. This likely reflects the growing number of references to environmentally conscious housing loans, suggesting a rising awareness of sustainability even in the context of real estate and housing loans.

Figure 7: List of TF-IDF on “real estate and housing lonas” by industry type

real estate/finance
housing loan
risk/exposure
environment/sustainability

*Note that the terms are derived from an analysis of the original Japanese text and then translated directly into English, the nuances may differ from those of original Japanese.

Major banks

	2020	2021	2022	2023	2024
1	real estate	real estate	real estate	real estate	real estate
2	trust	finance	finance	sector	risk
3	finance	asset	risk	finance	sector
4	asset	trust	corporate	value	finance
5	relevance	risk	ESG	emission	business
6	risk	management	environment	creation	value
7	provision	corporate	initiative	risk	environment
8	business	housing loan	profit	business	creation
9	management	business	individual	investment	corporate
10	corporate	individual	business	zero	emission

Regional banks

	2020	2021	2022	2023	2024
1	exposure	housing loan	risk	risk	risk
2	housing loan	launch	housing loan	group	result
3	trust	balance	balance	scenario	increase
4	balance	support	group	increase	group
5	group	group	service	housing loan	emission
6	support	exposure	support	balance	carbon
7	fund	fund	launch	real estate	sector
8	securitization	sales	exposure	support	value
9	lending	service	real estate	agent	support
10	real estate	risk	increase	service	balance

Others (Internet only banks)

	2020	2021	2022	2023	2024
1	housing loan	housing loan	housing loan	housing loan	exposure
2	launch	launch	launch	launch	housing loan
3	service	service	service	service	method
4	balance	exceeding	exceeding	balance	real estate
5	guarantee	rating	account	account	credit risk
6	special agreement	consolidation	balance	exceeding	eligible
7	interest rate	balance	consolidation	non-consolidated	for retail
8	exceeding	interest rate	provisionlaunch	provisionlaunch	asset
9	use	account	interest rate	deposti	delinquency
10	repayment	center	rating	interest rate	application

Next, text related to "real estate and housing loans" extracted from each bank's disclosure publication was analyzed using LLMs to compare descriptions from 2020 to 2024, by institution type. In this analysis, the LLMs were instructed to classify the nature of year-over-year changes as *positive*, *negative*, or *neutral*. It is important to note that these classifications reflect the LLMs' interpretation of textual trends within the disclosure documents and do not necessarily correspond to the actual lending stance or performance of the institutions. A review of the reasoning behind the LLMs' classifications suggests that the model makes holistic judgments. For example, an increase in loan balances may lead to a *positive* classification, while shifts in the volume of narrative content alone may also influence the assessment (Figure 8).

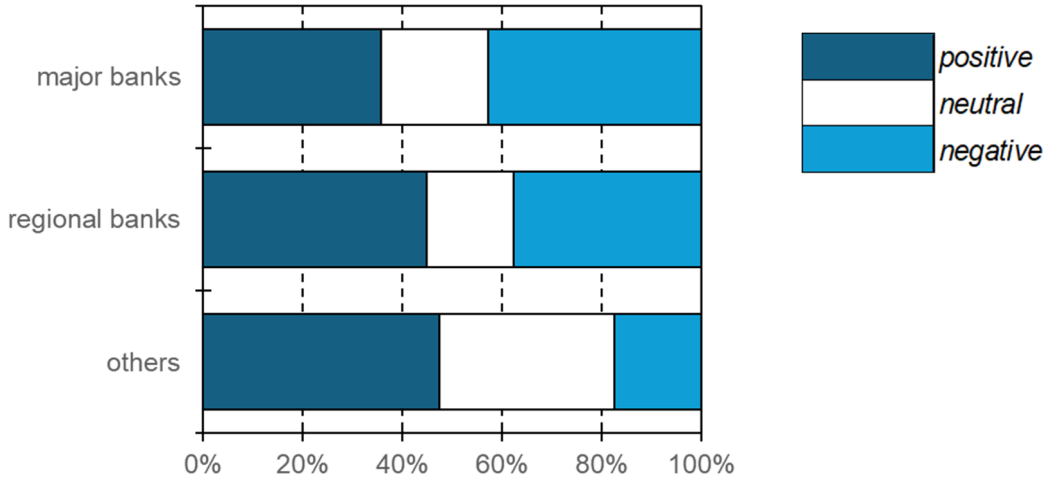
Figure 8: Reasoning of classification by LLMs

LLMs classification	Examples of rationale provided by LLMs
positive	The classification was based on the observation that the current year's content indicates an increase in housing loan balances compared to the previous year. Specifically, the text notes that the housing loan balance exceeded [XX] yen, and that initiatives such as [XX] made housing loans more accessible to a greater number of customers.
neutral	The classification was based on the observation that both this year and last year contain similar core information such as [XX], along with additional details such as [XX], and no significant differences were identified.
negative	The classification was based on the limited coverage of real estate and housing loans in this year's content. While last year's publication included references to initiatives such as [XX], no such descriptions were found this year. Furthermore, the document states [XX], and no references to real estate or housing loan-related lending activities were identified.

The analysis found that the institution type with the highest proportion of descriptions classified as *positive* was others (internet only banks), at approximately 48%. While others (internet only banks) also had a relatively high proportion of *neutral* compared to other institution types, this appears to be due to a number of disclosure publications containing little to no content related to the specified theme through years. The next highest was regional banks, with around 45% of descriptions classified as

positive. Major banks had the lowest proportion, at approximately 35%, although the share of descriptions classified as *negative* was not significantly higher (Figure 9). A review of the classification rationale shows that, for others (internet only banks), the main reasons cited were references to new services and increases in loan balances. For regional banks, increases in loan balances or the number of contracts were cited in approximately 50% of the cases. Major banks exhibited a wider variety of reasons, including factors such as portfolio shifts and management-related changes, which were rarely observed in other institution types.

Figure 9: LLM classification by industry type (5-year average)



Additionally, to assess the consistency between LLM-based classifications and actual data, a comparison was conducted for regional banks—the institution type with the largest number of financial institutions. Specifically, the year-over-year growth rates (arithmetic average of banks) of end-of-period balances for real estate and housing loans were examined against the LLM classifications (Figures 10 and 11). The results show that regional banks classified as *positive* by the LLMs exhibited higher year-over-year growth rates than those classified as *negative* , confirming certain consistency between the LLM assessments and the actual data.

Figure 10: Distribution of *positive/neutral/negative* for regional banks

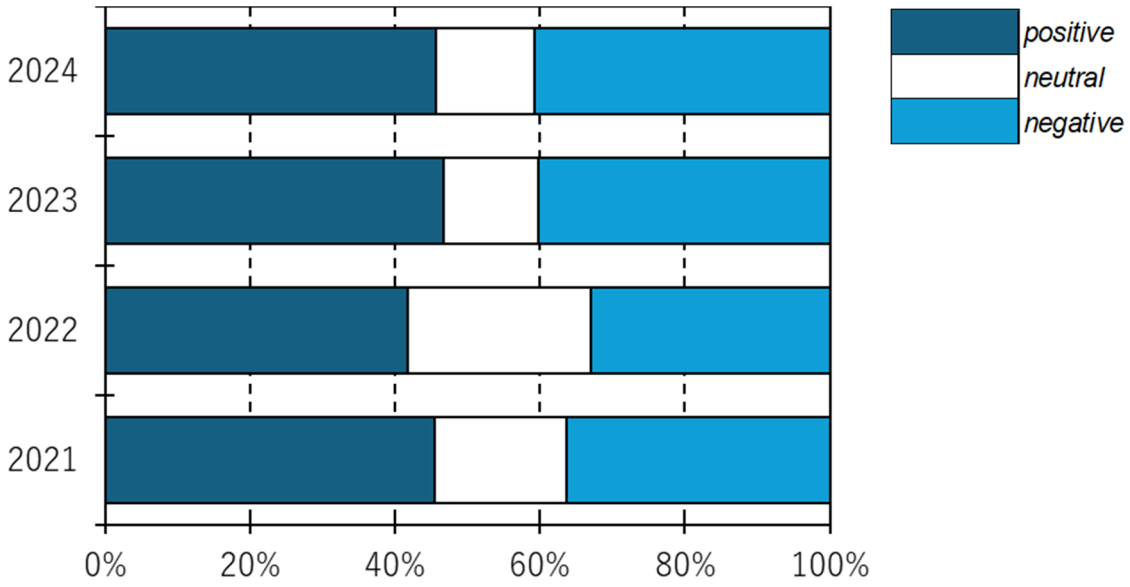


Figure 11: Growth rate of real estate and housing loan balance
(YoY, arithmetic average of banks classified as *positive/negative*)

	2022	2023	2024
positive	4.42	4.17	4.40
negative	2.96	3.02	2.64

III. Conclusion

This paper applied text analysis techniques, including LLMs, to disclosure publications in order to extract descriptions related to a specified theme from large volumes of text and to examine the appearance of characteristic terms and time-series changes in narrative content.

Focusing on the theme of “real estate and housing loans,” the TF-IDF analysis suggested that major banks frequently referenced terms such as “real estate,” “finance,” and “risk,” regional banks emphasized “housing loans” and “risk,” and others (internet only banks) primarily focused on “housing loans.” More recently, however, there has been an increasing emphasis on “sustainability,” and others (internet only banks) have begun to highlight “risk” as a key topic.

The LLM-based analysis of year-over-year changes in narrative content suggested that the

proportion of descriptions classified as *positive* was highest for others (internet only banks), followed by regional banks and then major banks. Disclsure publications classified as *positive* often included references to increased loan balances, growth in the number of customers, or initiatives that were well received by clients. Cross-checking these classifications with actual loan balance data confirmed their consistency with the LLMs' assessments.

These findings suggest that applying AI technologies to the analysis of text data—such as disclosure publications—can be effective in deepening understanding of specific themes. The analytical approach employed in this paper offers the potential to streamline the collection and interpretation of large volumes of unstructured text, which has traditionally required substantial time and human resources, thereby enhance the monitoring capabilities. At the same time, it is essential to be mindful of the specific risks associated with AI technologies such as LLMs, including limitations in model performance, prompt optimization, and model biases introduced during training. These factors can result in phenomena such as hallucination, where inaccurate or unfounded information is generated. To mitigate these risks, this analysis was designed to have the LLM output not only classification results but also the rationale behind its decisions, which were then reviewed and, if necessary, corrected by humans.

The FSA will continue efforts to advance the use of AI, including text data analysis, while managing its unique risks, with the aim of enhancing its monitoring capabilities.