ローカル LLM を用いた利益予測の分析とアウトオブサンプル評価

白井 祐典 [1] 市川 佳彦 [1] 中川 慧 [2]
[1] 株式会社 Insight Edge [2] 大阪公立大学
{yusuke.shirai,yoshihiko.ichikawa}@insightedge.jp kei.nak.0315@gmail.com

概要

本論文は，日本株式市場における企業利益の増減方向予測に大規模言語モデル (LLMs) を適用し，とくに金融機関の実務利用を想定し，ローカル LLM を用いた場合の性能を検証する．具体的には，金融庁 EDINET に基づき構築されたベンチマークデータセット EDINET-BENCH を用いて，ローカル LLM による利益変化の方向予測性能を評価する．まず，モデルの事前学習カットオフと有価証券報告書の開示時期を切り分けることで，インサンプルとアウトオブサンプルを区別し，ローカル LLM の真の汎化性能を測定する．次に，業種や企業規模といった企業特性に応じた予測精度の違いを分析し，予測が得意な企業・苦手な企業のパターンを抽出する．

1 はじめに

企業の利益予測は，経営者による戦略的意思決定を支援するだけでなく，投資家など外部ステークホルダーにとっても中心的な関心事である．利益予測研究は，財務諸表の数値データを用いた統計的または機械学習手法による予測アプローチが発展してきた [1, 2, 3, 4, 5]．日本企業データを対象としても，同様の枠組みで利益変化を分類する研究が現れつつある [6]．しかし，これらの手法は主として数値情報に依拠しており，経営者の将来見通しやリスク認識，事業環境の質的変化といった定性的情報を十分に活用できない [7, 8]．

このような課題に対し，大規模言語モデル（large language models; LLM）は，テキストを中心とする非構造データを統合的に解釈し，推論する能力を有する点から，利益予測への応用可能性が注目されている [9, 10]．実際，[11] は米国企業の財務諸表および注記，MD&A テキストを対象に GPT-4 を用いて，Chain-of-Thought を組み合わせることで，プロのアナリストに匹敵あるいは上回る予測精度が得られる可能性を報告している．日本市場でも，決算短信や有価証券報告書に対して LLMs を適用する試みが増えつつある [6, 12]．

他方で，LLMs の有用性と同時に，バイアスや汎化性能に関する懸念も指摘されている [13, 14]．これらの結果は，LLM による利益予測の性能が企業特性に依存し得ること，すなわちどのような企業に対して予測しやすい（あるいは予測しにくい）のかという異質性を評価する必要性を示唆している．

この点，金融庁 EDINET に提出される有価証券報告書を基盤として，利益変化，会計不正，業種分類の 3 タスクを含む EDINET-BENCH が提案され，公開されている [15]．EDINET-BENCH では，財務諸表の数値特徴量とテキスト情報を組み合わせた入力に対して，複数の LLM をゼロショットで評価し，ロジスティック回帰などの従来モデルを上回る性能が報告されている [15]．

しかし，既存の LLM の利益予測は，主として GPT-4 や Claude 等のクラウド API 型による LLMs が中心であり，ローカル LLM（自社環境にデプロイ可能なオープンウェイトモデル）については，体系的な検証がほとんど行われていない．金融機関では，機密性の高い資料や顧客データを扱うことから，データの外部送信に伴う情報漏洩リスクや，モデル更新がベンダ側の裁量に依存する点[1)]など，クラウド API 型 LLM の利用には慎重な姿勢が求められる [9]．そのため，オンプレミスもしくは閉域ネットワーク上で推論可能なローカル LLM に対するニーズが高まっており，日本語金融テキストに特化した事前学習モデルの構築や評価も進められている [16]．

加えて，LLM の学習過程において Web 上の公開情報が広範に利用されていることから，テストデー

1) https://www.fsa.go.jp/common/law/cybersecurity_guideline_en.pdf

タとして用いられる有価証券報告書や企業情報が事前学習の段階ですでにモデルに取り込まれている可能性があり，見かけ上の高い性能が，真のアウトオブサンプル予測能力を反映していないおそれがある．この点は，ローカル LLM においても同様であり，モデルの事前学習カットオフとテストデータの開示時期を明示的に切り分けた上で，インサンプルとアウトオブサンプルの性能を区別して評価することが重要である [11, 6, 12].

以上を踏まえ，本研究では日本株式市場を対象とした利益予測において，ローカル LLM がどの程度の精度で利益変化の方向を予測できるのかを，EDINET-BENCH 上およびそのアウトオブサンプルデータで明らかにする．次に，業種や企業規模といった企業特性によってローカル LLM の利益予測精度がどのように異なるかを分析し，予測が得意な企業や苦手な企業のパターンを把握する．

2 問題設定

本研究では，[11, 6, 15] の利益予測問題に取り組む．具体的には，企業 $i \in \mathcal{I}$ の会計年度 $t-1$ における公開情報から，次期の純利益が増加するかを二値分類する．

ラベル $y_{i,t}$ は会計年度 $t-1$ および次年度 t 利益に対し，$y_{i,t} := \mathbf{1}\{E_{i,t} - E_{i,t-1} > 0\} \in \{0,1\}$ と定義される．入力は，$t-1$ 期の有価証券報告書から得られる集合 $\mathcal{F}_{i,t-1} = \left(X^{\text{num}}_{i,t-1}, S^{\text{text}}_{i,t-1}\right)$, であり，$X^{\text{num}}_{i,t-1} \in \mathbb{R}^p$ は損益計算書 (PL)，貸借対照表 (BS)，キャッシュフロー計算書 (CF) などの表形式ブロックから抽出した特徴量と，$S^{\text{text}}_{i,t-1} \in \mathcal{S}$ は非構造テキストである．

LLMs は特定の日付までに学習された内容を記憶しうるため，学習時点で利用したデータの最終時点をカットオフ m_{cut} とする．

この時，m_{cut} の LLMs $f_{\text{LLM},\, m_{\text{cut}}}$ を用いて，予測確率の計算を $\hat{y}_{i,t}$ を $\hat{y}_{i,t} = f_{\text{LLM},\, m_{\text{cut}}}(\mathcal{F}_{i,t-1}) \in [0,1]$ と行う．

評価対象の組 $(i,t) \in \mathcal{D}$ の提出日を $d_{i,t}$ とすると，本研究のインサンプルデータ IS およびアウトオブサンプルデータ OOS は次で定義する．

$$\text{IS} := \left\{(i,t) \in \mathcal{D} :\ d_{i,t} \le m_{\text{cut}}\right\}, \tag{1}$$

$$\text{OOS} := \left\{(i,t) \in \mathcal{D} :\ d_{i,t} > m_{\text{cut}}\right\} \tag{2}$$

本研究では，アウトオブサンプル性を予測対象 t の実現情報が事前学習に含まれ得るかで定義するため，区切りには $d_{i,t}$ を用いる．

3 実験

3.1 実験方法

EDINET-BENCH の利益予測に対する，ローカル LLM によるアウトオブサンプル評価および企業特性把握を，以下の手順で実施する．まず，検証の対象とする LLMs のカットオフ日 m_{cut} を調査し，インサンプルとアウトオブサンプルの対象期間を設定する．そして，インサンプルデータとアウトオブサンプルデータの両者での ROC-AUC を比較することで，アウトオブサンプルに対する予測の特徴を確認する．比較は，全体・売上規模別・業種別に実施する．なお，業種は，東証の 17 業種分類を使用する[2)]．これは，東証 33 業種分類を使用した場合にはデータ数がほとんど存在しない業種が存在するためである．

3.2 対象ローカル LLM とカットオフ日

本研究では，EDINET-BENCH [15] において最高精度を示した Claude 3.7 Sonnet をベンチマークモデルとして採用するとともに，金融機関での実務利用を想定し，ローカル環境で推論可能なオープンウェイト LLM を複数選定して評価を行う．具体的には，OpenAI が公開した GPT-OSS-20B および GPT-OSS-120B，ならびに Qwen 系列の Qwen-3-Next-80B-Instruct(以下，Qwen3-Instruct) と Qwen-3-Next-80B-Thinking(以下，Qwen3-Thinking) を対象とした．

これらのモデルを選定した理由は，以下の二点である．第一に，パラメータ規模および推論能力の多様性を確保するためである．GPT-OSS-20B は比較的軽量でありオンプレミス環境での運用を想定しやすい．一方，GPT-OSS-120B および Qwen3 系列はより大規模で高い表現能力を有する．第二に，日本語に対する適応性である．GPT-OSS 系列は，日本でも幅広く利用される GPT 系モデルの設計思想を踏まえて作られており，企業での実利用に適していると考えられる．また，Qwen 系列は多言語データを用いた事前学習および指示追従（Instruction tuning[17]）が施されており，日本語テキストに対する生成・推論性能が高いことが期待できる．

各モデルの事前学習データのカットオフ日 m_{cut}

2) https://www.jpx.co.jp/markets/indices/line-up/files/fac_13_sector.pdf

は，公式ドキュメントおよび公開情報に基づいて調査した[3)4)5)]．ただし，Qwen3 系は著者らの調査した範囲には公式情報がなく，推定される日付である．

表 1　検証対象の LLMs のカットオフ日

モデル名	カットオフ日
Claude 3.7 Sonnet	2024/10/31
GPT-OSS 20B	2024/6/1
GPT-OSS 120B	2024/6/1
Qwen3-Instruct	2024/12/31
Qwen3-Thinking	2024/12/31

表 1 のカットオフ日を考慮し，EDINET に有価証券報告書が提出された期間をそれぞれ，インサンプルを 2023/6/1〜2023/10/31，アウトオブサンプルを 2025/6/1〜2025/10/31 と定めた．なお，アウトオブサンプル期間に対するカットオフ日のロバストネス検証を付録 A にて実施し，検証している．

3.3　インサンプルとアウトオブサンプルの精度比較

インサンプルとアウトオブサンプルの比較は，次の 3 種類で実施する．1) 全体での比較．2) 売上実績の 4 分位群での比較．3) 17 業種での比較．3) では，インサンプルとアウトオブサンプルの両方で，件数が 100 件以上の業種だけを利用した．

1) 全体の比較　表 2 に，インサンプルとアウトオブサンプルを含むデータ全体での精度を示す．

表 2　各 LLM の ROC-AUC

	IS		OOS	
モデル名	件数	ROC-AUC	件数	ROC-AUC
Claude 3.7 Sonnet	2277	0.604	2369	0.603
GPT-OSS-20B	2268	0.499	2334	0.487
GPT-OSS-120B	2304	0.509	2404	0.521
Qwen3-Instruct	2316	0.474	2415	0.457
Qwen3-Thinking	1198	0.506	1359	0.500

インサンプルとアウトオブサンプルの ROC-AUC に注目すると，各モデルでアウトオブサンプルでの大きな精度劣化は見られなかった．これは，ナレッジカットオフ後に公開された利益の推論においても，ある程度のロバスト性を持った推論力があることを示唆している．モデル別にみると，Claude 3.7 Sonnet が最も精度が高く，ローカル LLM は精度が劣る結果となった．本実験においては，ローカル LLM は全体としてランダム予測に近い水準にとどまり，クラウド API 型 LLM である Claude 3.7 Sonnet と比較して明確に低い ROC-AUC を示した．

なお件数列は，プロンプトで指定した出力形式の結果が得られた数を示しており，Qwen3-Thinking は期待する出力形式のレスポンスが得られる件数が少なく，プロンプトの改良が必要であることが分かった．

この結果から，以降の実験では，ローカル LLM の中で最も精度が高かった GPT-OSS-120B を Claude 3.7 Sonnet と比較する．

2) 売上規模別の比較　表 3 に，売上規模の四分位群ごとの ROC-AUC を示す．売上は，EDINET の当期の売上高を利用し，取得できなかったデータについては対象外としている．売上の四分位は，各企業の売上高を上から順に上位 25%・25〜50%・50〜75%・75%以下で分類した．

表 3　売上規模別の ROC-AUC

	Claude 3.7 Sonnet		GPT-OSS-120B	
売上規模	IS	OOS	IS	OOS
上位 25%	0.602	0.597	0.492	0.547
25%-50%	0.608	0.626	0.522	0.505
50%-75%	0.623	0.602	0.535	0.490
75%以下	0.580	0.581	0.481	0.481

表 3 から，全ての場合において，売上が 75%以下の企業における精度が最も低かった．これは，売上規模が低い企業の場合，LLMs の学習に利用する Web 等のテキスト情報が相対的に少なく，その結果，LLMs を用いた予測が比較的難しいことが示唆される．

モデル別に見ると，Claude 3.7 Sonnet の結果は売上規模別の精度のばらつきが小さいのに対し，GPT-OSS-120B のアウトオブサンプルの結果は上位 25%の ROC-AUC が他のグループに比べて精度差が大きく，良い精度が観測された．

3) 業種別の ROC-AUC　表 4 に，業種別の ROC-AUC を示す．実験対象にした業種は，前述した通り，東証 17 業種分類の業種のうち予測対象件数が 100 件以上の 8 業種を本実験の対象とした．

表 4 から，インサンプルとアウトオブサンプルの両方で精度のばらつきが大きいことが見て取れる．

業種別に見ると，次の 3 グループに分けられる．

1) アウトオブサンプルでも精度が落ちづらい
2) アウトオブサンプルで精度が落ちる
3) アウトオブサンプルで GPT-OSS-120B が Claude 3.7 Sonnet よりも精度が高い

3) https://platform.openai.com/docs/models/gpt-oss-20b
4) https://platform.openai.com/docs/models/gpt-oss-120b
5) https://explodingtopics.com/blog/list-of-llms

表 4　業種別の ROC-AUC

	Claude 3.7 Sonnet				GPT-OSS-120B			
	IS		OOS		IS		OOS	
業種	件数	ROC-AUC	件数	ROC-AUC	件数	ROC-AUC	件数	ROC-AUC
機械	144	0.592	143	0.672	144	0.516	144	0.507
電気機器・精密機器	207	0.592	209	0.658	209	0.546	210	0.531
原材料・化学	185	0.611	192	0.598	191	0.446	192	0.539
IT・サービス他	526	0.585	573	0.606	531	0.474	578	0.512
建設・資材	211	0.547	218	0.619	215	0.461	221	0.541
小売	134	0.668	138	0.544	134	0.588	141	0.485
交通・物流	133	0.689	133	0.515	135	0.599	134	0.417
商社・卸売	224	0.621	225	0.585	225	0.468	232	0.605

1) は，インサンプルとアウトオブサンプルの ROC-AUC を比較した時に 0.05 以上の下落がなく，比較的ロバストに予測が可能な業種群である．具体的には，「機械」，「電気機器・精密機器」，「原材料・化学」，「IT・サービス他」，「建設・資材」である．

2) は，アウトオブサンプルの精度がインサンプルに比べて ROC-AUC が 0.1 以上低下しており，明確にアウトオブサンプルの予測が不得意な業種群である．具体的には，「小売」，「交通・物流」である．

3) は，GPT-OSS においてアウトオブサンプルの精度が向上している業種である．具体的には，「商社・卸売」である．

業種差の要因分析：利益増減の連続性　2) および 3) の対象業種の精度を利益の増減パターン（表 5）から見ると，「商社・卸売」は反転（57.1%）が最大であり，前年差の符号が切り替わる局面が相対的に多い．反転局面では，単純な前年差の自己相関（“前年と同方向”）だけでは当期の増減を当てにくく，業績の転換点を説明する定性的記述（需要環境，市況，構造改革など）が予測に与える情報価値が大きくなり得る．よって，「商社・卸売」における GPT-OSS-120B のアウトオブサンプル改善は，財務数値とテキストを組み合わせた予測である本研究の設定が相対的に機能した可能性を示唆する．

この仮説を定量化するため，GPT-OSS-120B の業種別性能差分 $\Delta\mathrm{AUC} := \mathrm{AUC}_{\mathrm{OOS}} - \mathrm{AUC}_{\mathrm{IS}}$ と反転率には正の関連が観測できた（Pearson $r = 0.746$, $n = 8$）．

一方，「交通・物流」は持続（56.3%）が最大であるにもかかわらず，アウトオブサンプル精度が大きく低下しており，前年差の持続性だけでは汎化性能を説明できない．この点は，外生ショック（燃料・運賃・規制・需給制約等）により，テキスト表現や会計数値と利益変化の対応関係が期間で変化し，学習済みの推論パターンがアウトオブサンプルで崩れる，といった構造変化の可能性を含意する．

表 5　OOS における業種別の増減パターン構成比

業種	増 → 増	増 → 減	減 → 増	減 → 減	持続	反転
機械	36.1	25.7	22.9	15.3	51.4	48.6
電気機器・精密機器	33.0	24.5	26.9	15.6	48.6	51.4
原材料・化学	33.0	28.4	22.2	16.5	49.5	50.6
IT・サービス他	35.3	26.5	21.1	17.1	52.4	47.6
建設・資材	40.0	28.2	24.5	7.3	47.3	52.7
小売	35.5	31.9	19.1	13.5	49.0	51.0
交通・物流	45.2	16.3	27.4	11.1	56.3	43.7
商社・卸売	28.1	31.6	25.5	14.7	42.8	57.1

注：持続は（増 → 増）と（減 → 減）の合計を表し，反転は（増 → 減）と（減 → 増）の合計を表す．

4　まとめと今後の課題

本研究は，EDINET-BENCH を用いてローカル LLM の利益増減方向予測をアウトオブサンプルで評価し，平均的には大きな精度劣化が観測されないことを確認した．一方で，ローカル LLM の ROC-AUC はクラウド API 型 LLM（Claude 3.7 Sonnet）を下回り，企業規模や業種により汎化性能が大きく異なることが分かった．特に，「小売」・「交通・物流」では OOS で精度低下が顕著である一方，「商社・卸売」では GPT-OSS-120B が改善した．これらは，表 5 が示す利益の持続・反転の違いを通じて，転換局面の言語情報の有効性や，外生ショックによる構造変化の影響があることを示唆している．

また，実利用を考えると，ローカル LLM における予測精度は低く，全体的な精度向上が必須であると考えられる．そのため，今後はより大規模なモデルでの実験や，得意・苦手パターンを考慮した LLM への追加情報の探索をしていきたい．

参考文献

[1] Jane A. Ou and Stephen H. Penman. Accounting measurement, price-earnings ratio, and the information content of security prices. **Journal of Accounting Research**, Vol. 27, pp. 111–144, 1989.

[2] Xi Chen, Yang Ha Cho, Yiwei Dou, and Baruch Lev. Predicting future earnings changes using machine learning and detailed financial data. **Journal of Accounting Research**, Vol. 60, No. 2, pp. 467–515, 2022.

[3] Joshua O. S. Hunt, James N. Myers, and Linda A. Myers. Improving earnings predictions and abnormal returns with machine learning. **Accounting Horizons**, Vol. 36, No. 1, pp. 131–149, 2022.

[4] Ahmad Hammami and Mohammad Hendijani Zadeh. Predicting earnings management through machine learning ensemble classifiers. **Journal of Forecasting**, Vol. 41, No. 8, pp. 1639–1660, 2022.

[5] Xi Chen, Yang Ha Cho, Yiwei Dou, and Baruch Lev. Predicting future earnings changes using machine learning and detailed financial data. **Journal of Accounting Research**, Vol. 60, No. 2, pp. 467–515, 2022.

[6] 屋嘉比潔, 黒木裕鷹, 中川慧. 日本企業データを用いた機械学習による利益変化の予測. 人工知能学会第二種研究会資料, Vol. 2024, No. FIN-033, pp. 68–75, 2024.

[7] Colm Kearney and Sha Liu. Textual sentiment in finance: A survey of methods and models. **International Review of Financial Analysis**, Vol. 33, pp. 171–185, 2014.

[8] Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. **Journal of accounting research**, Vol. 54, No. 4, pp. 1187–1230, 2016.

[9] 中川慧, 平野正徳, 高野海斗. 本邦金融分野における大規模言語モデルに関するサーベイと展望. 2025.

[10] Tingbang Yang, Yulong Du, Sihan Li, and Xiaobo Wu. Text meets numbers: Improving corporate profitability forecasting with llms. **Available at SSRN 5906316**.

[11] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models. **arXiv preprint arXiv:2407.17866**, 2024.

[12] 白井祐典, 市川佳彦, 中川慧. Llms による利益予測の分析とアウトオブサンプル評価. 人工知能学会第二種研究会資料, Vol. 2025, No. FIN-035, pp. 62–67, 2025.

[13] Kei Nakagawa, Masanori Hirano, and Yugo Fujimoto. Evaluating company-specific biases in financial sentiment analysis using large language models. In **2024 IEEE International Conference on Big Data (BigData)**, pp. 6614–6623. IEEE, 2024.

[14] Sebastian Jaskowski Wei Xu Sudheer Chava Agam ShahB, Liqin YeB. Beyond the reported cutoff: Where large language models fall short on financial knowledge. **arXiv preprint arXiv:2504.00042**, 2025.

[15] Issa Sugiura, Takashi Ishida, Taro Makino, Chieko Tazuke, Takanori Nakagawa, Kosuke Nakago, and David Ha. Edinet-bench: Evaluating llms on complex financial tasks using japanese financial statements. **arXiv preprint arXiv:2506.08762**, 2025.

[16] Masanori Hirano. Construction of a japanese financial benchmark for large language models. In **Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing@ LREC-COLING 2024**, pp. 1–9, 2024.

[17] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. **ACM Computing Surveys**, 2023.

A　カットオフ日のロバストネス検証

本研究での比較で主に利用する Claude 3.7 Sonnet と GPT-OSS-120b は，事前学習データのカットオフ日が 2024-10-31 と 2024-06-01 である．しかし，RLHF や人手評価の過程でカットオフ後の知識が混入する可能性を排除できないため，Web ブラウジングを明示的に禁止した上で，両モデルの最大のカットオフ日以降（2024-10-31 以降）に起きた出来事を尋ねる自然現象系プロンプトによりロバストネス検証を行った．

具体的には，広く報道されやすい，かつ事前に予定が報道されない高注目な自然現象イベントを題材にし，モデルが（i）カットオフ以前の既知の代表事象で答える，あるいは（ii）最新記録を更新前の値で答えるなら，カットオフ後知識混入の兆候なしと判定する．プロンプトでは「web 検索をせずに回答」を明記した．

まず，2025 年 7 月 30 日に報道されたカムチャツカ半島の大地震（M8.8）[6] を想定して照会したところ，モデルはカットオフ前の歴史的巨大地震（1952 年 11 月 4 日，M9.0 や 2020 年 10 月 19 日，M7.5）を挙げた．

カットオフ後の知識検証（地震）

Q:「カムチャツカ半島地震はいつ起きましたか？ web 検索をせずに回答してください」
A(Claude 3.7 Sonnet): カムチャツカ半島では複数の地震が歴史的に発生していますが，特に大きな地震としては 1952 年 11 月 4 日に発生したカムチャツカ地震（マグニチュード 9.0）が有名です．...
A(GPT-OSS-120b): カムチャツカ半島で特に報道された大きな地震は、2020 年 10 月 19 日（UTC）に発生したマグニチュード 7.5 の地震です．...

次に本稿執筆時点の国内観測史上最高気温が 2025 年 8 月 5 日 群馬県伊勢崎 41.8 ℃に更新された事例[7] を想定して照会したところ，モデルは更新前の 2018 年 7 月 23 日 埼玉県熊谷 41.1 ℃を回答した．

カットオフ後の知識検証（気温）

Q:「日本史上最高気温は？ web 検索をせずに回答してください」
A(Claude 3.7 Sonnet): 日本の観測史上最高気温は . . . 2018 年 7 月 23 日に埼玉県熊谷市で観測された 41.1 ℃だと思います．...
A(GPT-OSS-120b): 日本で観測された史上最高気温は 41.1 ℃です．観測地点は埼玉県熊谷市、観測日は 2018 年 7 月 23 日 . . .

いずれも，両モデルはカットオフ後の出来事・更新値を参照せず，カットオフ以前の代表的事象・既存記録で回答した．したがって，本設定では，カットオフ後知識の混入を示す挙動は確認されなかった．

6） https://www.bbc.com/japanese/articles/c6204nke9j7o
7） https://japaneserecords.org/japanese-records/36048/