2024 年 11 月 1 日　全 11 頁

金融経済分析を変える自然言語処理の力②

大規模言語モデルの登場と今後の展望

高い精度や幅広い応用性に魅力も全ての従来手法は代替されないか

経済調査部　研究員　島本 高志

［要約］

■ 前回のレポートでは、従来型の自然言語処理が金融経済分析にどのようなメリットをもたらし、どのようなテーマに適用されてきたのかを紹介した。本レポートでは、GPT（Generative Pre-trained Transformer）モデルの一種である GPT-4 などの大規模言語モデル（Large Language Model、以下 LLM）の登場が、金融経済分析にもたらした影響について考察する。

■ LLM の金融経済分析への応用例としては、多言語能力を活用した異なる言語のセンチメント分析や、文書分類の精度向上などが挙げられる。大和総研では、各職業のタスクのテキスト文を GPT-4 に分類させた結果を利用して、生成 AI が日本の労働市場に与える影響を分析した事例などがある。また、財務諸表分析で、アナリストによる分析や従来型の機械学習モデルを上回る精度で、財務指標の変動を予測できたとする研究等もある。

■ ただし、LLM には従来型の自然言語処理と比較して、複数の課題がある。例えば、モデルの内部機構や訓練データの詳細な理解が困難なブラックボックス性は、説明責任が重い公的セクターでの指数開発への応用などに難をもたらす（透明性の問題）。また、多くのモデルの開発が企業依存であることは、一度開発した指数等を継続的に利用できないリスクが存在する（継続性の問題）。出力結果が複雑な場合は、人間による解釈が困難な問題（解釈可能性の問題）も生じる。この問題では、従来型の自然言語処理の中でも、解釈が比較的容易な原始的な手法が優れる。さらに、時系列データの予測においては、現時点では従来の時系列分析モデルに軍配が上がる場合もある。

■ 結論として、LLM は分類タスクや多言語での強みが確認されるものの、透明性や解釈性が求められる分野や時系列分析などでは、従来型の自然言語処理や時系列分析手法が依然として重要である。もちろん、LLM が抱える一部の課題は時間とともに解消されるだろうが、現時点（2024 年 10 月）においては、LLM に過度な期待をかけるべきではないだろう。金融経済分析には、LLM と従来型の自然言語処理や時系列分析手法を、タスクの要件に応じて適切に使い分けることが求められる。

はじめに

前回のレポート（島本（2024）「金融経済分析を変える自然言語処理の力①　従来の自然言語処理が与えたインパクト」）では、従来型の自然言語処理の金融経済分析への応用事例として、景気や市場に関するセンチメントスコア（テキストデータから抽出した、人々のポジティブ・ネガティブな心理状況を指数化したもの）の作成など、どのようなテーマに対して適用されてきたかを概観した。さらに、自然言語処理が金融経済分析にもたらしたメリットも紹介した。

本レポートでは、GPT-4 など近年の自然言語処理の発展の牽引役である大規模言語モデル（Large Language Model、以下 LLM）を活用した金融経済分析の具体例をまず紹介し、次に従来型の自然言語処理や時系列分析手法と比較しながら、LLM の課題を示す。最後に、総論として評価を行い、今後の展望について議論しながら結論をまとめる。

LLM を活用した金融経済分析の具体例

まず、LLM が活用された金融経済分析の具体例を概観する（**図表１**）。この章では、前回のレポートと同様に分かりやすさを重視して、手法よりもテーマに焦点を当てながら紹介していく。

センチメントスコアの構築や多言語への横展開

前回のレポートでは、従来型の自然言語処理の応用事例として、人々の心理状況を指数化したセンチメントスコアを取りあげ、様々なセンチメントスコアが開発されてきたことを紹介した。一方で、LLM の応用事例でもこうした傾向は変わらず、例えば ChatGPT が登場して 1 年も経たないうちに書かれた Ardekani et al.（2023）では、GPT-3 を用いて、センチメントスコアの構築を試みた。Ardekani et al.（2023）は、経済分野におけるセンチメントスコアの構築について「急速に成長している研究分野[1]」と指摘しており、LLM を含む自然言語処理の金融経済分析への応用テーマとして、近年強い関心を呼んでいることがうかがえる。

具体的に Ardekani et al.（2023）では、New York Times 誌の「インフレ」という単語を含む約 2,200 本の記事を使用し[2]、先行研究に従って、インフレに関するセンチメントスコアを用意した。そして、OpenAI 社が開発した GPT-3 がベースの「ada」と呼ばれる埋め込みモデル（テキストの表現を数量化するためのモデル）に 2022 年 1 月から 6 月の指数をファインチューニング（モデルの追加学習による調整）させて、同年 7 月から 12 月の同誌の記事からインフレに関するセンチメントスコアを測定させている。

結果としては、従来型の自然言語処理によるインフレに関するセンチメントスコアと、ada によ

[1] Ardekani（2023）p.2, 翻訳は筆者による。
[2] 先行研究の Barbaglia et al.(2022)によると、この特定の単語を含む記事を使用する手法には、特定の経済概念に関連するセンチメントをより正確に計算できる利点がある。

図表１　LLM が応用された研究の事例

研究	主な手法	テーマと概要	主要な示唆	国・地域
Ardekani et al.（2023）	GPT-3	NYタイムズの約2,200本の記事をデータとしたセンチメントスコアの算出と、ECBの英語の公表文を基にした学習した、多言語での経済センチメントスコアの算出	既存の有力なセンチメントスコアであるBarbaglia et.al.(2022) と約0.61の相関、フランス語とスペイン語で約0.5の相関	米国・ヨーロッパ
土井ほか（2024）	GPT-3.5, GPT-4	気候関連財務情報開示タスクフォース（TCFD）によるTCFD推奨開示項目の、ゼロショット分類	GPT-4による正解率（Accuracy）は92.8%、F1-Scoreは84.9%。	日本
新田（2024）	GPT-4	日本版O-NET上の456職種のタスクから、生成AIが職業ごとのタスクの代替や日本の労働市場に与える影響を試算	高付加価値化が進む「協働グループ」と自動化が進む「代替グループ」の各就業者グループが全体に占める割合は、どちらも約20%程度	日本
Pelster and Val（2024）	GPT-4	GPTモデルが学習に使用していない時期の新情報を与えた際に、GPTが株式銘柄の業績予想や魅力度を評価可能かの試行および、EPSや株式リターン、シャープレシオ等との関係の分析	ネガティブな追加情報に基づき魅力度を調整可能で、業績予想はコンセンサスを制御したうえでEPSと有意に相関。魅力度と30日後リターンには、分析期間24日中7日目から正の相関	米国
Carriero et al.（2024）	時系列言語モデル	大規模言語モデルの応用である時系列基盤モデルによる、ゼロショットでのマクロ経済指標の分析と予測	安定性や精度において、BVARやDFMなど、既存のマクロ計量経済モデルを上回らない	米国
Kim et al.(2024)	GPT-4	Chain-of-Thoughtというプロンプトを使用した財務諸表分析。	EPSの「増加」「減少」の二項予測で人間のアナリスト・コンセンサスを上回る精度	米国

（出所）参考文献（11 ページ掲載）より大和総研作成

るセンチメントスコアの間には、約 0.61 の相関 [3]が確認された。両指数の間には、それなりの関連性が見られたといえる。

ただ、Ardekani et al.（2023）で興味深いのは、LLM の活用によって多言語でセンチメントスコアを作成した点である。具体的には、モデル（DaVinci と呼ばれる別の GPT-3 ベースの埋め込みモデル）をファインチューニングする際に、欧州中央銀行（European Central Bank，以下 ECB）の月次の金融政策決定公表文のうち、2017 年から 2021 年の「英語」の公表文を使用した。そして、2022 年分のセンチメントスコアを算出する際には、学習した「英語」の公表文だけでなく、内容は同じでも別言語による別の文章であるフランス語、ドイツ語、スペイン語、ポルトガル語の公表文も使って、それぞれの言語でセンチメントスコアの計算を行った。

結果としては、フランス語やポルトガル語では、先行研究から計算されたセンチメントスコアとの間に、約 0.5 の相関が観察された。相関係数は決して高くないものの、GPT-4o や OpenAI o1 などが登場した現在から見れば発展途上の手法だった GPT-3 を使用して、英語データの学習のみで約 0.5 の相関が出た点は特徴的だ。これは、自然言語処理の最新手法である LLM が、他言語からの学習内容を活用できる特性をよく捉えている。

上場企業の開示文書の情報を適切に分類

次に紹介したいのは、土井ほか（2024）だ。この論文では、企業が公表する気候変動関連の莫大な開示文書について、LLM を使って人手に頼らずに適切に分類して、開示情報が当局の推奨す

[3] ここでは一般的に使われる「ピアソンの積率相関係数」ではなく、比べる変数間の関係が線形ではなく非線形でもよい、各変数を順位に変換して求めた「スピアマンの順位相関係数」が使われている。

る項目を充足しているかどうかを判定している。この研究では、ファインチューニングを行わずに、未知のデータを分類する「ゼロショットテキスト分類」という手法を、GPT-3.5 と GPT-4 で実施している。

対象となったデータは、G20 の要請で金融安定理事会が設立した Task Force on Climate-related Financial Disclosures（気候関連財務情報開示タスクフォース、以下 TCFD）による提言（通称、TCFD 提言）に基づいて開示された、約 2,200 社による 13,500 件ほどの開示情報である。土井ほか（2024）では、TCFD による 4 種類の構成要素（「ガバナンス」「リスク管理」「戦略」「指標と目標」）内の細目である 11 種類の推奨開示項目を、これら企業の開示情報が実際に充足しているか調査している。その際、推奨開示項目をさらに分解した 27 個の「クライテリア」と呼ばれる詳細な評価基準を設定し、基準を満たしているのかに関して LLM を使って判定させた。モデルの精度に関しては、TCFD 提言の 4 種類の構成要素に基づいて検証した。検証には、開示情報全体から構成要素ごとに 100 件ずつ無作為に抽出したテキストデータを採用した。

その結果、GPT-3.5 では正解率（全体のうちの正答した割合）が約 59.3%だったのに対し、GPT-4 では約 92.8%と大幅に向上したことが報告されており、モデルの性能を評価するための他の評価指標（適合率、再現率、F1-Score）[4]でも、全体的に GPT-4 が優れていることが報告されている。そして、この GPT-4 を用いて、開示情報のうち 27 個の推奨項目（クライテリア）を満たしている割合がそれぞれ何%か計算し、さらにその平均を取った結果、充足率の平均は 32.6%と判断されている。

日本の職業を生成 AI から受ける影響を踏まえて三つの職業グループに分類

当社でも、LLM による文章の分類という視点から金融経済分析へ応用している。新田（2024）は米国の先行研究[5]などに基づき、日本の各職業に関して、仕事の特徴[6]と生成 AI から受ける影響を踏まえて三つの職業グループに分類した上で、それぞれ協働グループ[7]、代替グループ[8]、その他の職業グループ[9]と定義した。生成 AI から受ける影響を定量的に評価するため、職業情報提供サイト「job tag」（日本版 O-NET）[10]に含まれる各職業のタスクのテキスト文を GPT-4 に読み込ませ、タスク全体のうち生成 AI の影響を受ける割合を計算した。この結果、協働グループには弁護士や経営コンサルタント、AI エンジニアなどの職業、代替グループにはプログラマーや一般事務、データ入力などの職業、その他グループには大工、電車運転士、美容師などの職

[4] 3 種とも機械学習の文脈でよく使われる評価指標。この論文では、GPT-3.5 では 36.3%だった適合率が GPT-4 では 77.2%に上がった様子や、適合性と再現率から算出される F1-Score が、GPT-3.5 で 52.7%だったのに対して GPT-4 では 84.9%に上がった様子が紹介されている。ただ、再現率は両者ほぼ同じで 95%程度だった。
[5] Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2023) “GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models,” https://arxiv.org/abs/2303.10130v4
[6] 「情報やデータを分析する」「厳密さ、正確さ」「人間関係を構築し、維持する」「同一作業の反復」「結果、成果への責任」「手と腕を使って物を取り扱い動かす」など、各職業の特徴を示すデータを用いた。
[7] 生成 AI と協力することで、単純作業を自動化し、より高度な仕事に集中できる職業グループを指す。
[8] 生成 AI によって、仕事の大部分が自動化されてしまう可能性が高い職業グループを指す。
[9] 生成 AI の影響が比較的少ない、もしくは影響が不明確な職業グループを指す。
[10] 厚生労働省が提供している、職業に関する必要スキルや仕事でのタスクを集めたデータベース。

業が分類された。さらに、「協働グループ」あるいは「代替グループ」に分類された就業者の割合はともに 20%前後、産業別では金融業や情報通信業などへの影響が大きいなどの分析を行った。

新田（2024）では、タスクのテキスト分析に GPT-4 を採用した理由として、①人的バイアスや主観的解釈を排除し、客観性の確保を図ることが可能となる点、②GPT-4 の高度な言語処理能力を活用することで、精度の高い分析が実現できる点、③数千件規模におよぶ大量のテキスト分類作業を自動化することで、作業効率の大幅な向上が見込める点、を挙げている。

LLM による株式銘柄の評価・分析や財務諸表分析に基づく EPS の変動の予測

金融分野に大きく関わる話として、LLM が株式銘柄の分析等をできるかに関する研究も進んでいる。

Pelster and Val（2024）では、GPT-4 が株式銘柄の業績予想や魅力度を評価できるかおよび、追加の情報を与えた際に評価を調整可能かについて試行した。さらに、各銘柄の魅力度と EPS や株式のリターン、投資の効率性を測る指標であるシャープレシオ [11]との関係についても分析している。ここでの魅力度とは、投資家が今後 1 カ月間、ある株式銘柄を保有する際に、他の S&P500 構成銘柄との相対的な魅力を▲5～＋5 の範囲で評価した数値だ。また、分析に際しては、GPT-4 の学習データに含まれない 2023 年第 2 四半期のデータ [12]を利用し、市場やファンダメンタルズに関する情報ベンダーからのデータやウェブ検索で得られた当該銘柄に関するテキスト [13]も追加で与えた。

結果としては、株式銘柄のポジティブなニュースに関して GPT-4 による魅力度の調整は観察されなかったが、ネガティブなニュースを与えられた際は、GPT-4 は魅力度を大幅に調整していた。業績予想については、市場のアナリストたちによる業績予想の中央値（コンセンサス予想）の影響を制御した上で、EPS と有意に相関していた。加えて、株式銘柄の魅力度と 30 日後リターンの間およびシャープレシオとの間には、分析期間 24 日間のうち 7 日目から最終日まで正の相関が見られたと報告された。

また、Kim et al.（2024）では、モデルに複数のステップを順番に考えさせる手法である Chain-of-Thought（以下、CoT）を使用して財務諸表分析を行った。具体的には、人間のアナリストの推論を模して複数の段階的なステップに分けたプロンプト（命令文）を入力し、財務諸表の数字を GPT-4 に分析させることで翌年の EPS の変動（上昇と下落の二項選択）を予測させるとともに、その予測の確信度を「高い」「適切」「低い」で回答させている。

結果としては、EPS の変動に関して単純な将来収益予測を要求するプロンプトでの正答率は

[11] 投資における、株式のリスク（変動の標準偏差）に対するリターンの指標。
[12] GPT-4 は訓練データとして 2021 年 9 月以前の情報を「知っている」ため、その期間のデータに基づく分析では、既存の知識に基づくバイアスがかかるリスクがある。
[13] WebChatGPT というブラウザの外部拡張機能を使用している。この研究では WebChatGPT を用いて銘柄に関連する情報を取得しそうな検索語句を ChatGPT に策定させ、その語句で Yahoo 検索してトップ 10 に出てくる URL 内のテキスト情報を取得し、ChatGPT でサマリーを生成して、その内容をプロンプトへと追加した。

52.33%だったが、CoT による複数の推論に分けて行った予測では 60.31%の正答率が出たと報告されている。これは、人間のアナリストの正答率が前年度財務諸表発表後の1カ月後で52.71%、3 カ月後で 55.95%、6 カ月後で 56.58%だったことを上回っている。先行研究に倣った機械学習モデルの正答率 60.45%ともほぼ同水準であり、かつ F1-Score では、機械学習モデルの61.62%に対して、CoT 推論ではこれを上回る 63.45%のスコアが出たと報告された。

興味深いのは、人間のアナリスト予想との相補性が指摘されていることだ。GPT の予測とアナリスト予想はそれぞれ単独でも将来の EPS の変動と正の相関があるが、両者をともに EPS の変動へ回帰しても各係数はともに統計的に有意で、モデルの説明力を示す指標である修正決定係数[14]も上昇している。つまり、両者は実際の収益変動の異なる側面を予想できており、組み合わせることでお互いを補える可能性が高い。

全ての分析は LLM に一本化されるのか？

前章では LLM が金融経済分析の多様なタスクに応用され、優れた精度を示していることを紹介した。LLM がこれだけ優れているなら、一見、従来の自然言語処理手法を全て LLM に置き換えるべきとも思えるが、果たして本当にそうだろうか。

本章では、いくつかの観点から従来の手法と LLM を比較し、LLM が金融経済分析の全ての面で優れているわけではないことを示していく。

予測精度において必ずしも従来手法を上回らないケースもある

まずは、そもそも予測精度において従来手法を上回らないケースがあることを紹介したい。

金融分野での LLM 等に関するサーベイ論文である Lee et al. (2024) は、「エキスパートモデル」と呼ばれるような、金融向けに特化するように調整された従来モデル手法も含めた 5 種類 8 つのモデルを比較している。そのうえで、センチメント分析では、使用したデータセットにおける GPT-4 の評価指標（F1-Score）が約 87%だったのに対して、それ以前に発表された従来型モデル（2022 年発表の FLANG-ELECTRA というモデル）は約 92%と GPT-4 を上回ったことが報告されている。テキスト分類においても、上記の従来型モデル（FLANG-ELECTRA）による評価指標（F1-Score）の平均が約 98%だったのに対して、2023 年 6 月に発表された最新の LLM モデル（Fin-MA（30B[15])）の評価指標の平均も約 98%と、従来型モデルと同等だった旨が報告された。同様に、Guo et al. (2024) でも、特にセンチメントの分類や分析などで、2023 年までに発表されている従来型モデル（FinBERT や RoBERTa、FLANG-BERT といった従来型の BERT の延長にあるモデル）が、GPT-4 のスコアを上回っている事例が報告されている。

---

[14] 「決定係数」はモデルがデータをどのくらい説明できているかを示す指標で、値が 1 に近いほど説明力が高いことを示す。ただし、解析に使う変数を増やすだけで値が大きくなる傾向があるため、使用する変数の数へペナルティを与える修正を施したのが「修正決定係数」である。

[15] この 30B はパラメータ数が 30billion（300 億個）であることを指す。他にも 7B などのモデルが存在する。

この論文では、一般的な質疑応答での性能などでは GPT-4 が群を抜いていた旨が報告されているが、金融経済分析においては、様々なタスクへの対応を考えると、必ずしも GPT-4 などの LLM モデルが全ての領域で従来的な自然言語処理を上回っていないことが分かる。

ただし、これは英語による研究であり、自然言語処理には言語間の違いの問題が常に影響することから、必ずしも日本語の環境において同様の結果が完全に再現されるというわけではない。しかし、現時点では英語圏を中心とするグローバルな環境において、必ずしも LLM だけが主流というわけではないことは、今後の潮流を考えるにあたっても重要だ。

時系列データの厳密な数値予想は未だ苦手な可能性

次に、LLM が時系列データなどに対する厳密な数値予想を、未だ苦手とする可能性もある。

言語モデルにおける LLM の成功を見て、この仕組みを時系列モデルとして利用できないかと考えて開発された「時系列基盤モデル（Time Series Foundational Models, 別称「時系列言語モデル」）」というモデル群が存在する[16]。Carriero et al.（2024）は、最新の様々な時系列基盤モデルをマクロ経済指標の予測に活用できないかを試した論文だ。

データとして 1959 年から 2023 年までの米国の月次変数 120 系列を使用し、1985 年以降の経済指標について、前年までの数値に基づき、複数の時系列基盤モデル（Nixtla 社が開発した TimeGPT や、Salesforce 社が開発した Moirai など）を使用して予測した。その結果、時系列基盤モデルの予測値は、一部のモデルで誤差の中央値が従来型の時系列モデル（ベイジアン VAR モデルやダイナミックファクターモデルなど）に並んだが、分散（予測誤差のばらつき）の少なさでは従来型モデルに及ばなかった。要するに、現時点ではベイジアン VAR やファクターモデルのような従来型モデルの方が、安定的に予測できるという点で優れていることになる。

ただし、こうした評価の違いは、主に分析手法と問題設定の相違に起因する可能性がある。第一に、予測方法の違いがある。Carriero et al.（2024）は数値の正確な予測を試みる一方、Kim et al.（2024）は EPS の上昇・下落という二値カテゴリでの予測を行っている。第二に、使用する LLM の種類による差異がある。OpenAI の GPT シリーズと、時系列予測に特化した TimeGPT や Moirai ではモデルの構造が完全に同一ではない。これらを踏まえると、LLM は二値分類では優れた性能を示すものの、数値予測では課題が残る可能性や、特定の LLM のみが高い性能を持つ可能性が示唆される。現時点では、マクロ経済指標の数値予測への LLM の適用には慎重な姿勢が求められるのではないか。

LLM にたちはだかる透明性や継続性の問題

LLM の金融経済分析への応用で、より本質的な部分に関する重要な問題として透明性と継続性の問題が挙げられる。ここでいう透明性の問題とは、例えば、LLM などのモデルの内部のアーキ

[16] Carriero et al.（2024）では、「時系列基盤モデル」と「時系列言語モデル（Time Series Language Models）」の両呼称を紹介した上で、後者を好んで使用している。

テクチャ（機構）やパラメータ、訓練データなど、その動作原理や仕組みに関して必ずしも明らかになっておらず、モデルの挙動について外部から理解することが困難になる問題を指す。

この透明性の問題に関しては、Ash and Hansen（2024）でも、LLMをはじめとしたTransformerを応用したモデルに関して指摘されている。LLMでは数千億以上のパラメータが用いられることも珍しくないゆえに、モデルの学習には莫大な数のGraphics Processing Unit（以下、GPU）などのハードウェアリソースが必要になる。しかし、そのようなモデルを開発する余裕は大規模な組織にしかなく、「完全な推定の流れをレプリケーション（論文等を同じデータや手法を用いて再現）することは不可能[17]」とする。こうした場合は、ChatGPTなどのウェブサイトやAPI（Application Programming Interface、アプリケーションを他のアプリケーションから呼び出すインターフェースの仕様）などを通じて大規模な組織が開発したモデルを利用する形となる。ところが、例えばOpenAI社のGPTシリーズがGPT-3.5以降は非公開化されブラックボックスとなるなど、透明性の問題を抱えるケースも多い。透明性の問題があることは再現性の問題にも関係することから、決して好ましいとは言えない。

さらに、新谷（2023）は継続性の問題について、「LLM等の最新言語モデルは進化の速度も速く近年益々複雑化されておりモデルが変更された場合の過去系列の遡及推計も含めて、指数の継続性は大きな課題であるといえよう」と指摘する。LLMはモデルのアップデートや提供企業の事情により、長期的な利用が保証されない可能性があるという継続性の課題も抱えている。

こうした透明性のなさや、継続性が担保されない点は、透明性や継続性が求められる行政セクターが指数開発等を行う際には問題になりかねない。新たな指数の開発等は、しばしば政府部門や中央銀行などの行政セクターによって行われる。LLMのようにどのように指数が開発されたのかがブラックボックスとなっていると、指数自体の透明性に問題が生じる。これは、国民に対する説明責任を負っていて透明性を重視せざるを得ない行政セクターでは、特に重たい問題となる。また、モデルを保有する企業が提供を終了してしまう可能性などを考えると、官民ともに継続性の観点からは厳しい側面もある。

もちろん、技術内容がオープンソースとして公開されている「オープンソースLLM」、パソコンやスマートフォンのローカル環境で実行できる「ローカルLLM」を用いるという選択肢もある。しかし現在、最先端のLLMは、一般的にクローズドかつオンラインを通じてのみ利用可能であるため、オープンソースあるいはローカルLLMは性能が制限されることが多い。

従って、もし最先端のLLMを活用して指数を継続的に作成するのであれば、例えば、それらの指数へ組み込む際に根幹を担わせるのではなく、あくまで一部について競合するモデルや後発モデルへの置き換えが可能な形での活用で留めるなどの工夫が必要だ。

このように、LLMの金融経済分野への応用では透明性や継続性の観点から課題が多いことから、手元でモデルを構築し、自己の計算資源で計算し続けることが可能な点を考慮すると、従来的な手法に軍配が上がる場合もある。

[17] Ash and Hansen（2024）p.667，翻訳は筆者による。

複雑な手法に潜む解釈可能性と予測精度のトレードオフ

さらに、出力結果の解釈可能性における問題が挙げられる。解釈可能性とは、モデルから何故そのような予測や回答が作られたのかということに関して、どのような要因がどの程度寄与したのかといった影響や、モデル自体の仕組みを理解できるか、ということに関するレベルを指す。特に、内部機構の理解に関する部分については、先述の透明性とも関連が深い。

Ash and Hansen（2024）は、予測精度と解釈性のトレードオフについて詳細に言及している。これは、複雑なモデルからの出力結果は、一般に、人間による直接的な解釈が難しいことに関するものだ。例えば、従来的な自然言語処理モデルの中でも後期のモデルにあたる、文脈も織り込んでテキストを数量化するようなモデルが、その代表例だ。新谷（2023）では、「純粋な予測精度向上の観点からは、文脈を含めたテキスト情報を有効に反映できる機械学習アプローチが望ましい」とする一方、辞書的なアプローチによる古典的なセンチメント分析に関して、機械学習的なアプローチよりも「経済学的な解釈の容易性の観点から、現在でも十分有用性が高い」としている[18]。そのうえで、五島ほか（2022）によって作成された「景気単語極性辞書」を使用した結果が、新型コロナウイルスの感染拡大期の景気センチメントをよく捉えていることを示している。既定の辞書に沿っていて基準が先決かつ明確で、どの単語がポジティブ・ネガティブに寄与しているかも明瞭という点で、よりシンプルな辞書的なアプローチは解釈性に優れている。また、内部構造が複雑な LLM についても、対話型のモデルを中心に、何故その数値を出したかを説明できるというメリットはあるものの、他の統計分析に組み込むためにテキストを数量化する際などには、解釈が難しいケースが依然として存在する。

つまり、機械学習モデルに比べて、先に決められた明瞭なルールに基づき、一目でデータを解釈できるという解釈可能性の点で、やはり従来型の辞書アプローチに軍配が上がる。LLM では改善されているものの、解釈可能性の面での課題はなくなっていない。この点からも、全てが LLM に置き換えられるとは考え難い。

おわりに

これまで 2 本のレポートを通じて、自然言語処理を用いた分析のメリットや、従来型の自然言語処理による分析や、最新型の LLM を用いた分析の例について提示してきた（**図表２**）。

結果としては、特に分類タスクや多言語能力の活用に LLM モデルが優位である点が明らかになった。さらに、両者は実際の収益変動の異なる側面を予想できている可能性が高いなど内容的には補完的でありつつ、精度でアナリストを上回る可能性が示された。だが、その一方で、透明性や継続性の問題、あるいは複雑すぎる手法における解釈可能性の問題などに関しても明らかになった。

[18] 双方ともに、新谷（2023）p. 128 より引用。

急速に活用が進んだとはいえ、ChatGPT が登場してからはまだ約 1 年 11 カ月だ。当時から現在まで大きな注目を浴び続けているとはいえ、LLM モデル自体の発展も活用も、未だ途上にある。

これから先の可能性を考えれば、例えば解釈性や透明性を強く要求しない分類タスクや、テキストを数量化した上での将来予測、多言語間での分析等に際しては、すでに始まっている従来型の自然言語処理モデルから LLM への置き換えが、これから益々加速していく可能性が指摘される。しかし、例えば透明性と継続性が強く求められるような分析や、解釈可能性を要求する文脈では、従来型の自然言語処理からの代替が進むとは考え難い。

また、未だに発展途上にある時系列分析に対する LLM の適用に関しても、応用研究等はますます盛んになっていくと予想される。だが、少なくとも今後数年は、従来型の時系列分析のパフォーマンスを上回るのは困難だろう。さらに、従来型の時系列分析の精度を上回ったとしても、先の解釈性の問題などから、依然として従来型の時系列モデルが優先され続ける可能性は高い。

結局は、LLM の革新的な性能に過度な期待を寄せるのではなく、分析の目的や要件に合わせて、従来の手法と LLM の特性を活かしながら適切に使い分けることが、金融経済分析においても求められるだろう。

以上

図表２　金融経済分析の力①②　両レポートのまとめ

前回レポート

- テキスト分析が金融経済分析へもたらすメリットの例
  - 「効率化」「精度向上」「新しい洞察」「適時分析によるリスク管理」「カスタマイズ」の５つの利点
  - 定量化手法としての「文書間の類似性」や「テキスト中の経済的概念」「テキスト中の概念の関連性」「テキストと定量的メタデータの関連付け」等の４つの測定
- 従来型の自然言語処理による分析テーマ例
  - 中央銀行の声明や経営者の将来見通しの、トーン分析による数量的解析
  - 新聞記事のデータを用いた、月次指標の日次代替指標の開発
  - 極性辞書による、人々の景気への考え等を数量化した、センチメント指数の作成

今回レポート

- ＬＬＭによる分析テーマ例
  - 多言語モデルによる、別言語での学習を反映した、センチメント指数の作成
  - 気候変動関連の莫大な量の情報開示文書を、人手に頼らず適切に分類
  - 財務諸表分析等に関する、内容的にはアナリストと補完的で、時に高精度な分析
- 今後の展望
  - 発展途上のＬＬＭにはまだ、精度面や、時系列データへの弱さに関しての不安も
  - 複雑なモデルであるＬＬＭには、透明性や継続性、解釈可能性の問題などもあり、上記の面においては、従来的なモデルや原始的モデルの方が優れることも多い
  - 分類タスクや多言語モデルにおいて優れ、今後の更なる発展が期待されるＬＬＭだが、あくまで分析の目的や要件に合わせた従来手法との使い分けが重要か

（出所）大和総研作成

参考文献

Ardekani, A.M., J. Berts, M.M. Dowling, and S. Long. (2023) “EconSentGPT: A Universal Economic Sentiment Engine?” https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4405779

Ash, E. and S. Hansen (2023) “Text Algorithms in Economics,” Annual Review of Economics, 15, pp.659-688

Barbaglia, L., S.Consoli, and S.Manzan (2022) “Forecasting with Economic News,” Journal of Business & Economic Statistics, 41(3), pp.708-719

Carriero, A., D. Pettenuzzo, and S. Shekhar (2024) “Macroeconomic Forecasting with Large Language Models,” https://arxiv.org/pdf/2407.00890

Guo, Y., Z. Xu, and Y. Yang (2023) “Is ChatGPT a Financial Expert? Evaluating Language Models on Financial Natural Language Processing,” https://arxiv.org/pdf/2310.12664

Kim, A.G., M. Muhn, and V.V. Nikolaev (2024) “Financial Statement Analysis with Large Language Models,” https://arxiv.org/pdf/2407.17866

Korinek, A. (2023) “Generative AI for Economic Research: Use Cases and Implications for Economists,” Journal of Economic Literature, 61(4), pp.1281-1317

Lee, J., N. Stevens, S.C. Han, and M. Song (2024) “A Survey of Large Language Models in Finance (FinLLMs),” https://arxiv.org/pdf/2402.02315

Pelster, M. and J. Val (2024) “Can ChatGPT assist in picking stocks?” Finance Research Letters, 59, 104786

五島圭一・新谷元嗣・高村大也（2022）「景気単語極性辞書の構築とその応用」,『自然言語処理』, Vol.29 No.4, pp.1233-1253, 言語処理学会

新谷元嗣（2023）「テキスト情報と機械学習を用いた景気動向分析」,『経済分析』, 第208号, pp.128-145, 内閣府経済社会総合研究所

土井惟成, 小田悠介, 中久保菜穂, 杉本淳（2024）「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, Vol.43, 日本取引所グループ, https://www.jpx.co.jp/corporate/research-study/working-paper/JPXWP_Vol43.pdf

新田尭之（2024）「生成ＡＩが描く日本の職業の明暗とその対応策～ＡＩと職業情報を活用した独自のビッグデータ分析～」,『大和総研調査季報』2024年春季号（Vol.54）, pp.38-71, 大和総研 https://www.dir.co.jp/report/research/economics/japan/20240425_030145.html