

AI 技術を用いたテキストデータの解析検証

(要旨)

本稿では、金融機関のディスクロージャー誌におけるテキストデータを大規模言語モデル（LLM）等の AI 技術を用いて解析することで、大量の文章から指定されたテーマに関する記述を抽出し、記述内容の業態毎の特徴や時系列変化を把握することを試みた。不動産・住宅ローンに着目して解析を行ったところ、注目単語や記載内容の変化の傾向が業態によって異なることが確認された。本稿の解析手法を応用することで、各業態・金融機関の特徴把握など情報収集の効率化や新たな兆候の発見が期待される。引き続き、AI 特有のリスクを十分踏まえたうえで、テキストデータ分析のモニタリングへの活用に取り組んでいく。

1. はじめに

近年、AI 技術が目覚ましく発展する中で、特にテキスト分野における解析技術の進化・普及が加速している。金融庁は、個別金融機関の経営状況、金融システム全体の脆弱性・強靱性、市場動向に関する実態把握を進めるためにデータ活用の高度化に取り組んでいる。金融機関や金融システムの状況は、貸出明細データや株式市場データなどの定量データのみならず、金融サービスに関するレポートや報告書等のテキストデータからも窺い知ることができる。こうしたテキストデータに対し AI 技術を活用することで、より効果的な情報収集や新たな着眼点の発見等、モニタリングの高度化につながると期待される。

本稿では大規模言語モデル（以下、「LLM¹」）等のテキスト解析手法（BOX 1 参照）を用いて、テキストデータから指定した情報を収集し解析することを試行した。具体的には、各銀行・グループが毎年発行しているディスクロージャー誌²のテキストデータを解析対象とし、記述内容の業態毎の特徴や、LLM 等を用いて指定したテーマに関する記述内容の時系列変化の把握を試みた。

¹ 「LLM」は、「Large Language Models」の略称であり、大量のテキストデータを入力データとする、深層学習の手法を用いて構築された自然言語処理モデルのこと。

² 本稿における「ディスクロージャー誌」は、ディスクロージャー誌及び統合報告書において、「資料編」等の詳細な財務データ及びそこに付随する文章以外の、業務内容等に関する記載部分を指す。なお、両者を公表している場合は原則統合報告書を採用している。

BOX 1: テキスト解析の分類

本稿では、AI 技術を用いたテキスト解析の領域を、①機械学習によるテキスト解析と②機械学習以外のテキスト解析に分けて考える。①はデータを機械学習させてモデルを作成する解析であり、②は主に算出式を含むルールによる解析である（図表 1）。本稿では双方の解析を活用しており、①は LLM、②は文章を単語単位に分解する形態素解析及び単語の重要度を評価する TF-IDF³を用いている。

図表 1 テキスト解析の分類



①機械学習によるテキスト解析の中でも、Deep Learning と呼ばれる機械学習手法を用いた解析は、他の機械学習と区別されることが多い。さらに Deep Learning による解析の中でも、膨大な量のテキストデータを学習させて作成されたモデルである LLM による解析は、他の Deep Learning による解析手法と分けて考えることができる。また、LLM による解析としては、Chat GPT 等を用いるクラウド上の LLM による解析と、公開されている AI モデルを用いるオフラインでの LLM による解析の 2 つが近年発展してきており（図表 2）、本稿では、オフラインでの LLM による解析を実施している。

図表 2 機械学習による解析



³ 「TF-IDF」は、「Term Frequency – Inverse Document Frequency」の略称であり、文章中においてどの単語が重要かを評価するための指標の一つ（詳細は BOX2 参照）。

II. ディスクロージャー誌を対象としたテキスト解析

ディスクロージャー誌は、銀行法施行規則等により開示項目が定められているものの、有価証券報告書のように様式が統一されている文章とは異なり、金融機関ごとに様式が異なるため、単純なルールベースでの解析では、指定テーマに関連する記述を網羅的に収集したり、金融機関横断的に解析したりすることは困難である。一方で、LLM 等の高度な判断が可能な AI では、様式が統一されていない文章であっても一定の解析が可能である。

第 1 節では、本稿で使用した AI を用いたディスクロージャー誌解析のプロセスを説明し、第 2 節では、主要行等、地域銀行及びネット銀行⁴の 2020 年から 2024 年（報告対象事業年度ベースでは 2019 年度から 2023 年度）までの 5 年分のディスクロージャー誌⁵に対して、「不動産・住宅ローン」をテーマとして指定した場合の解析の結果を説明する。

1. AI を用いたテキスト解析のプロセス

図表 3 は、本稿におけるテキスト解析のプロセスを示している。具体的には、初めにディスクロージャー誌のテキストデータを取得したうえで、LLM を用いたカテゴリ分類（図表 4）によって指定したテーマに関係する文章かどうかを判断し、指定テーマに関係する文章を抽出する。そして、抽出された文章を対象に TF-IDF（BOX2 参照）による登場単語の順位付けや LLM による記載内容の前年比較等の評価を行う。この一連のプロセスにより、指定したテーマに対して、他の金融機関と比べどのような単語（キーワード）が特徴的に登場し、それが時系列でどのように推移しているかや、指定テーマに関する記述が前年より充実した等の記載内容の変化を把握することができる。

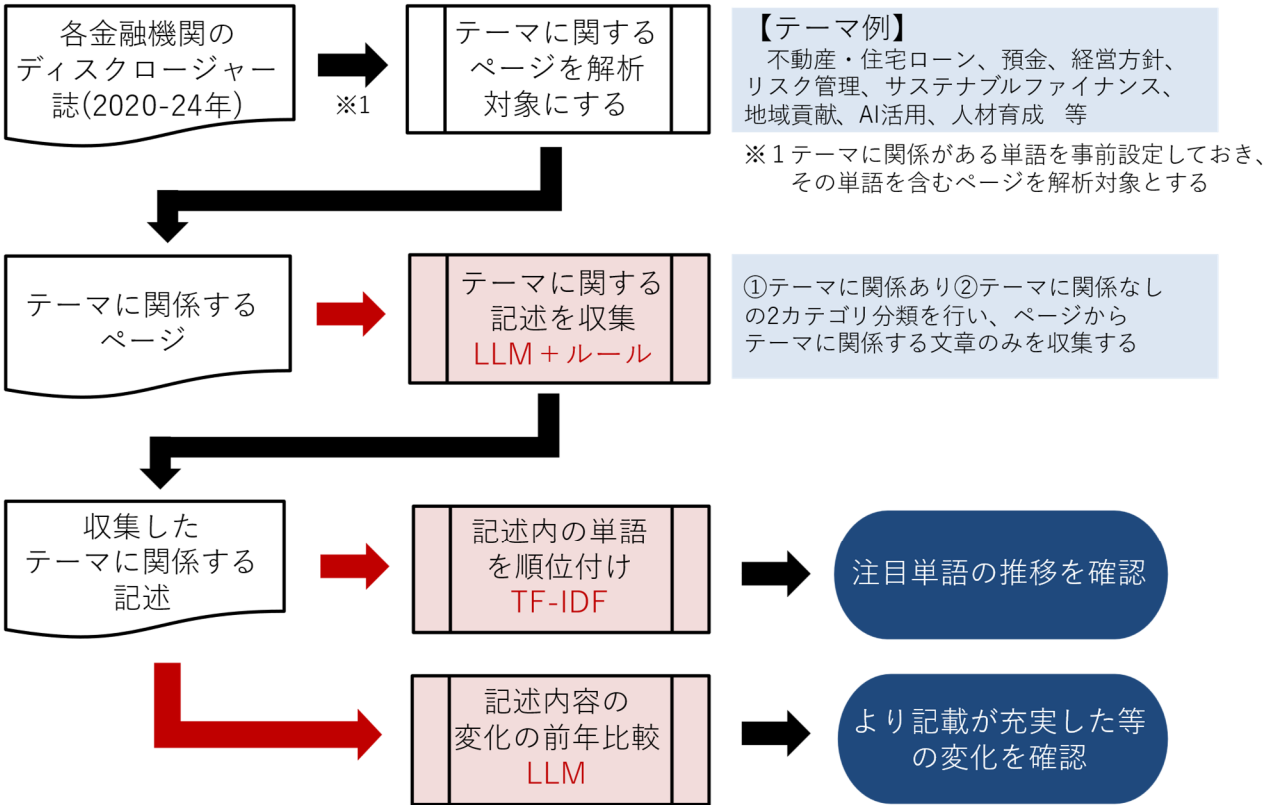
なお、LLM を用いたカテゴリ分類による判断や記載内容の評価の際は、LLM が判断した理由も併せて出力したうえで人間が確認し、必要に応じ修正をすることで、LLM が誤った情報を出力するハルシネーション等の影響を抑えている。

⁴ ディスクロージャー誌を公表している銀行・グループを対象に解析している。

本稿において、主要行等は、みずほフィナンシャルグループ、三菱 UFJ フィナンシャル・グループ、三井住友フィナンシャルグループ、りそなホールディングス、三井住友トラスト・ホールディングス、SBI 新生銀行、あおぞら銀行。地域銀行は、全国地方銀行協会及び第二地方銀行協会に加盟する銀行・グループ。ネット銀行は、PayPay 銀行、セブン銀行、ソニー銀行、楽天銀行、住信 SBI ネット銀行、au じぶん銀行、イオン銀行、大和ネクスト銀行、ローソン銀行、GMO あおぞらネット銀行。

⁵ 2025 年 1 月時点で各金融機関のウェブサイトから取得可能なディスクロージャー誌が対象。

図表3 本稿における LLM を用いたテキスト解析プロセス



図表4 LLM によるカテゴリ分類

カテゴリ	説明	LLM が出力する判断理由の典型例（「不動産・住宅ローン」をテーマに指定した場合）
テーマに関係あり	テーマに直接関係がある文章、もしくはテーマに関係する単語を含む文章	・理由は、アパートローンは不動産融資の一種であり、投資用や自己住居以外の目的でアパートやマンションを購入や建築する際に利用できるローンであるため
テーマに関係なし	テーマに対して間接的にしか関係がない、もしくは一切関係がない文章	・理由は、文章中には「不動産融資」「住宅ローン」に関する直接的な言及がなく、XXについて述べられているため。

BOX 2: TF-IDF について

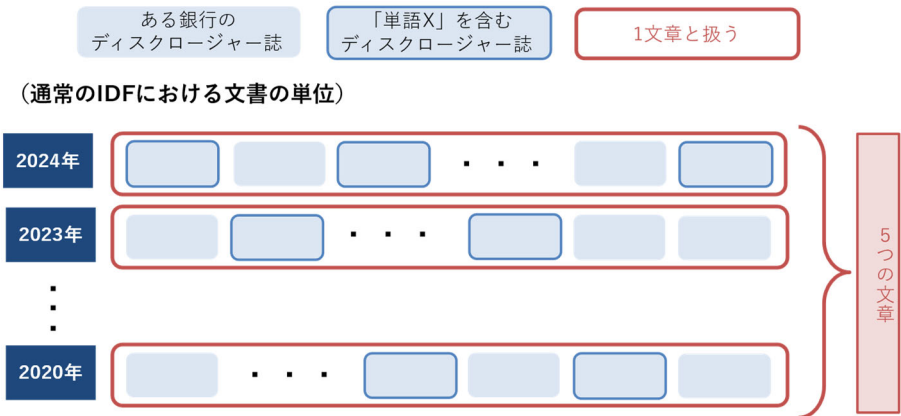
TF-IDF (Term Frequency-Inverse Document Frequency) とは、文章中においてどの単語が重要かを評価するための指標の一つであり、単語の出現頻度 (TF) と単語の希少度 (IDF) を掛け合わせて算出する。単語の出現頻度は、特定の単語の出現回数を全単語数で割って計算したものである。単語の希少度は、特定の単語が比較対象の他の文章には出現しない際に高値となる指標である。TF-IDF の値が高い単語は、その文章を特徴づける重要な単語であると考えられる (図表 5)。

図表 5 TF - IDF の考え方

TF × IDF	=	単語の出現頻度 × 単語の希少度
単語の出現頻度		文書Aにおいて「単語X」がどのくらい出現したのか
TF	=	$\frac{\text{文章Aにおける「単語X」の出現回数}}{\text{文章Aにおける「全単語」の出現回数}}$
単語の希少度		他の文章では出現頻度が少なくその文章固有の単語ほど高い値を示す
IDF	=	$\ln\left(\frac{\text{全文章数}}{\text{「単語x」を含む文章の数}}\right) + 1$

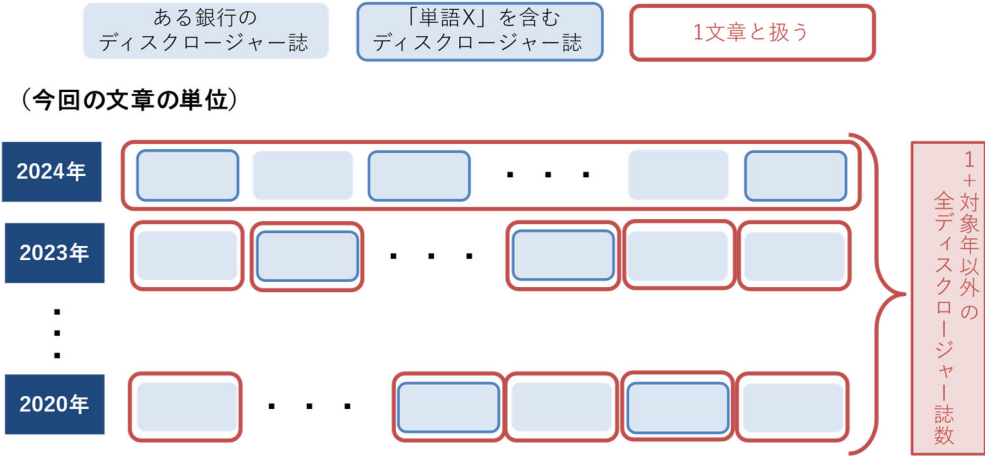
本稿の TF-IDF の算出に当たっては、銀行・グループ業態別の重要な単語を年毎 (2020~2024 年) に比較するため、文章の単位を解釈する際に図表 6 に示すような工夫をしている。これはディスクロージャー誌の内容を各年単位でまとめてしまうと、IDF の値が多くの単語において常に最小値となり、IDF の影響を評価に取り込めないためである。また、TF-IDF の算出に当たって、名詞以外の単語等を除外している。

図表 6 本稿における IDF 計算の工夫



各年のいずれかのディスクロージャー誌に「単語X」を含む場合、 $IDF = \ln\left(\frac{5}{5}\right) + 1 = 1$ の最小値となる。

各単語は、各年のいずれかの ディスクロージャー誌に含まれる可能性が高く、
IDF値が変化しづらいため、IDFの影響が取り込めない。



2024年の「単語X」にかかるIDFは、 $\ln\left(\frac{1 + 2024年以外のディスクロージャー誌数}{\text{「単語x」を含む文章の数}}\right) + 1 \geq 1$ で算出。

単語毎にIDFの値が変動することで、IDFの影響を取り込むことができる。

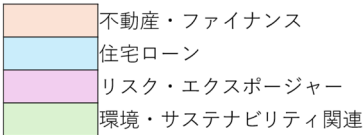
2. 「不動産・住宅ローン」をテーマとした解析結果

本章では、主要行等、地域銀行及びネット銀行のディスクロージャー誌に対して、業態横断で取組があり、ディスクロージャー誌にも一定の記載がみられる「不動産・住宅ローン」をテーマとして解析を行った例を紹介する。

図表 7 は、「不動産・住宅ローン」をテーマとして収集した各銀行の記述に対して TF-IDF を用いて解析した結果を業態別に集約したものである。主要行等では「不動産」の順位が安定して高いほか、ストラクチャードファイナンスや不動産ファイナンス等に関係する文章が一定数存在した結果、「ファイナンス」等の単語の順位を押し上げていた。また 2021 年を除いて「法人」が上位 10 位以内に入っている。地域銀行においては、「住宅ローン」が主要行等より上位に目立つほか、直近 3 年間では「リスク」が 1 位となっている。ネット銀行では、2023 年まで「住宅ローン」が 1 位であり、他にも「サービス」「開始」といった単語の順位が高いが、2024 年には「エクスポージャー」や「信用リスク」がランクインしている。

また、2024 年の主要行等、地域銀行の上位 10 位以内の単語に「排出」、「環境」、「炭素」が見られるなど、近年は環境に関連する単語もランクインしている。これは環境配慮型住宅ローン等に関する記述が増えていることが要因の一つであり、不動産・住宅ローンに関する文脈においても、サステナビリティに対する意識が高まっている可能性が考えられる。

図表 7 不動産・住宅ローンに関する記述における業態別の TF-IDF 推移



(主要行等)

	2020	2021	2022	2023	2024
1	不動産	不動産	不動産	不動産	不動産
2	信託	ファイナンス	ファイナンス	セクター	リスク
3	ファイナンス	資産	リスク	ファイナンス	セクター
4	資産	信託	法人	価値	ファイナンス
5	関連	リスク	ESG	排出	ビジネス
6	リスク	管理	環境	創造	価値
7	提供	法人	取り組み	リスク	環境
8	ビジネス	住宅ローン	収益	ビジネス	創造
9	管理	ビジネス	個人	投資	法人
10	法人	個人	ビジネス	ゼロ	排出

(地域銀行)

	2020	2021	2022	2023	2024
1	エクスポージャー	住宅ローン	リスク	リスク	リスク
2	住宅ローン	開始	住宅ローン	グループ	実績
3	信託	残高	残高	シナリオ	増加
4	残高	支援	グループ	増加	グループ
5	グループ	グループ	サービス	住宅ローン	排出
6	支援	エクスポージャー	支援	残高	炭素
7	資金	資金	開始	不動産	セクター
8	証券化	営業	エクスポージャー	支援	価値
9	融資	サービス	不動産	代理	支援
10	不動産	リスク	増加	サービス	残高

(ネット銀行)

	2020	2021	2022	2023	2024
1	住宅ローン	住宅ローン	住宅ローン	住宅ローン	エクスポージャー
2	開始	開始	開始	開始	住宅ローン
3	サービス	サービス	サービス	サービス	手法
4	残高	突破	突破	残高	不動産
5	保障	格付け	口座	口座	信用リスク
6	特約	連結	残高	突破	適格
7	金利	残高	連結	単体	リテール向け
8	突破	金利	提供開始	提供開始	アセット
9	利用	口座	金利	預金	延滞
10	返済	センター	格付け	金利	適用

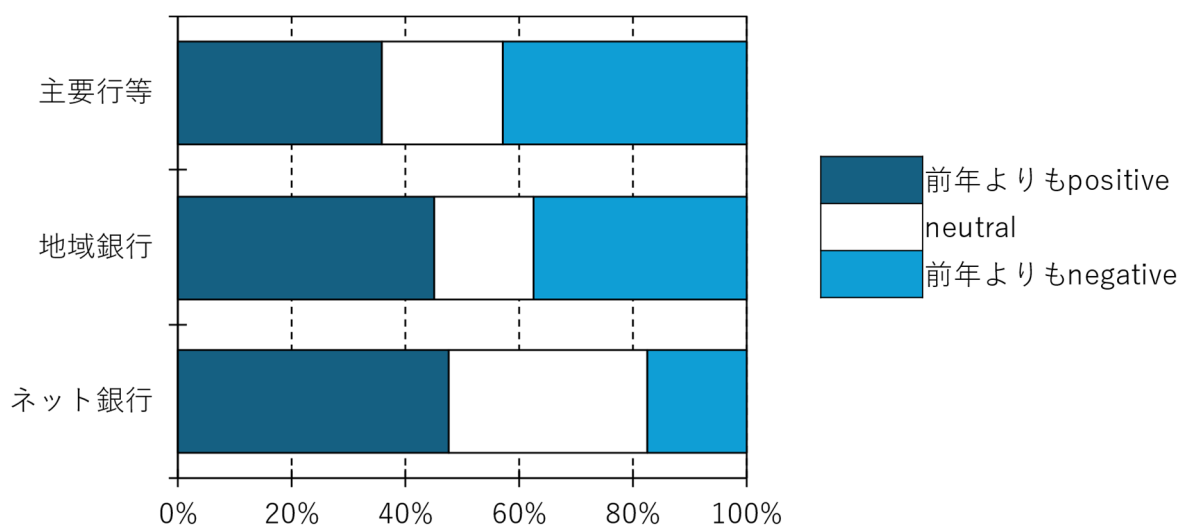
続いて、「不動産・住宅ローン」をテーマとして収集した各銀行の記述に対して、LLM を用いて 2020～2024 年の記述内容を比較し、変化が見られるかを業態別に集計した。今回は、観測された変化の内容が前年よりも positive か negative か（いずれでもない場合は neutral）を LLM に判断させたが、ここでいう positive/negative とは、あくまでも LLM にて判断されたディスクロージャー誌における記述の傾向であり、必ずしも各行の貸出姿勢等と一致するものではない点に留意が必要である。実際に、LLM による positive/negative の判断理由を確認すると、例えば貸出金残高が増加したといった記述があった場合には positive と判断されている例や、記述量の変化で判断されているケースなど、総合的に判断されている様子が窺える（図表 8）。

図表 8 LLM による判断理由の例

LLM による判断結果	LLM による判断理由の例
前年よりも positive	理由は、今年の内容が去年の内容に比べて、住宅ローンの貸出金残高が増加していることから判断した。具体的には、今年の住宅ローンの貸出金残高が XX 円を突破し、XX の取り組みで、より多くのお客さまに住宅ローンをご利用いただきやすくなったことが記載されている。
neutral	理由は、今年も去年も XX などの基本的な情報に加え、XX などが記載されており、特に大きな差異は見られないため。
前年よりも negative	理由は、今年の記述には不動産や住宅ローンに関する記述が少ないことから判断した。去年は XX を行うと記載されているが、今年はそのような記述がない。また、XX と記載されており、不動産や住宅ローンを含む融資関連の記述は見当たらなかった。

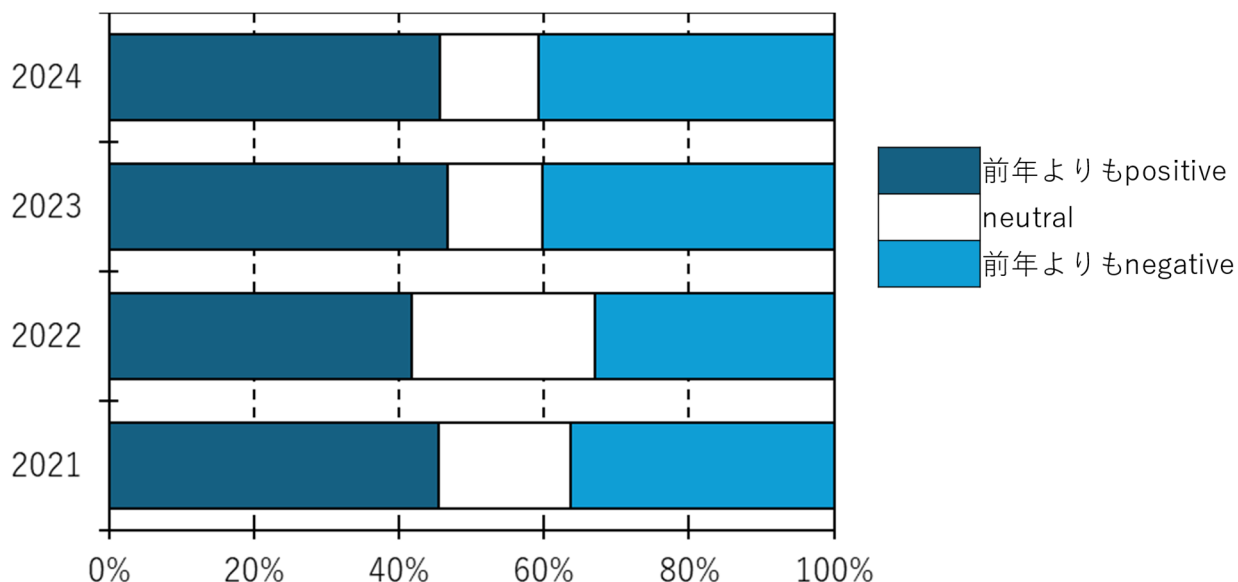
解析した結果、「前年よりも positive」と判断された割合が最も多い業態はネット銀行であり、その値は約 48%である。また、ネット銀行は他業態と比較し「neutral」と判断された割合も高いが、これは指定テーマに関する記載がほぼなく、記述を収集できないディスクロージャー誌が一定数存在したことが原因と考えられる。次に高かったのは地域銀行であり、約 45%の記述において「前年よりも positive」と判断された。主要行等は「前年よりも positive」と判断された割合が最も低く約 35%であったが、「前年よりも negative」と判断された記述の割合と大きな差は見られない(図表 9)。判断理由を確認すると、ネット銀行では「新規サービスに関する言及」や「貸出金残高等の増加」が主な理由を占め、地域銀行では、約 50%の回答において、「貸出金残高や契約者数等の増加」が理由の 1 つに挙げられていた。また、主要行等では多様な理由が見られたが、「ポートフォリオの変化」や「経営的な変化」等他の業態ではほぼ挙げられない理由も多く見られた。

図表 9 業態別 不動産・住宅ローンに関する記述内容の前年比較（5年平均）



加えて、LLM の判断と実データの整合性を検証するために、3 業態の中で最もデータ数が多い地域銀行において、不動産業向け貸出及び住宅ローン等の期末残高の動向（前年比伸び率の各行単純平均）と LLM による判断の対応関係を確認した（図表 10、11）。その結果、「前年よりも positive」と判断された地域銀行の方が、「前年よりも negative」とされた地域銀行よりも前年比伸び率が大きく、LLM による判断と実データは整合的であることが確認された。

図表 10 地域銀行のディスクロージャー誌における不動産・住宅ローンに関する記述内容の前年比較



図表 11 LLM の判断別 不動産及び住宅ローン等の期末残高の前年比伸び率（銀行別単純平均）

	2022年	2023年	2024年
前年よりもpositive	4.42	4.17	4.40
前年よりもnegative	2.96	3.02	2.64

III. 総括

本稿では、ディスクロージャー誌の内容に対して LLM 等のテキスト解析手法を用いて、大量の文章から指定テーマに関する記述を抽出し、特徴的な単語の抽出や、記述内容の時系列変化の把握を試行した。

「不動産・住宅ローン」をテーマにした場合、TF-IDF による解析結果から、主要行等は「不動産」、「ファイナンス」や「リスク」、地域銀行は「住宅ローン」や「リスク」、ネット銀行は「住宅ローン」に関する記載が中心であることが窺えるが、直近は「サステナビリティ」やネット銀行でも「リスク」を重要視していることが確認された。

また、LLM により記述内容の前年変化を確認したところ、ネット銀行、地域銀行、主要行等の順に「前年よりも positive」と判断されたケースが多かった。貸出金残高や契約者数の増加、顧客に支持されている取組等に関する記述があること等を「前年よりも positive」と判断した理由に挙げるケースが見られ、貸出金残高データと突合した結果、LLM による判断と整合的であることを確認した。

これらの解析事例を通じて、ディスクロージャー誌等のテキストデータに対する AI 技術を用いた解析は、指定テーマに対する理解深化に寄与するなど有効であるとの示唆を得た。本稿の解析手法を応用することで、従来は多くの時間・人的コストをかけていた大量文章からの情報収集や傾向把握が容易になり、モニタリング業務の大幅な効率化につながると期待される。一方で、LLM 等の AI 技術を用いる際には、使用している AI の性能の限界、入力指示・質問（プロンプト）の最適化の限界及び AI の学習時のバイアス等の理由で、誤った情報が出力されるハルシネーションといった特有のリスクに留意する必要がある。そのため、本稿における解析では、LLM に結論だけでなく判断理由を出力させ、人間が個別の確認・修正を行うことや整合性の検証を行うことで、リスク抑制を試みている。

引き続き、AI 特有のリスクにも留意した AI 活用の高度化、解析対象とするテキストデータの範囲の拡大等に取り組み、金融行政の進化・深化に努めていく。