JPX
JAPAN EXCHANGE GROUP

日本取引所グループ
JAPAN EXCHANGE GROUP

JPX WORKING PAPER

JPXワーキング・ペーパー

日本の上場会社における監査上の主要な検討事項の自然言語処理を用いた分析
―監査領域の自動分類・意味的類似性・記載内容管理の実証―

土井 惟成，信田 裕介，水野 豪

2025 年 11 月 27 日

Vol. 50

備考

JPX ワーキング・ペーパーは、株式会社日本取引所グループ及びその子会社・関連会社 (以下「日本取引所グループ等」という。) の役職員及び外部研究者による調査・研究の成果を取りまとめたものであり、学会、研究機関、市場関係者他、関連する方々から幅広くコメントを頂戴することを意図しております。なお、掲載されているペーパーの内容や意見は執筆者個人に属し、日本取引所グループ等の公式見解を示すものではありません。

日本の上場会社における監査上の主要な検討事項の自然言語処理を用いた分析
—監査領域の自動分類・意味的類似性・記載内容管理の実証— *

土井 惟成 †　　信田 裕介 ‡　　水野 豪 §

2025/11/27

概要

監査上の主要な検討事項 (Key Audit Matters: KAM) は、監査人が財務諸表等の監査において、職業専門家として特に重要だと判断した事項であり、監査報告書を通じて報告される。本稿では、KAM に対する自然言語処理を用いた分析を通じて、(1) 大規模言語モデル (Large Language Models: LLM) による監査領域のゼロショット分類、(2) KAM のテキストの意味的な類似性の評価手法、(3) 記載内容の管理の程度に関する分析、の 3 つの観点から実験を行った。(1) では、LLM を用いたゼロショットテキスト分類により、最大で 92.8% の精度を達成した。(2) では、KAM の意味的類似性を測定する複数の評価手法を比較した結果、単語の一致率に基づく指標がボイラープレート化の程度を測定するのに有効であることが示され、文脈埋め込みベクトルに基づく評価指標が KAM の意味的類似性を捉える上で有効であることが示された。(3) では、語彙多様性の評価指標と、監査人が所属する監査法人を推定する著者推定の精度を用いて、監査法人の規模別の傾向を分析した。この結果、大手監査法人は語彙が豊富でありながら、著者推定の精度が高い傾向があり、規模の小さい監査法人ほど、語彙が少ないにもかかわらず、著者推定の精度は低い傾向があった。この結果は、大手監査法人ほど組織的な品質管理が行われ、KAM の内容が管理されている可能性を示唆している。

本稿では、これらの結果を踏まえ、KAM における自然言語処理を用いた分析の有用性を示す。

* 本稿の作成に当たっては、日本取引所グループ等のスタッフから有益なコメントを頂いた。ここに深く感謝申し上げる。なお、本稿は、著者らの先行研究 [1, 2, 3] を拡張したものである。

† 株式会社日本取引所グループ 総合企画部 主任研究員， 東京大学大学院工学系研究科 博士後期課程 (n-doi [at] jpx.co.jp)

‡ 株式会社東京証券取引所 上場部 調査役

§ 株式会社 JPX 総研 クライアントサービス部 調査役

目次

1　はじめに

1.1　本稿の背景

証券市場は、様々な投資家がそれぞれの投資判断に基づいて株式などの売買を行う場であり、上場会社が開示する情報がその判断の重要な基盤となっている。これらの開示情報の一つに、金融商品取引法に基づく有価証券報告書がある。上場会社は、企業の概況や経理の状況等を記載した有価証券報告書を、事業年度末から 3 か月以内に内閣総理大臣に提出することが義務付けられている。また、有価証券報告書の経理の部に添付される財務諸表等は、監査人の監査を受けなければならない。監査上の主要な検討事項 (Key Audit Matters: KAM) は、監査人が財務諸表等の監査において、職業専門家として特に重要だと判断した事項 [4] であり、監査報告書を通じて報告される。従来の短文式監査報告書では、監査プロセスの透明性が欠けており、監査の質を評価するのが困難であった。そこで、日本では、国際的な動向を踏まえつつ、「監査プロセスの透明性を向上させること」[5] を目的として、2021 年 3 月期決算より、監査報告書への KAM の記載が全上場会社に対して強制適用されている。日本における KAM の位置付けを図 1 に、KAM のサンプルを表 1 にそれぞれ示す。

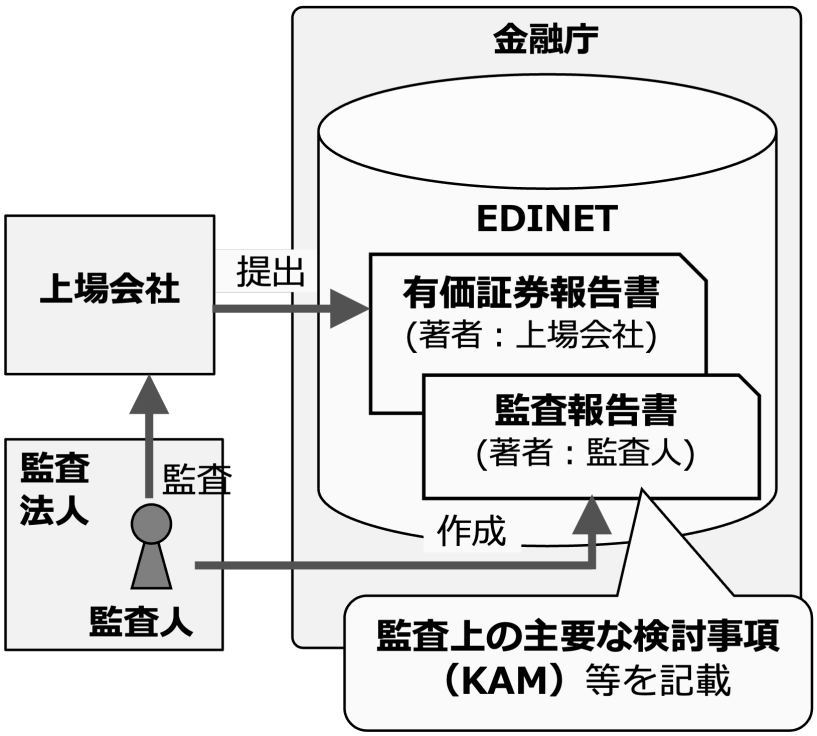

図 1　日本における KAM の位置付け

KAM の導入により、「財務諸表利用者に対して監査のプロセスに関する情報が、監査の品質を評価する新たな検討材料として提供されることで、監査の信頼性向上に資すること」「財務諸表利用者の監査や財務諸表に対する理解が深まるとともに、経営者との対話が促進されること」「監査人と監査役、監査役会、監査等委員会又は監査委員会 (以下「監査役等」という。) の間のコミュニケーションや、監査

表 1　KAM のサンプル（株式会社日本取引所グループ）

見出し	監査上の主要な検討事項の内容及び決定理由	監査上の対応
1 収益認識に関する IT 統制の評価	連結財務諸表注記「20. 営業収益」に記載されているとおり、当連結会計年度の取引関連収益は 64,515 百万円、清算関連収益は 34,445 百万円であり、連結損益計算書における営業収益の 61.0% を占めている。……以上より、当監査法人は当該事項を監査上の主要な検討事項に相当する事項に該当するものと判断した。	当監査法人は、IT 専門家と連携して、IT システム依存度の高い営業収益に関して、取引開始から収益計上に至るまでの IT システムにおける一連のデータフロー、処理プロセス及び自動化された内部統制を理解し、IT システム群の安定稼働のために構築された内部統制の有効性を評価した。主として実施した監査手続は以下のとおりである。……
2 ソフトウエア及びソフトウエア仮勘定の評価	連結財務諸表注記「13. のれん及び無形資産」に記載されているとおり、当連結会計年度末において、ソフトウエアが 32,556 百万円、ソフトウエア仮勘定が 1,751 百万円計上されている。……以上より、当監査法人は当該事項を監査上の主要な検討事項に相当する事項に該当するものと判断した。	当監査法人は、IT 専門家と連携して、ソフトウエア及びソフトウエア仮勘定の評価に係る内部統制の有効性を評価するとともに、開発中の新システムについて、減損の兆候の有無を検討するため、主として以下の監査手続を実施した。……

注）原文の表現を簡略化し一部省略（「……」）した。出典：株式会社日本取引所グループ『有価証券報告書―第 24 期（2024/04/01–2025/03/31）』「当期連結財務諸表に対する監査報告書」。

人と経営者の間の議論を更に充実させることを通じ、コーポレート・ガバナンスの強化や、監査の過程で識別した様々なリスクに関する認識が共有されることによる効果的な監査の実施につながること」[5]といった効果が期待されている。

もっとも、KAM の制度的意義が大きい一方で、日本の KAM には少なくとも三つの課題がある。第一に、監査領域に関するタグ等が付与されておらず、企業横断での比較や年次推移の把握に作業コストが発生する点である。第二に、KAM の内容が定型化する「ボイラープレート化」への懸念が制度当初から指摘されており [6]、同一企業の前年記載を踏まえた改稿がどの程度生じているかを定量的に把握する必要がある点である。第三に、監査法人ごとの記載文体や表現統一の管理の実態が十分に可視化されていない点である。これらの課題は、自然言語処理の手法の活用により、(1) LLM による監査領域のゼロショット分類、(2) KAM のテキストの意味的な類似性の評価手法、(3) KAM の内容の管理の程度に関する分析という 3 つの研究課題と捉えることが可能である。

1.2　本稿の内容

本稿では、日本国内の証券取引所に上場している、日本の上場会社の監査報告書に記載された KAM のテキストを対象に、自然言語処理の手法を用いて、以下の各項に述べる研究課題に取り組む。

1.2.1　KAM の監査領域の自動的な分類

第一の課題は、KAM の監査領域の自動的な分類である。KAM には「見出し」「監査上の主要な検討事項の内容及び決定理由 (以下、内容及び決定理由) 」「監査上の対応」が含まれるが、監査領域を示すタグが存在しないため、上場会社横断の集計や年次推移の把握を行うには大きな作業コストが発生する。

そこで本研究では、教師データや追加学習を要しない LLM のゼロショット分類を用いた、KAM の監査領域の自動的な分類手法を提案する。ゼロショットテキスト分類とは、事前にラベル付きデータを用いた学習を行わずに、新たなテキストをいずれかのラベルへ分類する手法である。これに当たって、KAM テキストのクラスタリングと人手レビューを通じた監査領域の定義を行った後、監査領域の定義を記載したプロンプトを入力することで KAM のゼロショットテキスト分類を実現し、この精度を評価した。

なお、先行研究 [1] でも同様の手法が提案され、その分類精度が検証されていたものの、分類精度を向上させる方法としてより細かい監査領域を定義することが示唆されているほか、使用した KAM の件数が 100 件と限定的であったことが限界として述べられている。

このことを踏まえ、本研究では、2022 年 4 月期から 2023 年 3 月期の KAM に対して埋め込み表現を取得した後、主成分分析による次元圧縮を行い、k-means++ によるクラスタリングと人手による修正を踏まえ、13 種類の監査領域の定義を得た。そして、2022 年 4 月期から 2023 年 3 月期の KAM のデータセット 4,202 件の内、全体の 5% 超である 250 件の KAM に対して、人手で監査領域を分類することで評価用データセットを作成した。その後、この評価用データセットを用いることで、提案手法の精度を評価した。

本研究の概要を図 2 に示す。

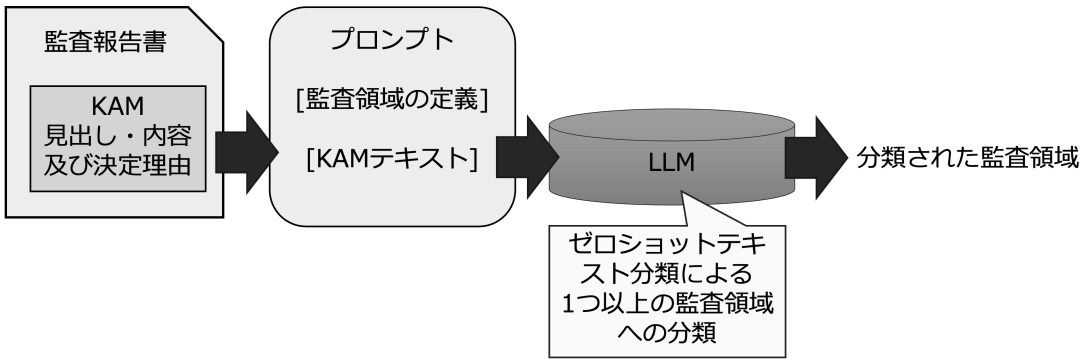

図 2　LLM を用いたゼロショットテキスト分類による KAM の監査領域の分類に係る研究の概要

1.2.2　KAM のテキストの意味的な類似性の評価手法の確立

第二の課題は、KAM のテキスト間の意味的な類似性の評価手法を確立し、同一企業における前年度からの KAM の変更の程度の把握や、他社の KAM との類似性の評価を可能とすることである。

前述のとおり、KAM の制度導入当初より、KAM の制度上の懸念として、KAM の内容が定型化又は画一化してしまう現象 (ボイラープレート化) が指摘されている。KAM のボイラープレート化が進むと、利用者の関心や監査機能の低下を招き、結果として KAM の形骸化につながる恐れがある [7]。KAM の内容がどういう状態であれば、あるいは、KAM にどういう類似性が生じていれば、その KAM はボイラープレート化していると言えるかといった点について、現時点では統一的な見解は無い。しかしながら、日本公認会計士協会の「監査上の主要な検討事項 (KAM) の適用 3 年目に関する周知文書」[8] を考慮すると、同一の上場会社の 2 期分の KAM の「内容及び決定理由」において、「対象論点が同じ」であり、「会社の状況やリスクに関する前年度からの変化が無い」場合においては、ボイラープレート化が生じていると解釈できる。

KAM の類似性の評価指標として、文の類似性の評価指標の一つである、単語の出現頻度のコサイン類似度に基づく手法が使用されている [9, 10, 11]。しかしながら、当該手法には、KAM の意味的な類似性を捉えることは難しいという欠点がある。昨今では、意味的な類似性の評価手法として脈埋め込みベクトルに基づく類似性が提案されているほか、自然言語処理では単語の出現頻度に基づく指標に加えて単語の一致率に基づく類似性の評価指標が活用されている。

そこで、本研究では、既存の類似性の自動評価指標を使用することで、KAM の意味的類似性の自動的な評価手法を検討する。まず、同一の上場会社の 2 期分の KAM のデータセットを作成し、後述する 0 から 5 までの 6 段階のスコアにより、類似性を人手で評価した。その後、様々な自動評価指標により 2 期分の KAM の類似性を評価し、評価用データセットの内容と比較した。本研究では、これらの手法を通じて、KAM の意味的な類似性の評価手法を提案する。

なお、先行研究 [2] では、KAM のテキストに前処理を行わずに、自然言語処理技術を利用した KAM の意味的類似性の自動評価に関する検討が行われた。これに対して、本研究では、KAM における数値表現に対してあらかじめマスク化の処理を施すことで、テキスト中に含まれる数値表現が意味的な類似度の評価に与える影響を低減し、文書全体の記述内容に起因する実質的な類似性の把握を可能とすることを狙った。本研究を通じて、KAM のテキストに適した意味的類似性の評価手法を提案する。

本研究の概要を図 3 に示す。

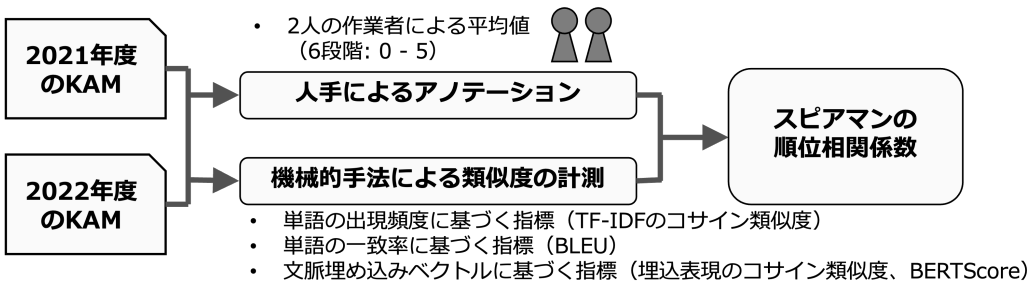

図 3　KAM のテキストの意味的な類似性の評価手法に係る研究の概要

1.2.3　監査法人による KAM の内容の管理の程度に関する分析

第三の課題は、監査法人による KAM の記載内容の管理の程度を、語彙多様性と著者推定の精度から推し量ることである。

日本の大手監査法人の監査品質に関する報告書等によると、KAM のボイラープレート化を避けるよう、各監査現場を支援する組織的な取り組みを行っていることが伺える。たとえば、PwC Japan 有限責任監査法人では、品質管理本部内に KAM 担当チームを設置し、各監査チームの相談対応や文面レビューを積極的に行っている [12]。有限責任監査法人トーマツにおいては、「過年度の有価証券報告書の開示データから過去の全上場企業の KAM のデータを作成したうえで自然言語処理技術を用いて KAM の記載内容や文字数が前年度からどの程度変化しているか等の分析を実施」するといった様々な取組みを行っている [13]。有限責任あずさ監査法人においても、外部のステークホルダーとの定期的な対話を通じたフィードバックの獲得や、重層的なレビュー体制の整備により、KAM の固定化を防ぎ、情報価値を高める工夫を行っている [14]。

これらの取り組みはいずれも、KAM の品質を高めるだけでなく、組織的な取り組みとして、文面統制や専門家としての判断を適切に反映するための仕組みを整えている点が特徴である。こうした仕組みの存在は、各監査法人の KAM において、記載内容が各社に応じたものとなり、文章表現のバリエーションが保持されつつも、一定の統一性や一貫性をもたらすと考えられる。しかしながら、このような監査法人による KAM の管理による影響に関する計量的な測定は、筆者らが知る限り先行研究 [3] を除いて行われていない。

そこで、本研究では、KAM の文章表現に着目し、監査法人単位での統一的な品質管理がどの程度行われているかを測定するために、語彙多様性と著者推定に基づく分析を行った。KAM における語彙多様性は、KAM の表現がどの程度豊富であるかを示す。本研究では、大手監査法人、準大手監査法人、中小規模監査事務所の間で、KAM 文書における語彙多様性がどのように異なるのかを比較し、大手監査法人ほど表現にバリエーションがあるかどうかを明らかにする。また、著者推定の精度が高いということは、書き手の特徴が類似している、つまり内部での文章統制が強く機能している可能性を示す。本研究では、著者推定を通じて、大手の監査法人ほど、KAM の文書表現が監査法人ごとに特徴的なものになっているかどうかを検証する。また、先行研究が 2021 年から 2023 年までが分析対象の期間であったのに対して、本研究では 2021 年から 2025 年までに拡張した。

本研究の概要を図 4 に示す。

1.3　本稿の貢献

本稿で実施した研究の概要を、表 2 に示す。

本稿の貢献は、(1) 手作業を要さない KAM の監査領域の自動分類手法の実運用可能性、(2) 数値マスキングを踏まえた KAM の類似性の評価手法の提案、(3) 語彙多様性と著者推定の精度に基づく KAM の記載内容管理の可視化の提示、の三点である。

一点目の貢献として、250 件の KAM に対して、LLM によるゼロショットテキスト分類による監査領域の推定により、最大で 92.8% の精度を達成し、実運用可能性を示した。二点目の貢献として、意味

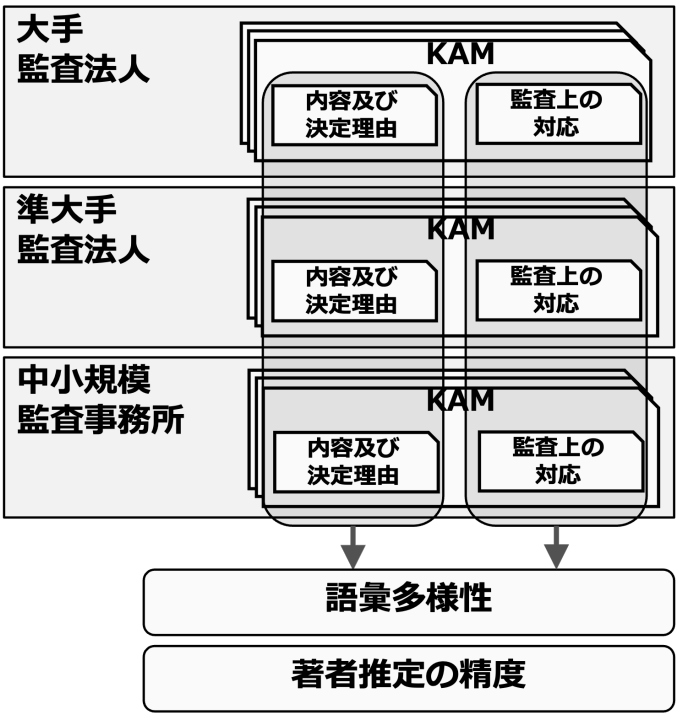

図 4　監査法人における監査上の主要な検討事項の記載内容の管理に関する分析

的類似性の評価において数値表現のマスキングの有用性を確認した。また、単語の一致率に基づく指標がボイラープレート化の程度を測定するのに有効であることが示され、文脈埋め込みベクトルに基づく評価指標が KAM の意味的類似性を捉える上で有効であることが示された。三点目の貢献として、語彙多様性の評価指標と、監査人が所属する監査法人を推定する著者推定の精度を用いることで、大手監査法人ほど組織的な品質管理が行われ、KAM の内容が管理されている可能性が示唆された。

本稿では、これらの貢献を通じて、KAM の分析における自然言語処理の手法の有用性を示す。

1.4　本稿の構成

本稿の構成は次のとおりである。第 2 章では KAM の監査領域の LLM によるゼロショット分類を扱い、第 3 章では KAM テキスト間の意味的類似性の測定を、第 4 章では KAM の記載内容の管理に係る分析を報告し、第 5 章で総括を述べる。

表 2　本稿で実施した研究の概要

テーマ	内容	結果
1. LLM による監査領域のゼロショットテキスト分類	KAM の「見出し」「内容及び決定理由」と 13 種類の監査領域定義を LLM に提示し、ゼロショットで監査領域を分類。 2022 年 4 月期から 2023 年 3 月期における 250 件の KAM で精度を検証。	完全一致 Accuracy の最大は 92.8% (GPT-5) 、Micro-F1 は 95.2%。GPT-4o は 92.0%、Micro-F1 は 94.8%。 複数監査領域の KAM は GPT-5 が高精度 (Accuracy 90.9%) 、単一監査領域の KAM は GPT-4o がやや優位 (95.4%) 。 残差カテゴリ「その他」は両モデルとも難しい。
2. KAM のテキストの意味的な類似性の評価手法	TOPIX Core 30 の同一企業の 2 期分の 50 ペアの KAM に対し、6 段階 (0–5) の人手アノテーションを作成。 金額・時期・割合等の数値表現を [MASK] に置換する前処理を導入。 複数の類似度の測定手法と人手のアノテーション結果におけるスピアマンの順位相関係数を分析。	数値マスキングにより全指標で人手評価との順位相関係数が一様に改善。 ボイラープレート化の程度の測定には BLEU 等の単語一致系が有効、より精緻な意味的類似性の評価には BERTScore 等の埋め込み表現に基づく手法が有効。
3. 記載内容の管理の程度に関する分析	語彙多様性の評価指標と、監査人が所属する監査法人を推定する著者推定の精度を用いて、監査法人の規模別に、KAM の記載内容の管理の程度を分析。	大手監査法人は語彙が豊富でありながら、著者推定の精度が高い傾向があり、規模の小さい監査法人ほど、語彙が少ないにもかかわらず、著者推定の精度は低い傾向。 大手監査法人ほど組織的な品質管理が行われ、KAM の内容が管理されている可能性を示唆。

注) 数値・設計は添付原稿 (概要、第 2–4 章) に基づく。モデル名や指標名は本文表記に準拠し、一部を簡略化した。

2 LLM を用いたゼロショットテキスト分類による KAM の監査領域の分類

本章では、LLM を用いたゼロショットテキスト分類による KAM の監査領域の分類に関する評価について述べる。

本章の構成は次のとおりである。第 2.1 節で本研究の関連研究について述べ、第 2.2 節で使用するデータセットについて述べる。第 2.3 節では、KAM のクラスタリングを通じた KAM の監査領域の候補一覧の作成手法について述べる。第 2.4 節では、KAM の監査領域のゼロショットテキスト分類手法の詳細について述べる。第 2.5 節では、評価用データセットの作成方法を述べる。第 2.6 節では、評価実験の結果を報告する。最後に、第 2.7 節で本研究の結論を小括する。

2.1 関連研究

日本の KAM について、監査領域をはじめとするトピック別に分類することで、KAM の全体的な傾向を把握しようとする取り組みは複数挙げられる。Doi et al.[1] は、LLM を用いたゼロショットテキスト分類による KAM のトピック分類手法を提案し、ラベル付きデータやモデルの学習を要さず、高い精度での KAM の分類を実現した。しかしながら、この報告では、実験に用いたデータセットの数が 100 件の KAM のみであり、実験結果の汎用性に限界がある。また、監査領域が「その他」の KAM の誤分類が目立ったことから、分類精度の改善にはより詳細な監査領域の候補の一覧が必要であると考えられる。

機械的な手法による KAM の分類に関する研究として、Alias et al.[15] は、マレーシアの上場会社における 2020 年 1 月 1 日から 2020 年 12 月 31 日までの KAM を対象として、t 分布型確率的近傍埋め込み法 (t-distributed Stochastic Neighbor Embedding: t-SNE 法) による次元削減を経て 10 種類の監査領域を定義した。その後、1,831 件の KAM に対して 10 種類のいずれかへのラベル付与を行った。そして、埋め込みベクトルに変換するためのモデルとして広く知られている手法である BERT (Bidirectional Encoder Representations from Transformers) [16] を金融テキストに特化させた FinBERT[17] という事前学習済みモデルをファインチューニングすることによって、このラベル付きデータの分類に対して 94% の分類精度を実現した。このように、教師あり学習により、機械的な手法による KAM の分類は高い精度が期待できるものの、このような機械学習モデルの構築には、学習データの構築に係るコストが大きい。これに加えて、金融文書では、年代に応じて出現する固有名詞や専門用語が変わりうるため、実務で利用するには機械学習モデルのメンテナンスのコストが生じる。そこで、本研究では、汎用的な LLM を活用することで、これらのコストを抑えることを目指す。

日本語の金融テキストにおける LLM を用いたゼロショットテキスト分類の応用として、土井ら [18] は、有価証券報告書のサステナビリティ開示情報を対象に、TCFD 推奨開示項目に関するクライテリアの判定のために、ChatGPT を用いたゼロショットテキスト分類の有用性を報告している。この手法を準用する形で、KAM の監査領域の分類においても、LLM を用いたゼロショットテキスト分類が有用性が期待される。

2.2 データセット

本研究に使用したデータセットは、2023 年 6 月 30 日時点で最新の、2022 年 4 月期から 2023 年 3 月期の間に、金融庁が提供する電子開示システムである EDINET で開示された上場会社の有価証券報告書の中から、3,312 件の「当期連結財務諸表に対する監査報告書」を収集することで作成した。収集した監査報告書のうち、KAM に関連する情報は「見出し」「内容及び決定理由」「監査上の対応」というセクションに分けて記載されており、各項目は XBRL 形式のタグでマークアップが施されている。そこで本研究では、XML パーサーを利用して、各監査報告書から KAM に関するテキストデータを抽出した。また、抽出したテキストに対して、NFKC 正規化を含むテキスト正規化を施した。最終的に、本研究では、4,202 件の KAM のデータセットを作成した。

2.3 監査領域の定義

LLM を用いたゼロショットテキスト分類により KAM の監査領域を分類する場合、まずは分類先の監査領域を定義する必要がある。そこで、本データセットの KAM のテキストの埋め込みベクトルとクラスタリング手法を利用して、KAM の代表的な監査領域を定義した。KAM の監査領域は「内容及び決定理由」に記載されていることから、以下では、内容及び決定理由の KAM のテキストを分析の対象とした。

まず、本データセットにおける全てのテキストを、高次元の埋め込みベクトルに変換した。埋め込みベクトルに変換するためのモデルとして BERT[16] が、日本語の金融テキストに特化した BERT モデルとして izumi-lab/bert-small-japanese-fin[19] が、それぞれ知られている。ただし、多くの BERT モデルは入力可能な文字列のトークンの上限数が 512 であるのに対して、多くの KAM のテキストはそれ以上の数のトークンで構成されている。そこで、本研究では、より多くのトークン数のテキストから埋め込みベクトルが生成できるモデルとして、OpenAI 社の text-embedding-3-large[*1] を利用した。text-embedding-3-large モデルは、トークン数の上限が 8,192 であり、入力テキストから 3,072 次元のベクトルを生成することができる。このような高次元なベクトルは、単語の出現頻度に基づくベクトルをはじめとする他の手法と比較して、テキスト間の微妙な意味の違いを捉える能力が高いと考えられる。

次に、3,072 次元のベクトルに変換したテキストに対して、主成分分析を利用して次元を削減した。高次元のデータには多くの場合、冗長な情報やノイズが含まれているため、有効な情報のみを抽出する目的で次元削減が行われる。本研究では、累積寄与率が 80% に達する次元数として、200 次元に削減した。この次元削減により、KAM のテキストが持つ本質的な意味を維持しながら、監査領域ごとの分類が可能となることが期待される。

そして、次元削減後のベクトルを使用して、クラスタリングアルゴリズムである k-means++ を適用した。k-means++ は、k-means アルゴリズムの初期値選定を改良した方法であり、より均等にデータ点をカバーする初期クラスタ中心を選択することが可能である。各テキストは意味的な表現を保つベクトルで表現されていることから、このクラスタリングにより、異なる監査領域を表すテキスト同士がそれ

[*1] https://platform.openai.com/docs/models/text-embedding-3-large

ぞれのクラスタへ効果的に分類されることが期待される。

k-means++ では、クラスタ数を予め定める必要がある。本研究では、Doi et al.[1] を踏まえ、KAM の監査領域として定義すべきクラスタの数は 10 よりも多くする必要があると考えた。図 5 のとおり、クラスタ内の凝集度とクラスタ間の分離度を測定する指標である、シルエットスコア [20] の傾向を見ると、クラスタ数を 31 以上とするとシルエットスコアが悪化していることから、本研究では、k-means++ ではクラスタ数を 30 として入力し、監査領域の検討の漏れを防ぐことを目指した。

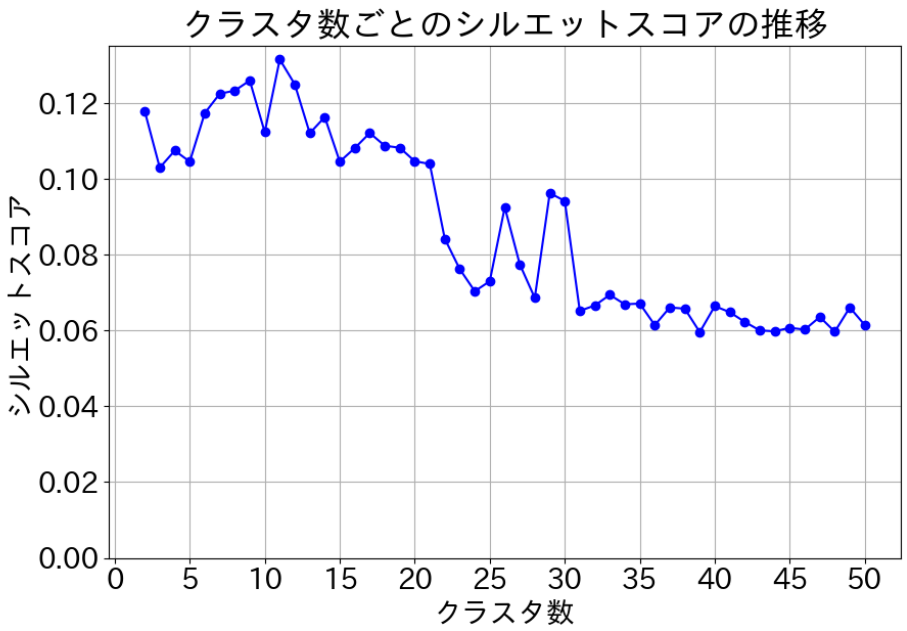

図 5　クラスタ数別のシルエットスコアの傾向

その後、k-means++ で得られた 30 個のクラスタに対して、人手により、各クラスタの代表的な監査領域を識別した。この過程において、複数のクラスタにおける代表的な監査領域の統廃合を行った。この例として、「有形固定資産の評価」と「無形固定資産の評価」が挙げられる。これらの監査領域は別のクラスタとして分類されていたものの、KAM によっては有形固定資産と無形固定資産の記載を明示的に分けている事例とそうではない事例が混在していたことから、これらの監査領域は「固定資産の評価」に統合した。また、他のものとは特筆される監査領域であっても、十分な件数が見込まれない監査領域については「その他」に集約することとした。このような、代表的な監査領域の統廃合を通じて、最終的に「その他」を含む 13 種類の監査領域を定義した。これらの監査領域の定義の詳細を表 3 に示す。

表 3　監査領域の定義

#	監査領域名	定義
1	固定資産の評価	のれんを除く有形固定資産・無形固定資産の評価に係る論点
2	のれんの評価	のれんの評価に係る論点
3	収益認識	売上高を含む収益認識に係る論点 (実在性・正確性に係る論点、期間帰属に係る論点、企業会計基準第 29 号「収益認識に関する会計基準」の適用開始に伴う論点、工事進行基準やソフトウェア開発等の一定期間にわたり履行義務が充足される契約に起因する収益認識に係る論点、等)
4	繰延税金資産の評価	回収可能性や妥当性を始めとする、繰延税金資産の評価に係る論点
5	棚卸資産の評価	在庫として保有している商品、製品、原材料、仕掛品等の、棚卸資産の評価に係る論点
6	債権の評価	売掛金等の営業債権の評価に係る論点及び、営業債権に対する貸倒引当金の見積りに係る論点
7	債務の見積り	引当金をはじめとする債務の見積りに係る論点 (ただし、貸倒引当金を除く)
8	組織再編	自社に関する企業結合や分社化を始めとする組織再編に係る論点及び、前記に実施された組織再編における配分手続き等を当期に完了したものに係る論点
9	継続企業の前提	継続企業の前提に重大な疑義を生じさせるような状況が存在しているが、現時点では継続企業の前提に関する重要な不確実性が認められない場合の継続企業の前提に係る論点
10	IT システムの評価	IT システムの信頼性や新 IT システムへの移行・稼働をはじめとする、IT システムの評価に係る論点
11	投融資の評価	非上場会社を始めとする投資有価証券の評価に係る論点
12	不正・不適切な会計処理	不正な財務報告、不適切な取引、内部統制の不備等に起因する、会計処理上の問題に係る論点
13	その他	上記以外の論点

2.4　提案手法

本研究では、LLM を利用した KAM の監査領域のゼロショットテキスト分類の手法を提案する。一般的に、LLM は、文字列が入力されると、それに自然に従うテキストを出力する。そのため、KAM のテキストと KAM の監査領域の定義に加え、その KAM のテキストに適した監査領域を出力するように指示するプロンプトを LLM に入力すると、その指示のとおり、入力した KAM のテキストに適した監

査領域が出力されることが期待される。

そこで、本研究では、LLM に入力するプロンプトを、KAM のテキストと、「その他」の監査領域を除いた、表 3 に示す監査領域の定義を含むように作成した。そして、そのプロンプトを LLM に入力し、出力されたテキストを用いて、KAM を 1 つ以上の監査領域に識別した。また、この時、いずれの監査領域にも該当しない場合は、「その他」に分類するよう、プロンプトに明示した。以下にて、本研究で使用するプロンプトのサンプルを示す。

以下に述べるのは、日本の上場会社の、監査上の主要な検討事項 (KAM) と、その監査領域の候補である。

KAM の見出し

...

KAM の内容及び決定理由

...

監査領域の候補の一覧

固定資産の評価, のれんの評価, 収益認識, 繰延税金資産の評価, 棚卸資産の評価, 債権の評価, 債務の見積り, 組織再編, 継続企業の前提, IT システムの評価, 投融資の評価, 不正・不適切な会計処理

監査領域の候補の定義の一覧

固定資産の評価: のれんを除く有形固定資産・無形固定資産の評価に係る論点

...

タスク

上記の監査領域の候補から、上記の KAM が該当する 1 つ以上の監査領域のみを出力してください。どれにも該当しない場合は、「その他」と出力してください。

該当する監査領域

2.5 評価用データセット

評価用データセットは、本データセットから全体の 5% 超である 250 件をランダムに選出して作成した。この時、5 年以上の証券又は監査関連の業務経験を持つ 3 人の作業者が、それぞれの KAM に対して監査領域の分類を行った。各作業者には、各 KAM の見出しと内容及び決定理由が与えられた。3 人の作業結果が異なる場合は多数決で決定し、それでも決まらない場合は 3 人で協議のうえ決定した。

評価用データセットの監査領域別の統計情報を表 4 に示す。

2.6 評価実験

定義した監査領域と提案手法の有用性を評価するため、評価用のデータセットを作成して評価実験を行った。本章では、評価用データセットの作成手順と、評価実験の環境を述べ、その後、実験結果とその考察を述べる。

表 4　評価用データセットの統計情報 (監査領域別)

監査領域	件数	内容及び決定理由の記述統計 (トークン)						
		平均	標準偏差	最小	第 1 四分位数	中央値	第 3 四分位数	最大
固定資産の評価	64	395.2	117.3	186	306.3	391.0	469.0	733
のれんの評価	23	391.1	100.1	187	321.0	403.0	446.0	600
収益認識	69	366.3	124.4	152	294.0	334.0	423.0	893
繰延税金資産の評価	25	331.2	95.4	226	265.0	301.0	360.0	607
棚卸資産の評価	28	353.6	66.2	201	318.5	345.5	381.5	516
債権の評価	15	562.7	211.7	168	428.0	577.0	723.0	935
債務の見積り	10	348.6	89.4	196	270.5	383.5	414.0	451
組織再編	4	385.5	83.0	329	337.3	352.5	400.8	508
継続企業の前提	5	434.0	102.5	329	369.0	418.0	459.0	595
IT システムの評価	7	340.1	115.2	163	265.0	354.0	434.5	465
投融資の評価	5	448.6	98.7	367	393.0	419.0	447.0	617
不正・不適切な会計処理	3	406.3	105.2	286	369.0	452.0	466.5	481
その他	3	338.7	78.4	250	308.5	367.0	383.0	399
全体	250	383.6	126.1	152	301.0	359.5	447.0	935

注) 件数は当該監査領域に割り当てられた KAM の件数を指し、複数の監査領域に該当する KAM は重複して計上した。

2.6.1　実験設定

評価実験では、提案手法である LLM を用いたゼロショットテキスト分類により、評価用データセットに含まれる 250 件の KAM を、13 種類の監査領域の内の 1 つ以上に分類し、その分類精度を人手による評価結果と比較した。

本研究では、LLM として、OpenAI 社の gpt-4o-2024-05-13[*2] (以下、GPT-4o) 、gpt-5-2025-08-07[*3] (以下、GPT-5) 、gpt-5-mini-2025-08-07[*4] (以下、GPT-5 mini) 及び gpt-5-nano-2025-08-07[*5] (GPT-5 nano) で試験した。なお、GPT-4o への入力においては、temperature (温度) を 0 として再現性を確保したが、GPT-5、GPT-5 mini、GPT-5 nano については temperature の設定が不可能であること[*6]を踏まえ、デフォルトの設定を利用した。

評価指標としては、KAM が複数の監査領域を取り得る点を踏まえ、TP=True Positive、FP=False Positive、FN=False Negative として、次を採用した。また、Accuracy については、250 件全体の Accuracy に併せて、単一の監査領域を持つ KAM(239 件) と複数の監査領域を持つ KAM(11 件) の Accuracy をそれぞれ計測した。

[*2] https://platform.openai.com/docs/models/chatgpt-4o-latest
[*3] https://platform.openai.com/docs/models/gpt-5
[*4] https://platform.openai.com/docs/models/gpt-5-mini
[*5] https://platform.openai.com/docs/models/gpt-5-nano
[*6] “The following parameters are not supported when using GPT-5 models (e.g. gpt-5, gpt-5-mini, gpt-5-nano): temperature ...” [21]

1. 厳密一致率 (Accuracy) ：予測集合と正解集合が完全一致したサンプルの割合。
2. 部分的一致率 (Any-hit) ：各 KAM について、推定された監査領域の集合と正解の監査領域の集合の積集合が空でない割合 (いずれか 1 つの監査領域でも当たっていれば正とする) 。
3. 適合率 (Precision) ：TP/(TP+FP) で算出。
4. 再現率 (Recall) ：TP/(TP+FN) で算出。
5. Micro-F1: Precision と Recall の調和平均で算出。

2.6.2　実験結果

実験結果のうち、全体的な傾向を表 5 に示す。最も高い完全一致率は GPT-5 の 92.8%、Micro-F1 は 95.2% であった。GPT-4o は僅差で続き、軽量モデル (GPT-5 mini と GPT-5 nano) は概ね 1-3 ポイント下回った。Any-hit はいずれのモデルも 96.8% 以上と高く、大半の KAM で主要な監査領域は捕捉できていることが分かる。

表 5　全体的な傾向

モデル	Accuracy (全体)	Accuracy (単一領域)	Accuracy (複数領域)	Any-hit	Precision	Recall	Micro-F1
GPT-4o	92.0	95.4	18.2	97.6	95.4	94.3	94.8
GPT-5	92.8	92.9	90.9	98.0	92.7	97.7	95.2
GPT-5 mini	90.0	87.9	90.9	97.6	87.8	96.6	92.0
GPT-5 nano	88.0	90.8	72.7	96.8	91.0	96.6	93.7

注) 各値は百分率 (%) 。

GPT-5 は GPT-5 mini と GPT-5 nano の結果を上回っていることを踏まえて、以下では、GPT-5 mini と GPT-5 nano に係る詳述は割愛し、GPT-5 と GPT-4o に焦点を当てる。

評価用データセットには、複数の監査領域の KAM が 11 件含まれる。これらに限定した完全一致率は、GPT-4o では 18.2%、GPT-5 では 90.9% であった。このように、GPT-5 は同時検出に長ける一方、GPT-4o は片方のみを出力しがちで Any-hit は満たすが完全一致に至らないケースが多かった。

一方で、単一の監査領域の KAM ($n = 239$) に限ると、GPT-4o の Accuracy は 95.4% で最も高く、GPT-5 は 92.9% であった。すなわち、複数の監査領域の推定については GPT-5 が優れるが、単一の監査領域の推定については GPT-4o が強いという傾向の差異が確認できる。

表 6 に、監査領域別の傾向を示す。表 6 を見ると、サンプル数の多い「収益認識 ($n = 69$) 」「固定資産の評価 ($n = 64$) 」は両モデルとも高水準 (F1 $\geqq$ 0.94) であった。その一方で、「IT システムの評価」ではモデル間の挙動が対照的で、GPT-4o は Precision が 100.0%、Recall が 42.9% (Micro-F1: 60.0%) と保守的に検出する傾向が見られ、GPT-5 は Precision が 70.0%、Recall が 100.0% (Micro-F1: 82.4%) と積極的に検出する傾向が見られた。残差カテゴリである「その他」 ($n = 3$) は両モデルとも難度が高い (GPT-5 の Micro-F1: 40.0%、GPT-4o の Micro-F1: 0.00%) ことが確認された。

表 6　監査領域別の傾向

		GPT-4o			GPT-5		
監査領域	**件数**	**Precision**	**Recall**	**Micro-F1**	**Precision**	**Recall**	**Micro-F1**
固定資産の評価	64	95.2	92.2	93.7	94.0	98.4	96.2
のれんの評価	23	82.1	100.0	90.2	85.2	100.0	92.0
収益認識	69	98.6	100.0	99.3	98.6	100.0	99.3
繰延税金資産の評価	25	100.0	100.0	100.0	100.0	100.0	100.0
棚卸資産の評価	28	100.0	100.0	100.0	100.0	100.0	100.0
債権の評価	15	100.0	93.3	96.6	100.0	80.0	88.9
債務の見積り	10	90.0	90.0	90.0	76.9	100.0	87.0
組織再編	4	75.0	75.0	75.0	80.0	100.0	88.9
継続企業の前提	5	100.0	100.0	100.0	100.0	100.0	100.0
IT システムの評価	7	100.0	42.9	60.0	70.0	100.0	82.4
投融資の評価	5	83.3	100.0	90.9	62.5	100.0	76.9
不正・不適切な会計処理	3	100.0	100.0	100.0	100.0	100.0	100.0
その他	3	0.0	0.0	0.0	50.0	33.3	40.0
全体	250	95.4	94.3	94.8	92.7	97.7	95.2

注) 各値は百分率 (%) 。件数は当該監査領域に割り当てられた KAM の件数を指し、複数の監査領域に該当する KAM は重複して計上した。

2.6.3　考察

本実験の結果から、KAM の監査領域の推定において、LLM を用いたゼロショットテキスト分類が高い精度で実現できることを示している。全体では、GPT-5 の完全一致率は 92.8%、Micro-F1 は 95.2%、GPT-4o の完全一致率は 92.0%、Micro-F1 は 94.8% と、いずれも高水準であり、Any-hit も両モデルで 96.8% 以上と極めて高い。すなわち、大半の KAM について主要な監査領域を推定できており、KAM の全社的な分析において、LLM を用いたゼロショットテキスト分類に基づく監査領域を暫定的な監査領域として与えることで、効率的に全ての KAM の監査領域を決定することが可能となると考えられる。

一方で、実験の結果からモデル間の挙動の差異が確認された。特に、GPT-5 は複数の候補を積極的に併記する挙動を示すのに対し、GPT-4o は単一の候補を割り当てるという保守的な挙動が確認された。GPT-4o は、OpenAI によると、発表当時の 2024 年 5 月 13 日時点では同社の LLM の中で最も高い水準の性能を示している [22]。また、GPT-5 は、より複雑な問題に対して複雑問題用モデル（深い推論）（GPT-5 thinking）を搭載している [23]。このようなモデルの仕様の差異が、実験結果の傾向の差異に影響を与えているものだと考えられる。運用面では、複数の監査領域を含みうる KAM は GPT-5 で広めに拾い、単一の監査領域であることが明確な KAM は GPT-4o で絞り込む、といった役割分担が考えられる。その他に、複数のモデルを用いた多数決のような手法で、複数のモデルのアンサンブルにより、Accuracy の一層の向上が見込まれる可能性がある。

なお、GPT-4o は temperature を 0 とすることで再現性が一定程度確保されている一方、GPT-5 はデフォルト設定での評価であり、出力ばらつきが残存する可能性がある。運用時には、複数回推論の多数

決を併用することで、実運用の安定性を確保できると考えられる。Doi et al.[24] は、KAM における監査領域の分類において、同一のモデルに対して同一のプロンプトを入力し、その多数決の結果を活用することで、分類精度が向上することを報告している。

監査領域別の傾向を踏まえると、モデル特性と監査領域の定義の差分に起因して、性能に差異が生じていることが確認された。サンプル数が多い「収益認識 ($n = 69$)」と「固定資産の評価 ($n = 64$)」は、GPT-4o と GPT-5 の両モデルとも極めて高精度で、ゼロショットでも安定することが分かる。他方、「IT システムの評価 ($n = 7$)」では、GPT-4o よりも GPT-5 の方が Micro-F1 をはじめとする性能が高く、一方で、「投融資の評価 ($n = 5$)」や「組織再編 ($n = 4$)」では、GPT-5 よりも GPT-4o の方が Micro-F1 をはじめとする性能が高い傾向が確認された。このように、監査領域に応じて得意なモデルが異なることが推察される。このような傾向の差異を踏まえると、複数の LLM の出力のアンサンブルにより、監査領域の推定精度の向上が期待できる可能性がある。

また、残差カテゴリの「その他 ($n = 3$)」については、先行研究 [1] と同様に、GPT-4o と GPT-5 の両モデルとも難度が高く、GPT-5 でも Micro-F1 は 40.0% に留まる。このことは、「その他」については、ゼロショットの定義だけでは判定根拠が希薄になりやすいことを示唆する。これを改善する方法としては、より多くの監査領域を定義することが考えられる。

以上を踏まえると、KAM の監査領域の推定において、LLM を用いたゼロショットテキスト分類が高い性能を持つことが確認された。実務的な運用としては、複数の LLM の出力アンサンブルが有効である可能性がある。また、「その他」の分類は先行研究に引き続き難しいことが確認されている。これを改善する方法としては監査領域の定義を増やすことが候補として挙げられる。

なお、本実験に関する限界として、全体の約 5% である 250 件の、単年度の KAM にのみ基づいていることが挙げられる。そのため、今後は、より広範囲の KAM に基づく分析が求められると考えられる。

2.7 小括

本章では、2022 年 4 月期から 2023 年 3 月期に EDINET で開示された有価証券報告書に添付される監査報告書から抽出した、4,202 件の KAM を対象に、LLM を用いたゼロショット分類で監査領域の自動推定について検証した。まず、text-embedding-3-large の埋め込み (3,072 次元) を主成分分析で 200 次元へ圧縮し、k-means++ で 30 クラスタを抽出し、作業者によるレビューにより統合・整理して 13 の監査領域を定義した。その後、見出し、内容及び決定理由、領域定義を組み合わせたプロンプトでゼロショットテキスト分類を行い、3 名の作業者が付与した 250 件の評価データで性能を評価した。

評価実験の結果、Accuracy は GPT-5 が 92.8% と最も良く、GPT-4o も 92.0% と高水準であった。複数の監査領域を持つ KAM では GPT-5 の Accuracy は 90.9% であり、単一の監査領域を持つ KAM では GPT-4o の Accuracy が最も高く、95.4% であった。監査領域別に見ると、GPT-4o と GPT-5 で異なる傾向が確認されたものの、残差カテゴリの「その他」は依然として難度が高く、GPT-5 の Micro-F1 は 40.0%、GPT-4o の Micro-F1 は 0.0% だった。

以上から、ゼロショットテキスト分類は、KAM の監査領域の推定を高精度に実現でき、モデル特性の差異の補完を狙ったアンサンブルが有効である可能性がある。

このように、KAM の監査領域の推定において、自然言語処理の手法が有用であることが示唆された。

3　KAM の意味的な類似性の測定方法

本章では、KAM の意味的な類似性の測定方法を提案し、その評価実験について述べる。

本章の構成は次のとおりである。第 3.1 節では、KAM とテキストの類似性に関する先行研究についてレビューする。第 3.2 節では、本研究で使用するデータセットについてまとめる。第 3.3 節では、人手による類似性のアノテーションについて述べる。第 3.4 節では、本研究で提案する、数値表現のマスキング方法について述べる。第 3.5 節では、本研究での比較対象である、KAM の類似性を評価手法についてまとめる。第 3.6 節では、KAM の類似性の自動評価に関する実験を行う。最後に、第 3.7 節で本研究の結論を小括する。

3.1　関連研究

各国の KAM や米国の Critical Audit Matters (以下、CAM) の内容が、投資家の意思決定や監査の質に及ぼす影響に関する調査では、様々な結果が得られている。Rautiainen[25] は、フィンランドの企業を対象として、KAM は必ずしも監査の質を向上させるものではないが、監査の有効性と監査人と経営者の間の協力関係を強化するものであると報告している。Chan and Liu[26] は、米国の CAM の開示が監査と投資家の精査の両方に影響を与えるだけでなく、投資効率にも影響を与える可能性があることを発見した。Suttipun[27] は、タイの企業を対象として、KAM の内容が監査の品質と正の相関があることが報告している。これらの調査結果は、KAM や CAM の内容が様々な影響を与える可能性があることを示唆している。

一方で、KAM 及び CAM の類似性に関する研究は限定的である。Carlé et al.[28] は、ドイツ企業の KAM の類似性を測定するために、レーベンシュタイン距離 [29] を利用した。KAM 及び CAM の類似性の評価には、レーベンシュタイン距離の他に、単語の出現頻度に基づくコサイン類似度がよく使用されている。Zeng et al.[9] は、中国の上場会社の KAM の類似性を評価するために、Burke et al.[10] は、米国の上場会社の CAM の類似性を評価するために、単語の出現頻度に基づくコサイン類似度を利用している。これらの類似性は、意味的な要素を直接考慮していないため、実際の文脈やニュアンスの違いを捉えきれない可能性がある。

この問題に対処するため、近年の研究では、文書や文の意味的な類似性を評価するために、単語の埋め込みベクトルや文脈を考慮したモデルが利用されている。例えば、BERT[16] のような事前学習された言語モデルは、単語の意味をその文脈の中で捉えることができ、これらによる埋め込み表現を類似度に活用することが期待される。更に、BERTScore[30] のような、言語モデルの出力を利用して文書間の意味的な類似性を評価する指標も提案されている。これらのモデルは、伝統的な単語の出現頻度に基づく手法やレーベンシュタイン距離と異なり、文脈上の意味を考慮に入れるため、より高度な意味的類似性の評価が可能である。

テキストの意味的な類似性の評価に関するタスクとして、Semantic Textual Similarity (STS)[31] が知られている。STS データセットは、異なる 2 つの文のペアと、その類似度を示す 0.0 から 5.0 のスコアで構成されている。日本語における文の類似性評価として、JGLUE[32] のサブタスクである JSTS があ

る。これらのタスクでは、一般的な文を評価対象としており、KAM のような特殊なドメインのテキストの類似性を評価することを目的としていない。

このような先行研究を背景として、Doi et al.[2] は、既存の類似性の自動評価指標を使用することで、KAM の意味的類似性の自動的な評価手法を検討した。まず、同一の上場会社の 2 期分の KAM のデータセットを作成し、後述する 0 から 5 までの 6 段階のスコアにより、類似性を 2 人の作業者が人手で評価した。その後、様々な自動評価指標により 2 期分の KAM の類似性を評価し、評価用データセットの内容と比較することで、KAM の意味的な類似性の評価手法を提案した。この既存研究では、KAM のテキストに対して特別な前処理を行わずに類似度を計測した。これにより、前年度と本年度において金額や日付等の数値表現のみが変わっている場合において、人手評価では 2 人の作業者が 5 (2 つの KAM は、数値や時期を除けば、完全に同等で、同じ意味を持つ。) と評価しているにもかかわらず、これらの差異は類似性の評価指標の計算に影響を与えている。そのため、実質的な意味の類似性を評価するためには、この数値表現に起因する影響を低減することが求められる。

3.2 データセット

KAM の意味的な類似性の評価手法の包括的な分析を行うため、同一の上場会社の 2 期分の KAM で構成されるデータセットを作成する。このために、2021 年 4 月期から 2022 年 3 月期まで (以下、2021 年度) と、2022 年 4 月期から 2023 年 3 月期まで (以下、2022 年度) の有価証券報告書に添付される監査報告書の収集と、監査報告書からの KAM の抽出を行った。監査報告書は、EDINET を通じて収集した。本研究では、代表的な日本企業の KAM に着目するため、2023 年 10 月 31 日時点で、時価総額や流動性の特に高い銘柄から構成される TOPIX Core 30 に選定されている上場会社 30 社を対象とした。

なお、日本の上場会社の監査報告書には、「当期連結財務諸表に対する監査報告書」と「当期財務諸表に対する監査報告書」の 2 種類がある。本研究では「当期連結財務諸表に対する監査報告書」を分析の対象とした。各 KAM は「見出し」「内容及び決定理由」「監査上の対応」で構成されている。本研究では、「内容及び決定理由」に注目し、データセットを作成した。

監査報告書には通常 1 つ以上の KAM が記載されている。2 期分の KAM において複数の KAM が記載されている場合、これらの比較のためにはそれぞれの KAM の紐付けが必要となる。そこで、人手で 1 つずつ KAM を確認し、より関連していると考えられる KAM 同士を紐付けた。2 期間において KAM の個数が変化し、紐付け先が存在しない場合、その KAM は除外した。また、監査報告書における KAM は、固有の XBRL タグによってマークアップされているが、このマークアップの範囲が明らかに誤っている KAM を除外した。

この結果、対象の上場会社 30 社において、2021 年度の KAM は合計 58 件、2022 年度の KAM は合計 54 件あり、上場会社単位の紐付けの結果、KAM のデータセットは 51 件となった。この KAM のデータセットのうち、1 件の KAM において明確なマークアップの誤りが含まれていたため、これを除外した。最終的に、50 件のデータセットを構築した。本データセットの統計情報を表 7 に示す。なお、トークン数の計算には、MeCab [33] と UniDic[34] を使用した。

表 7　意味的類似性の評価指標の分析に使用するデータセットの統計情報 (年度別)

年度	件数	KAM の内容及び決定理由の記述統計 (トークン)						
		平均	標準偏差	最小値	第 1 四分位数	中央値	第 3 四分位数	最大値
2021	50	434.3	171.5	163	309	379	532	849
2022	50	445.7	177.9	163	329	408	504	967

注) 各統計量は、同一企業の 2 期分の KAM(内容及び決定理由) のトークン数に基づく記述統計。

3.3　人手による類似性のアノテーション

前節の手続きによって作成した 50 件のデータセットを対象に、人手で類似性を評価し、アノテーションを付与した。

アノテーションの信頼性を確保するため、表 8 のとおり、Semantic Textual Similarity (STS) [31] に倣う形で、6 段階のクライテリアを設けた。1 から 5 のスコアは、「会社の状況やリスクに関する前年度からの変化」の評価結果を示しており、0 は KAM の対象論点が異なることを示している。

表 8　アノテーションにおけるクライテリア

スコア	クライテリア
(5)	2 つの KAM は、数値や時期を除けば、完全に同等で、同じ意味を持つ。
(4)	2 つの KAM は、数値や時期を除けばほぼ同等だが、重要でない細部が異なる。目安として、意味が異なる段落が 1 つ追加／削除されている。
(3)	2 つの KAM は、数値や時期を除けばほぼ同等だが、重要な情報が異なる／欠けている。これには新会計基準採用による変更も含まれる。目安として、意味の異なる段落が 1 つ追加／削除されている。
(2)	2 つの KAM は同等ではないが、一部の詳細を共有している。目安として、意味の異なる段落が 2 つ追加／削除されている。
(1)	2 つの KAM は同等ではないが、同じトピックについて述べている。
(0)	2 つの KAM は異なるトピックについて述べている。

各 KAM のペアは 2 名の作業者が独立してレビューした。そして、両者が付与したスコアの平均値をアノテーションとして採用した。なお、各作業者は、5 年以上の証券又は監査関連の業務経験を有している。

アノテーションの結果の分布を図 6 に示す。2 人の作業者によるアノテーションの一致度は、Cohen の Kappa 係数 [35] で測定したところ 0.88 であり、非常に良好な一致を示した。図 6 のとおり、本データセットの 78% においてアノテーションの結果は 4.0 から 5.0 の間となった。このことから、TOPIX Core 30 に選定されている上場会社の KAM において、ボイラープレート化の傾向は一定程度認められる。また、アノテーションの結果に対して、Shapiro-Wilk テストを使って正規性検定を行ったところ、正規性は認められなかった。

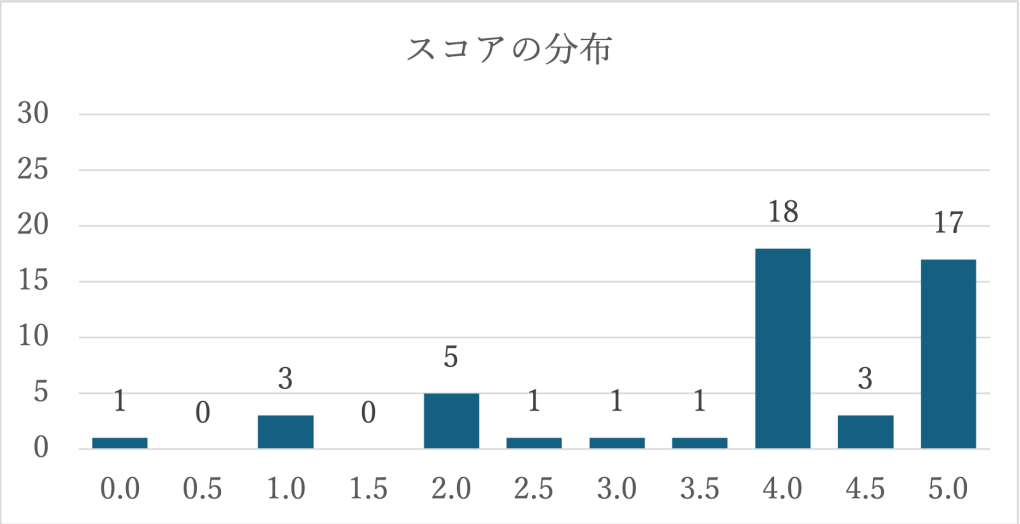

図 6　アノテーション結果に基づくスコアの分布

3.4　数値表現のマスキング

表 8 におけるクライテリアのとおり、本研究における KAM の意味的な類似性の評価においては、数値や時期の変化は意味の変化として捉えていないものとしている。このことから、自然言語処理の手法を用いた KAM の意味的な類似性の評価においても、数値表現の変化による影響は低減する方が望ましいと考えられる。

そこで、本研究では、KAM のテキストに含まれる数値表現が類似度算出に与える影響を低減することを目的として、数値表現のマスキングを行う。具体的には、財務諸表監査に関する金額、期日、割合、年度などの数値表現を特定し、これらを [MASK] という統一的な文字列に置き換える。これにより、前年度と本年度で数値だけが異なる場合においても、それらが類似度の評価を過度に左右することを防ぎ、KAM が本質的に意味する内容に基づいた類似性評価を可能にする。

表 9 に、実際の KAM のテキストを用いたマスキング結果のサンプルを示す。左側がマスキング前、右側がマスキング後のテキストである。なお、実際の監査報告書では金額や年度の情報がさらに複雑に記載されている場合も多いが、基本的な方針は同様である。

表 9　数値表現のマスキングの例

マスキング前	マスキング後
会社は 2015 年 3 月期から 2018 年 3 月期までの 4 事業年度につき、	会社は [MASK] 期から [MASK] 期までの [MASK] 事業年度につき、
2022 年 3 月末現在の連結財政状態計算書に、のれんを 339,904 百万円計上しており、総資産の 12.7% を占めている。	[MASK] 末現在の連結財政状態計算書に、のれんを [MASK] 円計上しており、総資産の [MASK]% を占めている。

3.5 類似度の評価指標

KAM のテキストの類似性を評価するため、3 種類の自然言語処理技術に基づく評価指標を採用した。その内訳は、単語の出現頻度に基づく評価指標、単語の一致率に基づく評価指標、文脈埋め込みベクトルに基づく評価指標である。以下では、それぞれの算出方法について詳細を述べる。

3.5.1 単語の出現頻度に基づく指標

第一の指標として、TF-IDF (Term Frequency-Inverse Document Frequency) [36] に基づくコサイン類似度を用いた指標を用いた。これは、文書内の単語の出現頻度と、文書集合全体におけるその単語の希少性を考慮に入れた重み付けを行い、文書間の類似度を計算するものである。通常、この類似度は 0 から 1 の範囲で評価され、1 に近いほど文書間の類似度が高いことを意味する。

この指標の計算に当たっては、データセット内の全 KAM に対して前処理を行った。この前処理には、形態素解析による単語の分割と単語の基本形への正規化が含まれる。その後、TF-IDF ベクトルを計算し、それらのベクトル間のコサイン類似度を用いて、KAM 間の類似性を測定した。コサイン類似度は、算出した各 TF-IDF ベクトルに対して、それらのベクトルの内積を、それぞれのベクトルのユークリッドノルムで割ったものとして定義される。

3.5.2 単語の一致率に基づく指標

第二の指標として、BLEU (BiLingual Evaluation Understudy) [37] を用いた。これは、翻訳の質を測るために開発された指標であり、元々は機械翻訳の評価に用いられていたが、本研究では KAM 文書間の類似性を評価するために応用する。BLEU は 0 から 100 の範囲で評価され、100 に近いほど、文書間の類似度が高いことを意味する。BLEU スコアは、ある文書の単語列が別の文書の単語列とどの程度一致するかを測るもので、n-gram (連続する n 個の単語の列) の一致率を基に計算される。しかし、BLEU スコアは単語の出現頻度や順序のみを考慮し、文脈や意味の類似性は直接反映されないという限界がある。

3.5.3 文脈埋め込みベクトルに基づく指標

第三の指標として、事前学習済み言語モデルを用いて生成された文脈埋め込みベクトルのコサイン類似度と、BERTScore を採用した。

事前学習済み言語モデルによって生成された文脈埋め込みベクトルは、単語やフレーズの意味をその使用される文脈に応じて捉える能力を持つ。本研究では、これらのモデルから埋め込みのベクトルを生成し、それらのコサイン類似度を KAM 間の意味的類似度と見なした。この値は 0 から 1 の範囲で出力され、1 に近いほど、文書間の類似度が高いことを意味する。

更に、本研究では BERTScore を評価指標として採用した。BERTScore は、文脈埋め込みベクトルを使用して、二つの文書間の意味的類似性を評価する指標である。BERTScore では、入力する 2 文はトークン単位に分割され、それぞれ対応する埋め込みベクトルを抽出し、それらのペアの中からコサイン類似度の高い組み合わせを求めて文間の類似度が計算される。BERTScore には Recall、Precision および

F1 スコアがあるが、本研究では F1 スコアを用いる。BERTScore は 0 から 1 の範囲で出力され、1 に近いほど、文書間の類似度が高いことを意味する。

事前学習済み言語モデルとしては、入力可能なトークン数の上限を考慮して、BERT[16] と Decoding-enhanced BERT with disentangled attention (DeBERTa) [38] に基づくモデルを採用する。ただし、これらのトークン数の上限は 512 であり、本データセットにおける KAM の最大トークン数はこの値を超過している。そこで、これらのスコアを算出する時は、入力トークンをまずは句点で分割し、ウィンドウをオーバーラップしながら複数のトークンに分割してモデルに入力し、それぞれの平均値を計算した。

3.6 評価実験

本節では、3.2 節と 3.3 節を通じて作成したデータセットと、3.4 節と 3.5 節に述べた手法を用いて、以下の項に述べるとおり、KAM のテキストの類似性の自動評価指標に関する評価実験を行った。

3.6.1 実験設定

実験の主な目的は、自動評価指標による類似性の評価が、3.3 節で作成した人手によるアノテーションとどの程度一致するかを調査することである。このため、各評価指標による KAM 文書間の類似度スコアを計算し、それらのスコアと人手によるアノテーション結果との相関を分析した。この分析には、アノテーションのスコアに正規性が認められなかったことを踏まえ、スピアマンの順位相関係数を用いて、評価指標の精度を定量的に評価した。

TF-IDF と BLEU におけるトークナイズには、MeCab[33] と UniDic[34] を用いた。また、BLEU スコアの算出には SacreBLEU[39] を使用し、KAM のテキストの短い変化と長い変化の評価バランスを考慮し、デフォルト値である n=4 とした。Embeddings のコサイン類似度と BERTScore の算出時に使用したモデルには、日本語に最適化された BERT モデル [40] と DeBERTa モデル [41] を採択した。

3.6.2 実験結果

表 10 に、各自動評価指標と人手アノテーションのスピアマンの順位相関係数の一覧を示す。

全体的な傾向として、マスキングの前後で全ての指標が改善した。マスキング前は順位相関係数の範囲は約 0.65 から約 0.79 であったが、マスキング後は約 0.84 から約 0.86 の範囲に収まっている。

個別の自動評価指標を見ると、まず、TF-IDF は、マスキングの前後のいずれにおいても順位相関係数が最も低い傾向が見受けられた。また、マスキング前は、BERT に基づく BERTScore の順位相関係数が最も高く、マスキング後は BLEU の順位相関係数が最も高かったものの、BERT に基づく BERTScore と大きな差は見受けられなかった。

3.6.3 考察

本研究では、年次間で金額、日付、割合といった数値情報のみが変化しても、KAM の実質的な意味が同等ならば高い類似と見なす評価系の構築を目指すものである。

実験の結果として、マスキングの前後で全ての指標において順位相関係数が一様に改善した。TF-IDF は 0.196 と最大の改善を示し、BLEU も 0.069 上昇した。これは、数値差分が表層一致を不必要に攪乱

表 10　各自動評価指標と人手アノテーションのスピアマンの順位相関係数の一覧

自動評価指標	モデル	スピアマンの順位相関係数	
		数値表現のマスク前	数値表現のマスク後
TF-IDF	MeCab + UniDic	0.649	0.845
BLEU	SacreBLEU (n=4)	0.787	0.856
BERTScore	BERT	0.794	0.855
BERTScore	DeBERTa	0.770	0.853
埋め込み表現のコサイン類似度	BERT	0.783	0.853
埋め込み表現のコサイン類似度	DeBERTa	0.759	0.844

注) 係数は、同一企業の 2 期分の KAM ペア ($n = 50$) に対する人手アノテーションとの順位相関 (Spearman の ρ) 。「数値表現のマスク」は、金額・日付・割合等を `[MASK]` に置換する前処理を指す。

していたこと、マスキングによってそのノイズが除去されたことを示唆する。埋め込み表現に基づく評価指標においても 0.06 から 0.09 程度の改善が得られており、意味ベースの手法でも数値差分の影響は無視できないことが分かる。

BLEU、BERTScore (BERT、DeBERTa) 、BERT による埋め込み表現のコサイン類似度を見ると、順位相関係数は約 0.85 から約 0.86 の範囲に収まった。これは、本データセットでの人手スコア分布が高類似領域 (4–5 点) に偏在することとも整合的で、年次での記述様式が近いケースが多いことを反映する。実務上は、大規模な KAM のデータセットを対象とした表層的なボイラープレート化の程度の計測には計算資源の軽い BLEU を、精密な類似性の評価には BERTScore や BERT による埋め込み表現のコサイン類似度を採用するなど、目的に応じた使い分けが考えられる。

なお、本分析は TOPIX Core 30 による 50 件のデータセットに基づくため、銘柄裾野の拡大や KAM の多様性が増す場合の頑健性の評価が課題となる。また、マスキングは数量変化自体の重要性を評価から切り離す設計であるため、数量変化の重大性を検知したいユースケースでは、数量変化を別途で把握するような仕組みが求められると考えられる。

3.7　小括

本章では、KAM の年次間の意味的な類似性を定量化する方法を提案し、TOPIX Core 30 の採用銘柄 30 社を対象に、2021 年度と 2022 年度の KAM を紐付けた 50 件のデータセットを構築したうえで、STS に倣う 0–5 の 6 段階基準で 2 名の有識者がアノテーションを付与し、提案手法とアノテーション結果の整合性を順位相関で検証した。併せて金額、日付、割合といった数値表現のマスキングを導入し、表層的な差分の影響を低減する手法の有用性を検証した。

実験の結果、全ての評価指標において数値表現のマスキング後に一様な改善が見られたことから、意味的な類似性の評価におけるマスキングの有用性が確認された。

また、本研究の結果から、BLEU において高い相関を示したことは注目に値する。これは、KAM のテキストが多くの場合、前年度の内容を基に若干の修正を加えたものであることが影響していると考えられる。その一方で、文脈埋め込みベクトルに基づく指標、特に BERTScore は、KAM のテキストが持

つ意味的な微妙な違いを捉える上で有効である可能性がある。また、マスキング後における BLEU と文脈埋め込みベクトルに基づく指標の間の差異は僅少であることから、運用目的に応じた使い分けが有効である。

以上より、KAM の意味的類似性測定において、数値マスキングを施した上での既存自動指標の活用が実務的に有効であることが示された。一方、データセットが大型銘柄に偏る点が限界として挙げられる。そのため、今後は対象の銘柄の拡大が課題となる。また、近年では、最大 8,192 トークンまで入力可能な ModernBERT-Ja-130M[42] といった言語モデルが出現していることから、このようなモデルの活用も今後の課題として挙げられる。

このように、KAM の意味的類似性の測定において、自然言語処理の手法が有用であることが示唆された。

4　監査法人における KAM の記載内容の管理に関する分析

本章では、語彙多様性と著者推定の精度に基づく、監査法人における KAM の管理の程度に関する分析について述べる。

本章の構成は次のとおりである。第 4.1 節では、本研究に関する先行研究についてレビューする。第 4.2 節では、分析視点と仮説の設定についてまとめる。第 4.3 節では、使用するデータセットや指標をはじめとする研究デザインについて述べる。第 4.4 節では、語彙多様性に関する実験を行い、第 4.5 節では著者推定に関する実験を行う。第 4.6 節では、本研究に関する限界について述べる。最後に、第 4.7 節で本研究の結論を小括する。

4.1　関連研究

本節では、監査品質に関する関連研究、語彙多様性に関する関連研究、著者推定に関する関連研究についてそれぞれ述べる。

4.1.1　監査品質に関する関連研究

監査品質は、一般的に「財務諸表の重要な虚偽表示を発見し、それを是正するよう依頼人に要請する監査人の能力と意図」と定義される [43]。この定義によれば、監査人が十分な専門知識と独立性を保つことで、高品質な監査が行われると考えられる [44]。監査品質に影響を与える要因としては、監査法人の規模 (いわゆる Big 4 と非 Big 4 の比較) 、監査報酬、監査チームの専門性、内部統制の整備状況などが挙げられてきた [45, 46]。

これまでの研究では、大手監査法人ほど監査の専門性や名声による損失回避インセンティブが高く、より厳格な監査を行う傾向があると報告されている [47]。また、大手監査法人は世界的なネットワークと統一された監査手法を有しているため、法人全体で監査品質を一定水準以上に保つ仕組みを整えていると考えられている。一方で、大手監査法人ではない場合、リソースが限られていることや専門分野の人材確保が難しいことなどから、監査品質にばらつきが生じやすいという指摘もある [48]。

4.1.2　語彙多様性に関する関連研究

語彙多様性は、文書に含まれる単語の多様性を定量化する代表的な指標である [49]。一般に、語彙多様性が高い文書ほど、文体や表現が多岐にわたり、より詳細かつ柔軟な説明が行われている可能性がある。一方で、専門分野における文書では、使用する用語の専門性や頻出語彙が限定的であるため、語彙多様性が低くなりがちであるとも指摘される [50]。

語彙多様性の代表的な評価指標として、Type Token Ratio (TTR) が知られているが、TTR はテキスト長に大きく依存する性質がある。そのため、テキスト長が大きく異なりうるテキスト間の語彙多様性の比較においては、TTR は適切ではない。McCarthy & Jarvis[49] によると、Measure of Textual Lexical Diversity (MTLD) [49]、vocd-D[51]、Hypergeometric Distribution Diversity (HD-D) [52]、Maas[53] は、独自の語彙情報を捉えており、語彙多様性に関する研究ではこれらの使用が推奨されている。

4.1.3　著者推定に関する関連研究

著者識別は、与えられたテキストの書き手が誰であるかを推定する手法であり、計量文献学の一分野として発展してきた [54]。初期には、単語の出現頻度や n-gram に基づく手法が提案されていたが、word embeddings を始めとする深層学習の手法が普及すると、より高次元の文書ベクトルを用いた著者識別が試みられるようになった。例として、BERT 系モデルによる文章の特徴ベクトルを抽出し、機械学習手法で著者識別を行う事例も報告されている [55]。また、KAM の著者推定に関する研究として、トピックモデルを使用して、監査人が所属する監査法人単位での KAM の類似性の評価を行った事例が報告されている [56]。

4.2　分析視点と仮説の設定

本研究では、KAM における語彙多様性と著者推定精度を手がかりに、監査法人が KAM の内容と表現をどの程度組織的に管理しているかを検証する。先行研究の知見を踏まえると、大手監査法人ほど組織的な品質管理を行うインセンティブとリソースを有しており、より一貫した文書作成方針が徹底されていると考えられる。一方で、中小規模監査事務所では、個々のパートナーやスタッフの裁量に左右される度合いが大きいため、文面上の統一性が相対的に低いことが予想される。

これらを踏まえて、本研究では、次のように 2 つの仮説を設定する。

仮説 1: 大手監査法人、準大手監査法人、中小規模監査事務所の順で、KAM の語彙多様性が高い。

語彙多様性は、文書内でどの程度多彩な言葉を用いているかを示す指標である。大手監査法人は顧客規模が大きく、業種も多岐にわたるため、取り扱う会計論点やリスク評価の内容も多様である。そして、大手監査法人では、KAM の品質管理の結果として、監査対象の各社に応じた内容が記載されると考えられる。これらの結果として、大手監査法人では、KAM の記載に用いられる語彙も幅広くなることが予想される。一方で、それ以外の監査法人では、担当企業の規模や業種が限られる場合が多く、取り上げる会計論点にも類似点が多いため、語彙多様性は相対的に低くなると考えられる。

仮説 2: 大手監査法人、準大手監査法人、中小規模監査事務所の順で、KAM に関する著者推定の精度が高い。

大手監査法人は、組織としてマニュアルや体制を整備し、KAM の記載内容が最終的に法人の標準的な表現やスタイルに修正されている可能性がある。一方、中小規模監査事務所は人材の流動性やリソース不足等の要因から、監査人個々の文書作成スタイルがそのまま KAM に反映されやすく、結果として組織内部で統一された文体が生まれにくい可能性がある。そのため、著者推定モデルの精度は、大手監査法人の方が高く、中小規模監査事務所の方が低くなると考えられる。

これらの仮説が立証されると、大手監査法人においては、各社の事情に即した内容の KAM が記載されつつも、その文体は各監査法人で統一化されていることを意味し、すなわち、大手監査法人では KAM の内容が管理されていることを示すことに繋がる。

4.3 リサーチデザイン

本節では、本研究におけるデータとサンプルの選定方法、語彙多様性の測定手法、著者推定モデルの構築方法について述べる。

4.3.1 データセット

本研究では、2021 年 3 月期 (以下、2021 年度) 、2022 年 3 月期 (以下、2022 年度) 、2023 年 3 月期 (以下、2023 年度) 、2024 年 3 月期 (以下、2024 年度) 、2025 年 3 月期 (以下、2025 年度) の 5 期分の KAM のテキストを分析の対象とする。具体的には、EDINET で公表された監査報告書から、KAM の「内容及び決定理由」と「監査上の対応」のテキストを抽出した。なお、日本の上場会社の監査報告書には、「当期連結財務諸表に対する監査報告書」と「当期財務諸表に対する監査報告書」の 2 種類がある。本研究では「当期連結財務諸表に対する監査報告書」を分析の対象とした。

また、一定数の KAM を記載している監査法人を対象とするため、原則として 5 年間において 10 社以上の上場会社の監査を行っている監査法人を対象とした。更に、公認会計士・監査審査会の報告における定義 [57, 58] を踏まえ、4 法人を大手監査法人、5 法人を準大手監査法人、8 法人を中小規模監査事務所に分類した。なお、2023 年 12 月 1 日、PwC あらた有限責任監査法人は PwC 京都監査法人を吸収合併し、PwC Japan 有限責任監査法人に改称した。そのため、本研究では、2023 年度以前の準大手監査法人は 5 法人、2024 年度以降の準大手監査法人は 4 法人として取り扱う。

本研究では、特定の監査法人へのバイアスを除去するため、大手監査法人を A1–A4、準大手監査法人を B1–B5、中小規模監査事務所を C1–C8 にそれぞれ匿名化した。本データセットにおける各監査法人の KAM の件数の推移を表 11 に示す。

4.3.2 語彙多様性の評価指標

本研究の一つ目の分析視点である語彙多様性は、年度及び監査法人ごとの「内容及び決定理由」と「監査上の対応」を対象として測定する。

まず、同一年度及び同一監査法人の KAM をランダムに 10 件ずつ選択する。そして、それぞれの KAM から「内容及び決定理由」と「監査上の対応」のテキストを抽出する。そして、10 件の「内容及び決定理由」のテキストを 1 つのテキストへ連結し、同様に 10 件の「監査上の対応」も 1 つのテキストへ連結する。そうして、各年度の各監査法人別に、10 件の KAM で構成される「内容及び決定理由」と「監査上の対応」のテキストを得る。これにより、KAM の個数によるテキスト長の増大によるバイアスを防ぐこととする。

また、この時、同じ規模の監査法人において偏りが生じないよう、表 11 における各監査法人の KAM の件数を踏まえて、大手監査法人では 10 件× 11 グループ、準大手監査法人では 10 件× 3 グループ、中小規模監査事務所では 10 件× 1 グループの KAM をランダムに抽出した。その後、NFKC 正規化をはじめとする正規化処理をこれらのテキストに対して行う。そして、MeCab[33] 及び UniDic[34] を用いて、これらのテキストを形態素単位に分割する。

次に、形態素単位に分割された各テキストについて、複数の語彙多様性の評価指標の値を計算す

表 11　監査法人別の KAM の個数の推移

監査法人	2021 年	2022 年	2023 年	2024 年	2025 年
A-1	688	647	609	1174	1136
A-2	612	565	540	1018	1010
A-3	592	564	540	1010	970
A-4	139	118	117	294	292
B-1	162	172	181	364	334
B-2	53	61	49	94	82
B-3	47	50	49	110	106
B-4	34	34	37	72	64
B-5	40	34	34	–	–
C-1	32	34	31	70	72
C-2	25	26	27	28	30
C-3	18	15	17	28	28
C-4	16	16	20	36	34
C-5	15	14	16	26	28
C-6	14	16	15	26	20
C-7	14	14	10	20	18
C-8	10	12	12	18	16

注) 数値は各年度の KAM 件数。– は該当データなしを示す。

る。本研究では、McCarthy & Jarvis[49] に基づき、Maas、HD-D、MTLD の 3 指標を採用する。なお、McCarthy & Jarvis[49] によると、HD-D は vocd-D の代替となることが述べられているため、vocd-D による仮説の検証は省略した。

以下に、Maas、HD-D、MTLD の 3 指標の概要と計算について述べる。

■Maas　Maas は、Type-token ベースの評価指標の一つであり、TTR と比較して文字列長に対する依存度が比較的低いと捉えられている。Maas の値 ($Maas's\ \alpha$) は、式 (1) のとおり計算される。

$$Maas's\ \alpha = \frac{\log N - \log V}{(\log N)^2} \tag{1}$$

ここで、N は文書の総形態素数、V は文書中の異なり形態素数 (types) を表し、$Maas's\ \alpha$ が小さいほど文書内で重複しない語彙が多く、語彙多様性が高いと解釈される。本稿では、以下、$Maas's\ \alpha$ を単に Maas と記す。

■HD-D (Hypergeometric Distribution Diversity)　HD-D は統計処理に基づく語彙多様性の指標である。テキスト内の各単語について、固定サイズ n の無作為標本においてその語が少なくとも 1 回出現する確率を、ハイパージオメトリック分布を用いて計算する。この確率を全語種にわたって合計し、n で割ることで HD-D 値が得られる。これは n トークンからなる任意のサンプルにおける異なる単語の種類の期待値である。この値は期待値から導かれるため、テキスト全体の長さにほぼ影響されず、規模が大きく異なるテキスト間でも語彙の多様性を信頼性をもって比較できる。HD-D は、式 (2) のとおり計算さ

れる。

$$HD\text{-}D = \frac{1}{n}\sum_{i=1}^{V}\left(1 - \frac{\binom{N-f_i}{n}}{\binom{N}{n}}\right) \tag{2}$$

ここで、N は文書における総トークン数、V は文書中の異なり形態素数、f_i はそれぞれのユニークトークン i の出現頻度、n はサンプリングサイズを表す。また、$\binom{x}{y}$ は二項係数を表し、「x 個の中から y 個を取り出す組み合わせの総数」を意味する。なお、一般的には、サンプリングサイズは 42 が選択されている。HD-D の値が大きいほど、多様な単語が含まれている、すなわち語彙多様性が高い文書とみなされる。

■MTLD (Measure of Textual Lexical Diversity)　MTLD は、文書を単語列 (形態素列) の塊 (segment) に分割し、それぞれの塊に含まれるユニークな単語の比率が一定の基準に達した段階で次の塊へ移動していく手順を通じて、多様性を測定する手法である。MTLD の値は、一定レベルの語彙の多様性を維持するために必要なテキストの平均的な長さと捉えられる。MTLD の算出は、次の 3 ステップで構成される。

1. 文書の冒頭から 1 トークンずつ読み込み、TTR がしきい値に達するまでを 1 セグメントとする。
2. しきい値に達すると、同様に次のセグメントが作成される。
3. 全セグメントの平均セグメント長が計算される。
4. 平均セグメント長をテキストの総長で割ることで MTLD が算出される。

具体的な算出式は、式 (3) のように表される。

$$MTLD = \frac{N}{S + \left(\frac{1-TTR_{\text{end}}}{1-\theta}\right)} \tag{3}$$

ここで、N は文書における総トークン数、S はしきい値に達したセグメントの数、θ はしきい値、TTR_{end} は最終セグメントの TTR を表す。しきい値の θ としては、0.72 が採用されることが多い。MTLD の値が大きいほど、語彙多様性が高いと解釈される。

4.3.3　著者推定の手法

本研究のもう一つの分析視点である、著者推定の精度は、KAM 文書を「どの監査法人が作成したものか」を判別するタスクとして位置付けることで測定する。つまり、各 KAM を特徴ベクトルに変換し、監査法人をラベルとした分類問題を設定する。著者推定モデルの精度が高いほど、「監査法人ごとに文書の書きぶり (スタイル) が一貫している」ことを意味すると考えられる。

本研究では、著者推定モデルの学習のため、事前学習済みモデルとして日本語の BERT モデルを利用する。ただし、KAM の「内容及び決定理由」と「監査上の対応」のテキストは、BERT の一般的な入力上限である 512 トークンを超える傾向がある。そのため、本研究では、KAM のテキストの先頭近辺と末尾近辺に監査法人ごとの特徴がより強く現れると考え、これらのテキストについて先頭 512 トークンと末尾 512 トークンを抽出し、それぞれを同一ラベルとして学習データに取り込む方法を採用した。

これにより、監査法人ごとの KAM の特徴を捉えながら、著者推定モデルの学習を実現している。本手法の概要を、図 7 に示す。

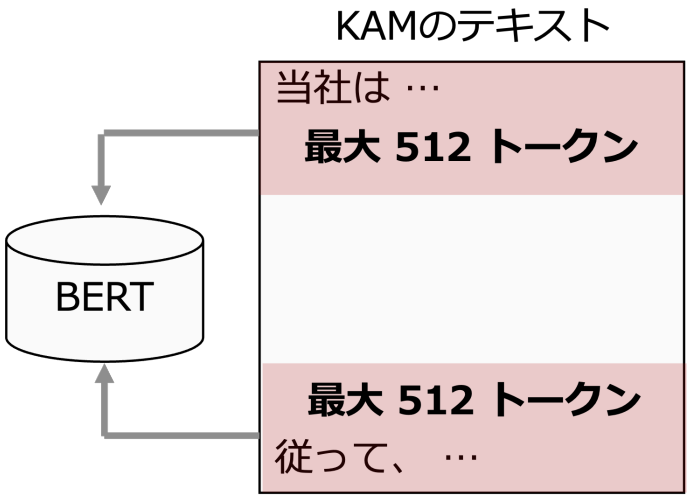

図 7　著者推定における BERT の学習に使用する KAM のテキストの概要

より安定したモデル評価を行う目的で、5 分割の交差検証を導入する。すなわち、KAM のデータセットの 5 分の 1 を著者推定の精度の評価に使用し、全ての KAM がその評価に使用されるよう、これを 5 回繰り返す。最終的に、監査法人別にこの 5 回分の平均値を計算することで、監査法人別の著者推定の精度を導出する。

4.4　語彙多様性に関する実験

本節では、本研究が採用した 3 種類の語彙多様性指標 (Maas、HD-D、MTLD) を用いて KAM における語彙多様性を測定し、その結果を報告する。

4.4.1　実験設定

本データセットにおける各 KAM のテキストに対しては、NFKC 正規化を行った後、MeCab と UniDic を用いて形態素解析を行い、名詞・動詞・形容詞・副詞などの形態素レベルに分割した。

先行研究に倣い、HD-D のサンプリングサイズの n は 42、MTLD で使用する TTR のしきい値は 0.72 とした。

4.4.2　実験結果

各監査法人について、5 期度分のサンプルから得られた語彙多様性 (Maas、HD-D、MTLD) の平均値を算出し、大手監査法人 (A1–A4) 、準大手監査法人 (B1–B5) 、中小規模監査事務所 (C1–C8) のクラス別に集計した。表 12 及び表 13 は、3 種類の指標について年度ごとに平均値をまとめたものである。

Maas は、値が低いほど語彙多様性が高いと解釈され、HD-D と MTLD は値が高いほど語彙多様性が高いと解釈される。表 12 及び表 13 より、いずれの評価指標でも、3 つのクラスの中で、大手監査法人の語彙多様性が最も高い傾向を示した。特に、HD-D は、一貫して、大手監査法人＞準大手監査法人＞

中小規模監査事務所の順に語彙多様性が高い傾向を示しており、年度差による大きな変動は見られなかった。

統計学的にも検証するため、さらに、Tukey–HSD による検定 (5% 水準) を実施した。この結果を表 14 に示す。表 14 のとおり、特に HD-D において、大手監査法人と中小規模監査事務所、大手監査法人と準大手監査法人、準大手監査法人と中小規模監査事務所のいずれのペアにおいても有意差が認められた。

表 12　語彙多様性指標の年度別推移 (内容及び決定理由)

評価指標	規模	2021 年	2022 年	2023 年	2024 年	2025 年	平均
Maas	大手監査法人	0.0273	0.0272	0.0269	0.0269	0.0270	0.0271
	準大手監査法人	0.0274	0.0271	0.0272	0.0271	0.0275	0.0272
	中小規模監査事務所	0.0287	0.0285	0.0280	0.0284	0.0284	0.0284
HD-D	大手監査法人	0.8118	0.8111	0.8117	0.8133	0.8126	0.8121
	準大手監査法人	0.8080	0.8102	0.8079	0.8091	0.8084	0.8087
	中小規模監査事務所	0.8026	0.8010	0.8032	0.8038	0.8027	0.8026
MTLD	大手監査法人	58.4935	58.1984	58.1420	57.6568	57.8547	58.0691
	準大手監査法人	58.2540	58.1833	57.3869	58.0288	57.5774	57.8933
	中小規模監査事務所	57.0173	56.3265	59.2736	57.4718	57.9392	57.6057

注) Maas は値が**小さい**ほど語彙多様性が高く、HD-D・MTLD は値が**大きい**ほど語彙多様性が高いと解釈する。

表 13　語彙多様性指標の年度別推移 (監査上の対応)

評価指標	規模	2021 年	2022 年	2023 年	2024 年	2025 年	平均
Maas	大手監査法人	0.0300	0.0297	0.0297	0.0298	0.0298	0.0298
	準大手監査法人	0.0305	0.0301	0.0300	0.0299	0.0304	0.0302
	中小規模監査事務所	0.0315	0.0310	0.0309	0.0312	0.0308	0.0311
HD-D	大手監査法人	0.7646	0.7661	0.7667	0.7669	0.7664	0.7661
	準大手監査法人	0.7599	0.7633	0.7639	0.7651	0.7637	0.7631
	中小規模監査事務所	0.7567	0.7573	0.7547	0.7568	0.7607	0.7572
MTLD	大手監査法人	49.6956	49.3724	49.3798	49.1045	49.4004	49.3905
	準大手監査法人	48.3184	48.4697	48.9176	49.2503	48.8091	48.7290
	中小規模監査事務所	47.0216	46.9647	46.9794	48.0687	48.3416	47.4752

注) Maas は値が**小さい**ほど語彙多様性が高く、HD-D・MTLD は値が**大きい**ほど語彙多様性が高いと解釈する。

4.4.3　考察

本実験の結果、3 種類の語彙多様性指標 (Maas、HD-D、MTLD) のいずれにおいても、大手監査法人が最も語彙多様性が高くなる傾向が確認された。特に、HD-D においては 大手監査法人、準大手監査法

表 14　語彙多様性に関する Tukey–HSD の結果 (5% 水準)

評価指標	区分 1	区分 2	p 値 (内容及び決定理由)	p 値 (監査上の対応)
Maas	大手監査法人	準大手監査法人	0.3418	**0.0001**
	大手監査法人	中小規模監査事務所	**<0.0001**	**<0.0001**
	準大手監査法人	中小規模監査事務所	**<0.0001**	**0.0001**
HD-D	大手監査法人	準大手監査法人	**<0.0001**	**0.0212**
	大手監査法人	中小規模監査事務所	**<0.0001**	**<0.0001**
	準大手監査法人	中小規模監査事務所	**<0.0001**	**0.0014**
MTLD	大手監査法人	準大手監査法人	0.6360	0.0921
	大手監査法人	中小規模監査事務所	0.8775	**<0.0001**
	準大手監査法人	中小規模監査事務所	0.9037	**0.0168**

注) 太字は 5% 水準で有意 (Tukey–HSD) 。

人、中小規模監査事務所の間に統計的に有意な差が認められたことから、語彙多様性の観点において、監査法人の規模間には明確な差異があると考えられる。

この結果は、仮説 1 (「大手監査法人、準大手監査法人、中小規模監査事務所の順で、KAM の語彙多様性が高い」) を概ね支持するものである。先行研究では、大手監査法人は世界規模のネットワークと統一された監査手法を有するため、幅広い業種・会計論点に対応する豊富なリソースを備えていることが指摘されてきた。KAM においても、各業種や企業特有のリスク評価や論点を踏まえたバリエーション豊かな記載が行われることで、結果として語彙多様性が高まった可能性がある。一方、中小規模監査事務所は担当する企業の規模や業種が限定的であるため、使用語彙が特定の会計論点に偏りやすいことから、語彙多様性が相対的に低いと推察される。

以上より、KAM における語彙多様性という観点からも、大手監査法人ほど KAM の内容をより多様な表現で作成している可能性が示唆される。これは、大手監査法人が組織的な品質管理体制を整備しつつ、各クライアント企業の特性に応じたきめ細かい記述を行っていることの表れとも考えられる。

4.5　著者推定に関する実験

本節では、著者推定に関する実験について、実験設定、実験結果、考察を述べる。

4.5.1　実験設定

本実験では、前節と同様、本データセットを利用する。ただし、4.3 節で述べた通り、本実験では、各 KAM の先頭と末尾の 512 トークンを学習に使用した。また、5 分割の交差検証を実施し、全ての監査法人に対して平均的な推定精度を算出した。

学習には Transformers 系のライブラリを用い、日本語の事前学習モデルとして、tohoku-nlp/bert-base-japanese-v3[59] を使用した。そして、事前学習済みモデルを初期化し、Classification head により監査法人ラベルを分類する。最適化には AdamW を用い、学習エポック数は 10、バッチサイズは 2 とした。各監査法人の推定に係る評価指標としては、Accuracy を採用した。

4.5.2　実験結果

5 期分の各監査法人の平均 Accuracy を表 15 及び表 16 に示す。

表 15 及び表 16 のとおり、「内容及び決定理由」と「監査上の対応」のいずれにおいても、大手監査法人 (A-1–A-4) の推定精度が相対的に高く、特に A-2 に至っては 2021 年度から 2025 年度の各年度で 80–90% 台という高い Accuracy を示した。一方、準大手監査法人 (B-1–B-5) および中小規模監査事務所 (C-1–C-8) は、いずれの年度においても、大手監査法人より Accuracy が低く、全体的に 20–40% 前後にとどまるケースが多かった。

規模別平均を見ると、大手監査法人が「内容及び決定理由」で約 61.92%、「監査上の対応」で約 71.53% と高い値を示すのに対し、準大手監査法人ではそれぞれ約 21.28%、約 27.52%、中小規模監査事務所ではそれぞれ約 12.32%、約 15.13% となった。このことから、規模別の傾向として、著者推定の推定精度は大手監査法人＞準大手監査法人＞中小規模監査事務所の順で高い傾向があった。

統計学的にも検証するため、さらに、Tukey–HSD による検定 (5% 水準) を実施した。表 17 にこの結果を示す。この結果、いずれの規模の間においても、有意差が認められた。

表 15　著者推定の精度 (内容及び決定理由)

監査法人	2021 年	2022 年	2023 年	2024 年	2025 年	平均	規模別平均
A-1	76.45	71.99	62.87	61.67	62.50	67.10	61.92
A-2	93.79	90.48	84.66	82.32	80.59	86.37	
A-3	63.03	60.49	55.80	57.13	48.45	56.98	
A-4	44.57	36.44	36.32	37.07	31.85	37.25	
B-1	48.77	44.86	34.14	41.48	35.33	40.92	21.28
B-2	26.42	17.74	13.27	6.38	9.76	14.71	
B-3	17.39	6.86	10.00	8.18	11.32	10.75	
B-4	26.47	6.06	20.27	8.33	23.44	16.91	
B-5	32.50	19.12	17.65	–	–	23.09	
C-1	37.50	16.18	22.58	11.43	13.89	20.31	12.32
C-2	12.50	7.41	7.14	0.00	6.67	6.74	
C-3	38.89	26.67	23.53	7.14	0.00	19.25	
C-4	9.38	15.63	15.00	19.44	2.94	12.48	
C-5	26.67	14.29	37.50	15.38	7.14	20.20	
C-6	14.29	15.63	6.67	0.00	0.00	7.32	
C-7	28.57	17.86	10.00	5.00	0.00	12.29	
C-8	0.00	0.00	0.00	0.00	0.00	0.00	

注) 各値は百分率 (%)。– は該当データなしを示す。

4.5.3　考察

著者推定に関する実験の結果は、仮説 2 (「大手監査法人、準大手監査法人、中小規模監査事務所の順で、著者推定の精度が高い」) を概ね支持するものである。すなわち、大手監査法人ほど組織的なマニュ

表 16　著者推定の精度 (監査上の対応)

監査法人	2021 年	2022 年	2023 年	2024 年	2025 年	平均	規模別平均
A-1	86.19	85.19	83.36	81.69	76.23	82.53	71.53
A-2	95.75	96.56	93.53	72.30	87.92	89.21	
A-3	73.77	76.90	73.57	52.67	61.44	67.67	
A-4	54.35	54.24	51.28	36.05	37.67	46.72	
B-1	62.65	73.43	66.94	41.76	52.10	59.37	27.52
B-2	18.87	31.45	16.33	12.77	4.88	16.86	
B-3	19.57	23.53	12.00	5.45	20.75	16.26	
B-4	38.24	15.15	24.32	8.33	12.50	19.71	
B-5	35.00	23.53	17.65	–	–	25.39	
C-1	43.75	29.41	22.58	20.00	27.78	28.70	15.13
C-2	31.25	22.22	25.00	0.00	0.00	15.69	
C-3	27.78	13.33	11.76	7.14	7.14	13.43	
C-4	0.00	6.25	5.00	0.00	5.88	3.43	
C-5	33.33	50.00	37.50	0.00	14.29	27.02	
C-6	14.29	0.00	26.67	0.00	10.00	10.19	
C-7	28.57	14.29	10.00	10.00	0.00	12.57	
C-8	50.00	0.00	0.00	0.00	0.00	10.00	

注) 各値は百分率 (%) 。– は該当データなしを示す。

表 17　著者推定に関する Tukey–HSD の結果 (5% 水準)

区分 1	区分 2	p 値 (内容及び決定理由)	p 値 (監査上の対応)
大手監査法人	準大手監査法人	**<0.0001**	**<0.0001**
大手監査法人	中小規模監査事務所	**<0.0001**	**<0.0001**
準大手監査法人	中小規模監査事務所	**0.0460**	**0.0179**

注) 太字は 5% 水準で有意 (Tukey–HSD) 。

アルやレビュー体制が整備されている影響で、最終的に監査法人固有の文書スタイルへと集約されている可能性が高いと考えられる。一方、中小規模監査事務所では、監査チームごとの裁量が相対的に大きく、文書レビューやフィードバック体制などのリソースも限られるため、統一されたスタイルが形成されにくい傾向がある可能性がある。

ただし、それぞれの監査法人の規模に属する全ての監査法人について、一貫した傾向を結論づけることは難しいと考えられる。例えば、大手監査法人における A-4 の Accuracy は、大手監査法人の中でも相対的に低く、準大手監査法人の中で相対的に Accuracy が高い B-1 と同程度であった。同様に、中小規模監査事務所の C-5 は、準大手監査法人と同程度の Accuracy が確認された。このことは、それぞれの規模の監査法人の中でも、組織的なマニュアルやレビュー体制の程度が異なる可能性を示唆している。そのため、今後の研究においては、監査法人の品質管理体制の質や手続きの詳細、クライアント企業の属性などをさらに統制したうえで、より精緻に著者推定を行う研究が求められると考えられる。

4.6　本研究の限界

本研究には、主に以下に述べる 2 点の限界がある。

1 点目は、著者推定における語彙への偏重に関する恐れである。本研究では、事前学習済みモデルを用いた分類タスクとして著者推定を行い、監査法人を識別するモデルを構築した。この手法は、KAM が有する監査法人固有の文体を捉えられる点は有用であるが、KAM の文体ではなく語彙の特徴を強く捉えて監査法人を識別している可能性が否定できない。そのため、今後の研究では、語彙に左右されず、文体で著者推定を行う著者推定モデルによる著者推定が求められる。

2 点目は、監査品質との因果関係の不確実性である。語彙多様性や著者推定精度の高低と、監査品質そのものの高低は必ずしも同義ではない。テキストの特徴量から監査品質を直接的に測定することは困難であり、今回の分析はあくまで「KAM の内容が監査法人内部でどの程度統一管理されているか」を推察する一助と位置付けられる。今後は、KAM と監査品質の関係を検証する研究が求められる。

4.7　小括

本研究では、日本の上場会社の監査報告書に含まれる KAM について、語彙多様性と著者推定精度に基づく分析の結果、大手監査法人ほど多様な語彙を用いながらも統一的な文書スタイルを持つ傾向が確認された。この結果は、大手監査法人が組織的なレビュー体制やマニュアルを通じて KAM の内容を管理している可能性を示唆している。一方、中小規模監査事務所では、語彙の幅が狭く著者推定の精度が低いことがわかり、組織的に KAM を管理するよりも、監査チームに委ねられている側面が大きいことが推察される。

ただし、本研究の結果は、あくまで語彙多様性や著者推定の観点からの示唆であり、実際の監査品質そのものを直接評価するものではない。今後の研究では、文体要素を細分化した著者推定手法の導入や、監査報酬や将来の不正発覚リスクなどとの関連を検証することで、KAM が監査や投資判断にもたらす影響をさらに明らかにすることが期待される。

このように、監査法人における KAM の記載内容の管理に関する分析において、自然言語処理の手法が有用であることが示唆された。

5　おわりに

本稿では、KAM に対する自然言語処理を用いた分析を通じて、(1) LLM による監査領域のゼロショット分類、(2) KAM のテキストの意味的な類似性の評価手法、(3) 記載内容の管理の程度に関する分析、の 3 つの観点から実験を行い、KAM の分析における自然言語処理の手法の有用性を検証した。第 2 章の結果から、LLM を用いたゼロショット分類により、人手によるデータセットを要さず、高精度に監査領域を推定できることが確認された。第 3 章の結果から、KAM の類似性の測定において、数値表現のマスキングにより人手評価との整合性が大きく向上し、また、単語の一致率に基づく指標がボイラープレート化の程度を測定するのに有効であることが示され、文脈埋め込みベクトルに基づく評価指標が KAM の意味的類似性を捉える上で有効であることが示された。第 4 章の結果から、大手監査法人は語彙が豊富でありながら、著者推定の精度が高い傾向があり、規模の小さい監査法人ほど、語彙が少ないにもかかわらず、著者推定の精度は低い傾向があった。この結果は、大手監査法人ほど組織的な品質管理が行われ、KAM の内容が管理されている可能性を示唆している。

本稿の結果は、KAM の活用に関する実務に対して、少なくとも次の三点の含意を与える。第一に、監査領域の自動分類は、人手審査の「前処理」として活用できる可能性がある。ゼロショットテキスト分類と人手レビューを組み合わせた二段階運用は、作業負荷と正確性のバランスが良いと考えられる。また、複数のモデルの出力結果のアンサンブルを採用することで、一層の精度の向上が期待される。第二に、KAM の意味的類似性の計測においては、目的に応じて適した手法があることが示唆されている。すなわち、KAM のボイラープレート化の測定においては BLEU をはじめとする単語の一致率に基づく手法が有用であり、より厳密な意味的類似性を捉えるためには BERTScore をはじめとする文脈埋め込みベクトルに基づく手法が有用である。第三に、KAM の品質管理に関する可視化の手法として、語彙多様性と著者推定精度の有用性が示唆されている。ただし、これらの指標は監査品質そのものの代理ではないため、その解釈には慎重になる必要があると考えられる。

もっとも、実務への展開に当たっては、本稿の分析設計に起因する限界を踏まえた慎重な運用が求められる。第 2 章のゼロショットテキスト分類は、2022 年 4 月期〜2023 年 3 月期に EDINET から抽出した 4,202 件の KAM に基づき、そのうち約 5% (250 件) に人手ラベルを付して評価した単年度の分析であるため、必ずしも KAM 全体の傾向を把握できているものではなく、また、年度固有の影響を完全には排除できない。更に、残差カテゴリである「その他」は依然として判定難度が高いことが確認されている。第 3 章の意味的類似性の検証については、TOPIX Core 30 を対象に 2021 年度と 2022 年度の対応付けから得た 50 件の KAM で構成しており、数値表現のマスキングが人手評価との整合性を高めることを示した一方、銘柄の裾野の拡大は今後の課題として残る。第 4 章のうち、著者推定に関しては、KAM を監査法人ラベルで分類する枠組みに基づくものであり、学習したモデルが文体そのものではなく語彙特徴に強く依存して識別している可能性、また推定精度と監査品質の因果関係を直接には示さないという制約がある。従って、著者推定の結果は「文面の管理の度合いを示唆する指標」として解釈し、監査品質の優劣の断定には用いるのは慎重になる必要がある。

今後の研究課題としては、第一に、ゼロショットテキスト分類において、監査領域の階層化と細分化を行い、「その他」への分類を減らすよう、監査領域を再定義することが挙げられる。第二に、意味的類

似性の測定において、長文への対応力が高いモデルの活用や分割入力戦略の最適化により、KAM 全体の文脈を途切れなく扱うようにすることが挙げられる。第三に、著者推定において、文体要素に焦点を当てることで、語彙依存性を下げたスタイルの識別を検討する余地がある。第四に、年度をまたいだ長期パネルを構築し、意味的類似性、語彙多様性、著者推定の精度といった指標が、時系列でどう推移するかを観察することが有益である。第五に、監査領域別に、意味的類似性、語彙多様性、著者推定の精度といった指標の傾向を確認することが挙げられる。

総じて、本稿は、KAM における「監査領域の自動分類」「意味的類似性」「記載内容管理の実証」という三つの観点に対し、自然言語処理が現実的な精度で寄与しうること、そして実務上の応用可能性があることを示した。本稿の知見が、KAM の開示実務の高度化と監査の透明性と信頼性の一層の向上、さらにはデータ駆動の監査研究の進展に資する端緒となることを期待する。

参考文献

[1] Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. Topic Classification of Key Audit Matters in Japanese Audit Reports by Zero-shot Text Classification. In *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 540–545, 2023. doi: 10.1109/IIAI-AAI59060.2023.00108.

[2] Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. Measuring Semantic Similarity in Japanese Key Audit Matters. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 468–475, 2024. doi: 10.1109/IIAI-AAI63651.2024.00091.

[3] Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. Do Audit Firms in Japan Manage the Content of Their Key Audit Matters? Evidence Based on Lexical Diversity and Author Identification Accuracy. In *2025 18th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2025. to appear.

[4] 日本公認会計士協会. 独立監査人の監査報告書における監査上の主要な検討事項の報告. **監査基準委員会報告書** *701*, 2024. https://jicpa.or.jp/specialized_field/2-24-701-2-20240926.pdf (accessed on 2025-10-30).

[5] 企業会計審議会. 監査基準の改訂に関する意見書, 2018. https://www.fsa.go.jp/news/30/sonota/20180706/1.pdf (accessed on 2025-10-30).

[6] 企業会計審議会. 「監査報告書の透明化」についての主な論点 (１), 2018. https://www.fsa.go.jp/singi/singi_kigyou/siryou/kansa/20180126/20180126/2.pdf (accessed on 2025-10-30).

[7] 金融庁. 監査上の主要な検討事項 (KAM) の特徴的な事例と記載のポイント 2022, 2023. https://www.fsa.go.jp/news/r4/sonota/20230217/01.pdf (accessed on 2025-10-30).

[8] 日本公認会計士協会. 監査上の主要な検討事項 (KAM) の適用３年目に関する周知文書. **監査基準報告書** *701* **周知文書第２号**, 2023. https://jicpa.or.jp/specialized_field/files/2-24-24-2-20230403.pdf (accessed on 2025-10-30).

[9] Yamin Zeng, Joseph Zhang, Junsheng Zhang, and Mengyu Zhang. Key audit matters reports in china: Their descriptions and implications of audit quality. *Accounting Horizons*, 35(2):167–192, 06 2021. doi: 10.2308/HORIZONS-19-189.

[10] Jenna J. Burke, Rani Hoitash, Udi Hoitash, and Summer (Xia) Xiao. The Disclosure and Consequences of U.S. Critical Audit Matters. *The Accounting Review*, 98(2):59–95, 03 2023. ISSN 0001-4826. doi: 10.2308/TAR-2021-0013. URL https://doi.org/10.2308/TAR-2021-0013.

[11] 佐々木 貴司. 銀行業の監査における貸倒引当金に関する KAM の類似性の分析. **経営分析研究**, (36):19–41, 2023.

[12] PwC Japan 有限責任監査法人. 監査品質に関する報告書 2024, 2024. https://www.pwc.com/jp/ja/about-us/member/assurance/assets/pdf/transparency-report-2024-01.pdf (accessed on 2025-10-30).

[13] 有限責任監査法人トーマツ. 監査品質に関する報告書 2025, 2025. https://www.

deloitte.com/content/dam/assets-zone1/jp/ja/docs/services/audit-assurance/2025/jp-aa-audit-quality-report-2025_a4.pdf (accessed on 2025-10-30).

[14] 有限責任あずさ監査法人. AZSA Quality 2025/26, 2025. https://assets.kpmg.com/content/dam/kpmg/jp/pdf/2025/jp-azsa-quality-2025.pdf.

[15] Mohd Shafiq Alias, Muhamad Haziq Fuad, Xavier Leong Foo Hoong, and Edward Goh Yoon Hin. Financial text categorisation with finbert on key audit matters. In *2023 IEEE Symposium on Computers & Informatics (ISCI)*, pages 63–69, 11 2023. doi: 10.1109/ISCI58771.2023.10391878.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. URL https://arxiv.org/abs/1810.04805.

[17] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. FinBERT: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020. URL https://arxiv.org/abs/2006.08097.

[18] 土井 惟成, 小田 悠介, 中久保 菜穂, and 杉本 淳. 大規模言語モデルを用いたゼロショットテキスト分類による TCFD 推奨開示項目の自動判定. *JPX* **ワーキング・ペーパー**, 43:1–30, 3 2024.

[19] Masahiro Suzuki, Hiroki Sakaji, Masanori Hirano, and Kiyoshi Izumi. Constructing and analyzing domain-specific language model for financial text mining. *Information Processing & Management*, 60 (2):103194, 2023. doi: 10.1016/j.ipm.2022.103194.

[20] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[21] OpenAI. Using GPT-5. https://platform.openai.com/docs/guides/latest-model (accessed on 2025-10-30).

[22] OpenAI. GPT-4o System Card, 2024. https://cdn.openai.com/gpt-4o-system-card.pdf (accessed on 2025-10-30).

[23] OpenAI. GPT-5 System Card, 2025. https://cdn.openai.com/gpt-5-system-card.pdf (accessed on 2025-10-30).

[24] Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. Zero-shot text classification using large language models for key audit matters in japanese audit reports. *International Journal of Smart Computing and Artificial Intelligence*, 9(1), 2025.

[25] Antti Rautiainen, Jani Saastamoinen, and Kati Pajunen. Do key audit matters (kams) matter? auditors' perceptions of kams and audit quality in finland. *Managerial Auditing Journal*, 36(3):386–404, 2021. doi: 10.1108/MAJ-11-2019-2462. URL https://doi.org/10.1108/MAJ-11-2019-2462.

[26] Derek K. Chan and Nanqin Liu. The Effects of Critical Audit Matter Disclosure on Audit Effort, Investor Scrutiny, and Investment Efficiency. *The Accounting Review*, 98(2):97–121, 03 2023. ISSN 0001-4826. doi: 10.2308/TAR-2020-0121. URL https://doi.org/10.2308/TAR-2020-0121.

[27] Muttanachai Suttipun. Impact of key audit matters (kams) reporting on audit quality: evidence from

thailand. *Journal of Applied Accounting Research*, 22(5):869–882, January 2021. ISSN 0967-5426. doi: 10.1108/JAAR-10-2020-0210. URL https://doi.org/10.1108/JAAR-10-2020-0210.

[28] Tobias Carlé, Nicolas Pappert, and Reiner Quick. Text similarity, boilerplates and their determinants in key audit matters disclosure. *Corporate Ownership and Control*, 20:49–62, 01 2023. doi: 10.22495/cocv20i2art4.

[29] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10(8):707–710, 02 1965. URL https://api.semanticscholar.org/CorpusID:60827152.

[30] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019. URL http://arxiv.org/abs/1904.09675.

[31] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In Mona Diab, Tim Baldwin, and Marco Baroni, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/S13-1004.

[32] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.317.

[33] Taku Kudou. Mecab: Yet another part-of-speech and morphological analyzer, 2005. https://taku910.github.io/mecab/ (accessed on 2025-10-30).

[34] The UniDic Consortium. Unidic for contemporary written japanese (unidic-cwj). https://clrd.ninjal.ac.jp/unidic/, 2023. Version 2023.03 (accessed on 2025-10-30).

[35] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 0013-1644. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104. Publisher: SAGE Publications Inc.

[36] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 12 2003.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 07 2002.

[38] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *2021 International Conference on Learning Represen-*

tations, May 2021. URL https://www.microsoft.com/en-us/research/publication/deberta-decoding-enhanced-bert-with-disentangled-attention-2/. Under review.

[39] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

[40] Tohoku NLP Group. BERT base Japanese (IPA dictionary) . https://huggingface.co/tohoku-nlp/bert-base-japanese (accessed on 2025-10-30).

[41] Language Media Processing Lab at Kyoto University. Model Card for Japanese DeBERTa V2 base . https://huggingface.co/ku-nlp/deberta-v2-base-japanese (accessed on 2025-10-30).

[42] Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. ModernBERT-Ja. https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a, 2025. URL https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a.

[43] Linda Elizabeth DeAngelo. Auditor size and audit quality. *Journal of Accounting and Economics*, 3(3):183–199, 1981. ISSN 0165-4101. doi: https://doi.org/10.1016/0165-4101(81)90002-1. URL https://www.sciencedirect.com/science/article/pii/0165410181900021.

[44] Jere R. Francis. What do we know about audit quality? *The British Accounting Review*, 36(4):345–368, 2004. ISSN 0890-8389. doi: https://doi.org/10.1016/j.bar.2004.09.003. URL https://www.sciencedirect.com/science/article/pii/S0890838904000769.

[45] Dan A. Simunic. The pricing of audit services: Theory and evidence. *Journal of Accounting Research*, 18(1):161–190, 1980. ISSN 00218456, 1475679X. URL http://www.jstor.org/stable/2490397.

[46] Jere R. Francis, Edward L. Maydew, and H. Charles Sparks. The role of big 6 auditors in the credible reporting of accruals. *AUDITING: A Journal of Practice & Theory*, 18(2):17–34, 09 1999. ISSN 0278-0380. doi: 10.2308/aud.1999.18.2.17. URL https://doi.org/10.2308/aud.1999.18.2.17.

[47] Ann L Watkins, William Hillison, and Susan E Morecroft. Audit quality: A synthesis of theory and empirical evidence. *Journal of accounting literature*, 23:153, 2004.

[48] Aloke Ghosh and Steven Lustgarten. Pricing of initial audit engagements by large and small audit firms. *Contemporary Accounting Research*, 23(2):333–368, 2006. doi: https://doi.org/10.1506/927U-JGJY-35TA-7NT1. URL https://onlinelibrary.wiley.com/doi/abs/10.1506/927U-JGJY-35TA-7NT1.

[49] Philip M. McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392, may 2010. ISSN 1554-3528. doi: 10.3758/BRM.42.2.381. URL https://doi.org/10.3758/BRM.42.2.381.

[50] Paul Baker. *Using Corpora in Discourse Analysis*. Bloomsbury Discourse. Bloomsbury Academic, 2006. ISBN 9780826477255.

[51] G. McKee, D. Malvern, and B. Richards. Measuring vocabulary diversity using dedicated software.

Literary & Linguistic Computing, 15:323–337, 2000. doi: 10.1093/llc/15.3.323.

[52] P. M. McCarthy and S. Jarvis. Vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4): 459–488, 2007. doi: 10.1177/0265532207080767.

[53] H. D. Maas. Zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 8:73–79, 1972.

[54] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. doi: https://doi.org/10.1002/asi.21001. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001.

[55] Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. BertAA : BERT fine-tuning for authorship attribution. In Pushpak Bhattacharyya, Dipti Misra Sharma, and Rajeev Sangal, editors, *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India, December 2020. NLP Association of India (NLPAI). URL https://aclanthology.org/2020.icon-main.16/.

[56] 土井 惟成. トピックモデルによる監査上の主要な検討事項（KAM）の類似性の検証. In **じんもんこん** *2022* **論文集**, volume 2022, pages 199–206. 情報処理学会, Dec 2022.

[57] 公認会計士・監査審査会. 令和５年版 モニタリングレポート, 2023. https://www.fsa.go.jp/cpaaob/shinsakensa/kouhyou/20230714/2023_monitoring_report.pdf(accessed on 2025-10-30).

[58] 公認会計士・監査審査会. 令和６年版 モニタリングレポート, 2024. https://www.fsa.go.jp/cpaaob/shinsakensa/kouhyou/20240719/2024_monitoring_report.pdf(accessed on 2025-10-30).

[59] Tohoku NLP Group. BERT base Japanese (unidic-lite with whole word masking, CC-100 and jawiki-20230102). https://huggingface.co/tohoku-nlp/bert-base-japanese-v3 (accessed on 2025-10-30).