

EDINET-Bench：有価証券報告書を用いた日本語金融ベンチマーク

2025年8月21日19:25 ~ 19:50

W&Bミートアップ #24 「生成AIベンチマーク最前線、
高度な推論能力とドメイン特化性能に迫る」

Sakana AI

石田 隆

 <https://takashiishida.github.io>



AI分野における金融ベンチマークの課題

- AI研究者が金融データで研究を進めるには？
 - AI分野のベンチマークはCV、NLP、一般常識、コーディング、数学が多い
 - 金融のデータ（セット）の多くは有料。金融AI研究は金融研究室や金融機関が中心

Chemistry (general)

A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?

A) A metal compound from the fifth period.
B) A metal compound from the fifth period and a non-metal compound from the third period.
C) A metal compound from the fourth period.
D) A metal compound from the fourth period and a non-metal compound from the second period.

Astrophysics

Astronomers are studying a star with a T_{eff} of approximately 6000 K. They are interested in spectroscopically determining the surface gravity of the star using spectral lines (EW < 100 mÅ) of two chemical elements, E1 and E2. Given the atmospheric temperature of the star, E1 is mostly in the neutral phase, while E2 is mostly ionized. Which lines are the most sensitive to surface gravity for the astronomers to consider?

A) E2 I (neutral)
B) E1 II (singly ionized)
C) E2 II (singly ionized)
D) E1 I (neutral)

GPQA (Rein他 2023)

<https://arxiv.org/abs/2311.12022>

*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

✓ waterfall, ✗ bridge, ✗ valley, ✗ pebble, ✗ mountain

*Where can I stand on a **river** to see water falling without getting wet?*

✗ waterfall, ✓ bridge, ✗ valley, ✗ stream, ✗ bottom

*I'm crossing the **river**, my feet are wet but my body is dry, where am I?*

✗ waterfall, ✗ bridge, ✓ valley, ✗ bank, ✗ island

CommonsenseQA (Talmor他 2019)

<https://aclanthology.org/N19-1421.pdf>

Sample problem 2: Find the degree 19 polynomial

Construct a degree 19 polynomial $p(x) \in \mathbb{C}[x]$ such that $X := \{p(x) = p(y)\} \subset \mathbb{P}^1 \times \mathbb{P}^1$ has at least 3 (but not all linear) irreducible components over \mathbb{C} . Choose $p(x)$ to be odd, monic, have real coefficients and linear coefficient -19 and calculate $p(19)$.

Answer: 1876572071974094803391179

MSC classification: 14 Algebraic geometry; 20 Group theory and generalizations; 11 Number theory generalizations

FrontierMath (Glazer他 2024)

<https://arxiv.org/abs/2411.04872>

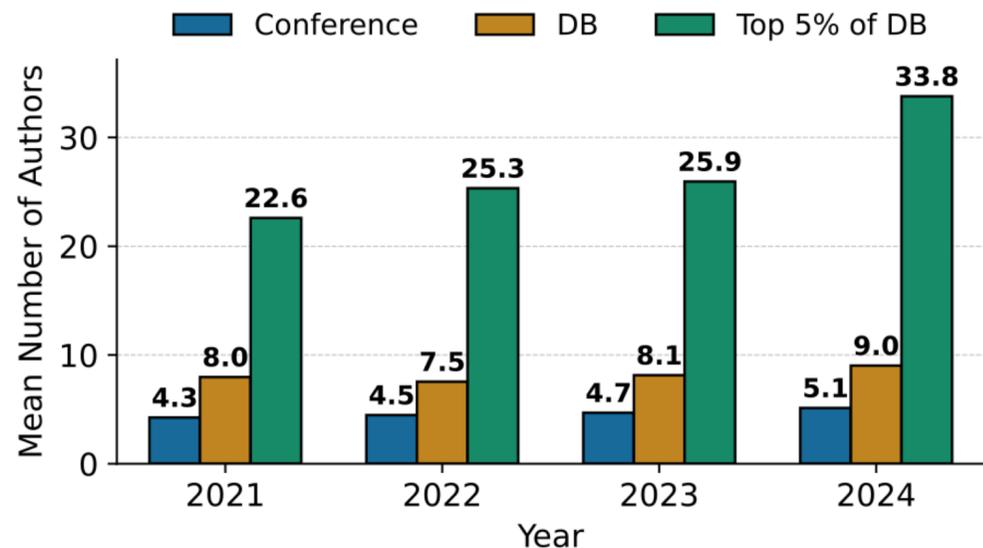
```
def incr_list(l: list):  
    """Return list with elements incremented by 1.  
    >>> incr_list([1, 2, 3])  
    [2, 3, 4]  
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])  
    [6, 4, 6, 3, 4, 4, 10, 1, 124]  
    """  
    return [i + 1 for i in l]
```

HumanEval (Chen他 2021)

<https://arxiv.org/abs/2107.03374>

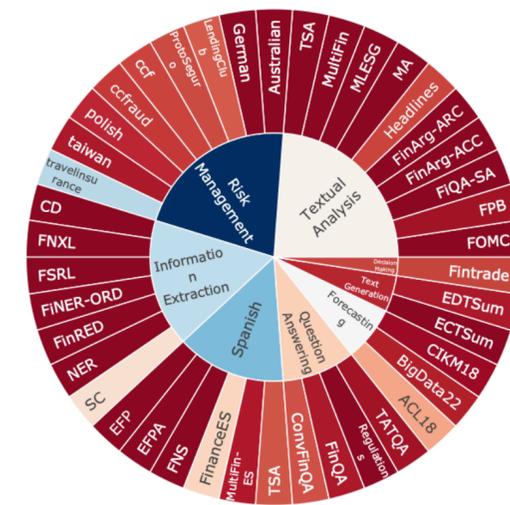
金融領域のLLMベンチマーク動向

NeurIPS main track, datasets & benchmarks trackの著者数



Ishida他 (2025)

24個の金融タスクベンチマーク



FinBen: Xie他 (2024)

中国語の金融ベンチマーク

Open book式金融QA

What is the year end FY2019 total amount of inventories for Best Buy? Answer in USD millions.

✗ The year-end FY2019 total amount of inventories for Best Buy is \$11,395 million.

✗ Sure! [...] the year-end FY2019 total amount of inventories for Best Buy was \$8,144 million, in USD millions.

✓ The total amount of inventories for Best Buy is \$5,409 million.

FinanceBench: Islam他 (2023)

日本語資格試験から構築したQA

An example of cma_basics
Please answer the letter corresponding to the appropriate choice for the following question.

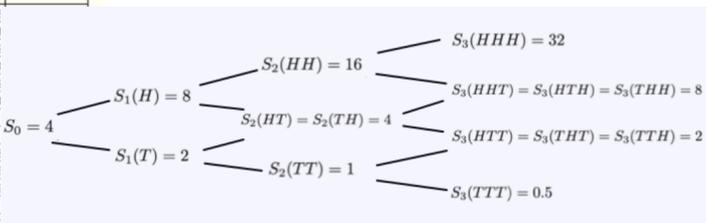
Question:
Which of the following statements about the Japanese economy is incorrect?
A: Real GDP (real gross domestic product) is the level of production activity excluding the effects of price fluctuations.
B: Inflation implies a sustained increase in the general price level.
C: Indirect finance is a form of financial intermediation in which banks and other financial intermediaries play a central role in mediating money lending and borrowing.
D: The fiscal policy of the Bank of Japan adjusts the price level through an increase or decrease in money supply.

Answer:
D

Hirano (2024)
Masuda他 (2023)

Balance Sheet of Clever Company		
Assets	2015 (\$)	2014 (\$)
Current assets:		
Cash	13,862	16,339
Net accounts receivable	23,887	25,778
Inventory	54,867	43,287
Total current assets	92,616	85,404
Fixed assets:		
Plant, property, and equipment	101,541	
Less: Accumulated depreciation	-34,331	
Net fixed assets	67,210	
Prepaid expenses	1,914	
Other assets	13,052	
Total assets	174,794	
Liabilities and Equity	2015 (\$)	
(details omitted)...		
Total current liabilities	34,321	
Long-term debt	22,036	
Total liabilities	56,357	
Equity		
Common stock	38,000	
Paid-in capital	12,000	12,000
Retained earnings	68,435	63,758
Total equity	118,435	113,758
Total liabilities and equity	174,794	167,991

図表を含むベンチマーク



FAMMA: Xue他 (2025)

Anomalous Event Attribution

您是一个知识渊博且乐于助人的金融领域专家，需要基于<时间>和<新闻>回答用户问题，给出令用户满意的答复。以下是一个参考样例：

序号	时间	新闻
1	2023年9月5日	公司半年度业绩交流会上中提到，目前隆基绿能BC产能主要为30GW HPBC（高效复合钝化背接触技术）电池项目，并已于上半年投入量产。
2	2023年11月15日	据报道，特斯拉现已大量发出邀请函，确认Cybertruck交付活动将于11月30日在德州超级工厂举行，这意味着发布四年的Cybertruck终于要交车。
3	2023年11月14日	国盛证券指出，BC电池或成为行业未来新选择，技术加速渗透有望拉动激光设备采购需求，建议重点关注在BC电池布局的龙头电池企业隆基股份。
4	2023年11月14日	当地时间11月14日，美国劳工统计局公布数据显示，CPI环比不变，同比增速降至3.2%；去掉食品、能源价格的核心通胀环比上升0.2%，同比增速降至4.0%。中信证券指出，预计美联储或已完成本轮加息，10年期美债利率短期或走弱，但仍将在4%以上水平运行。

用户问题：哪些新闻导致了特斯拉股票在2023-11-15 这个时间大涨这个事实。

答案：{ "相关新闻序号": [2,4], "具体分析": "特斯拉股票在2023年11月15日大涨，主要受到两条新闻的影响。第一，新闻2提到特斯拉确认Cybertruck交付活动将于11月30日在德州超级工厂举行，这是备受期待的事件，对市场情绪产生了积极影响，反映了市场对新车发布和销售业绩增长的乐观预期。第二，新闻4显示美国核心通胀环比上升0.2%，但总体CPI同比增速降至3.2%，市场对美联储结束加息周期的预期增强，带来了整体市场的反弹。这种宏观利好尤其有助于科技股和成长型公司的上涨，其中特斯拉作为标杆性企业受益显著。"} }

BizFinBench: Lu他 (2025)

金融における複数ターン日本語生成

Roleplay の例題

Turn 1: あなたは資産運用アドバイザーです。クライアントから「老後資金を準備するために、長期的な資産運用を始めたいと思います。おすすめのポートフォリオを教えてください。」という相談を受けました。クライアントの年齢やリスク許容度を確認し、適切な資産配分を提案してください。

Turn 2: クライアントから「提案されたポートフォリオを維持するために、どのくらいの頻度でリバランスを行うべきでしょうか？また、リバランスの重要性を教えてください。」という質問がありました。リバランスの必要性と適切な頻度について説明してください。

平野,今城 (2024)

サーベイ論文

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

中川 慧*, 平野 正徳†, 高野 海斗‡
May 21, 2025

Abstract

本論文は、日本国内における金融分野での大規模言語モデル (LLMs) の研究と応用動向を包括的に調査・整理したサーベイである。2022年11月のChatGPT公開以降、2025年3月までに発表された日本語の関連論文69本を収集し、タスク軸(分類・生成・予測・構築・評価・その他)および目的軸(意思決定の高度化・業務の効率化・モデル性能の評価)に基づいて分類・可視化を行った。最多の研究は「分類」タスクに集中しており、特にセンチメント分析やESG評価が中心のテーマとして浮上している。次いで「生成」タスクが多く、レポートや要約の自動生成による業務の効率化を目指す研究が目立った。一方で、予測や評価タスクは相対的に少なく、今後の研究課題とされる。さらに、実務への応用に際しての判断基準と検討課題を提示した。

Keywords: LLMs, 大規模言語モデル, 金融応用

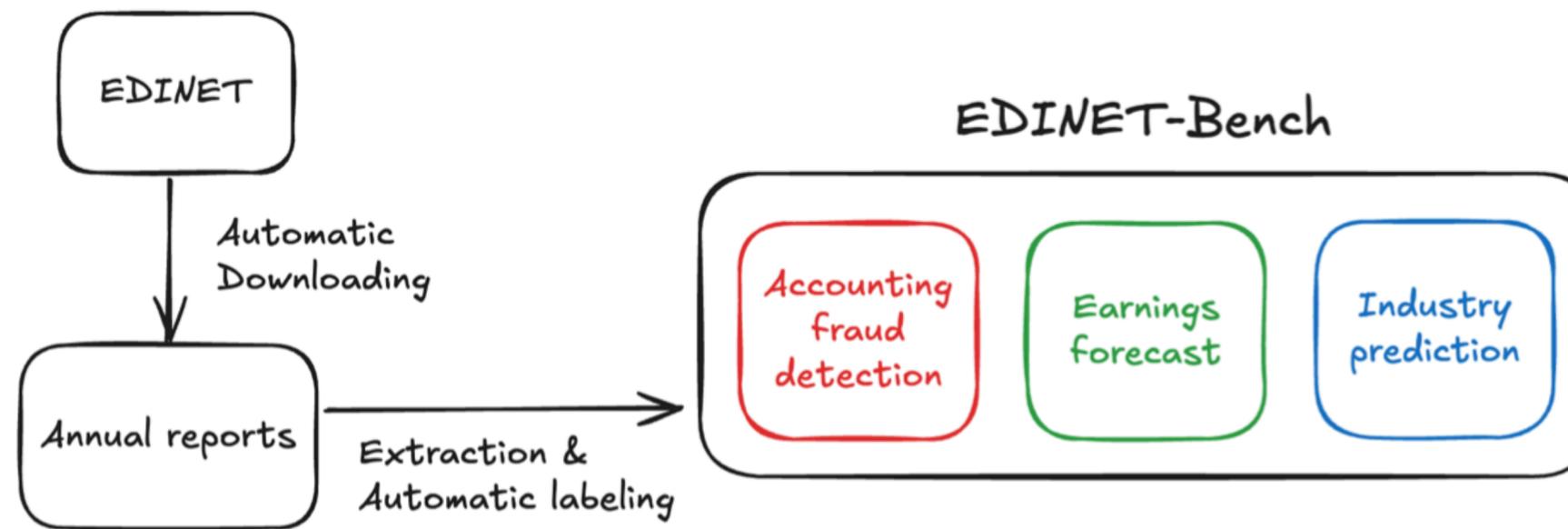
中川,平野, 高野 (2025)

EDINET-Benchの紹介

EDINET-Bench: Evaluating LLMs on Complex Financial Tasks using Japanese Financial Statements

Issa Sugiura^{1,2}, Takashi Ishida¹, Taro Makino¹, Chieko Tazuke¹, Takanori Nakagawa¹, Kosuke Nakago¹ and David Ha¹

¹Sakana AI, ²Kyoto University



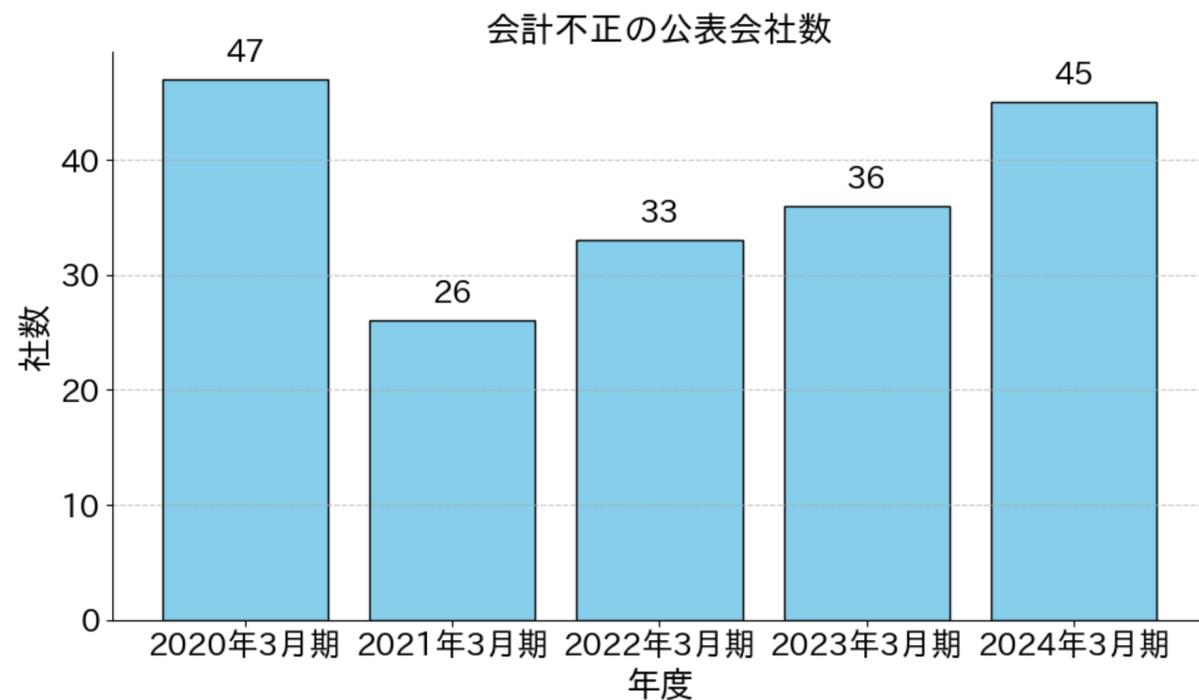
<https://arxiv.org/abs/2506.08762>

EDINET-Benchの概要

- 有価証券報告書の3タスクを提案
 - **粉飾検知**：不正や粉飾が含まれるかどうかを判定
 - **業績予測**：次年度の純利益が増加・減少のどちらかを予測
 - **業種分類**：16業種の中から当てはまる業種に分類
- 特徴
 - 日本語・日本企業
 - LLMにとって難易度が高い
 - AI研究者・開発者にも理解しやすい（3つとも分類タスク）
 - 専門家非依存：自動ラベリングのため将来データに拡張可
 - ツール公開：カスタマイズ可（例：異なる指標、異なるタスク）

タスク①：粉飾検知

- 架空の売上によって業績を良く見せる等の会計不正行為がたびたび発生
- 投資家や金融機関の意思決定を誤らせる粉飾会計は、市場の信頼性を損ない、日本経済に悪影響
- 有価証券報告書の訂正は後から発覚するケースが多く、早期に不正を検知する技術が不可欠
- LLMの活用：不自然な記述や矛盾を抽出できる可能性



会計不正の公表会社数の推移

(日本公認会計士協会「上場会社等における会計不正の動向(2024年版)」図表II-1をもとに作成)

The Detection of Earnings Manipulation

Messod D. Beneish
(corrected January 2004)

Presented are a profile of a sample of earnings manipulators, their distinguishing characteristics, and a suggested model for detecting manipulation. The model's variables are designed to capture either the financial statement distortions that can result from manipulation or preconditions that might prompt companies to engage in such activity. The results suggest a systematic relationship between the probability of manipulation and some financial statement variables. This evidence is consistent with the usefulness of accounting data in detecting manipulation and assessing the reliability of reported earnings. The model identifies approximately half of the companies involved in earnings manipulation prior to public discovery. Because companies that are discovered manipulating earnings see their stocks plummet in value, the model can be a useful screening device for investment professionals. The screening results, however, require determination of whether the distortions in the financial statement numbers result from earnings manipulation or have another structural root.

Beneish (1999)等、古くから研究されているテーマ

時事解説

AIによる不正会計検知・予測の可能性 ——会計監査の未来を探る

宮川大介 一橋大学大学院准教授

宇宿哲平 有限責任 監査法人 公認会計士・公認不正検査士

〈はじめに〉

後を絶たない不適切会計事案を背景に、会計不正を見逃さないことが監査の重要なテーマとなっている。こうした問題に対処するため、あずさ監査法人(以下、あずさ)は、国内大

取引などの金融取引における意思決定や価格形成を歪め、個人、企業レベルにとどまらず、社会全体に大きな損失、不効率をもたらす。また、不正会計により企業の開示情報の信

RIETI Discussion Paper Series 21-J-049

機械学習手法を用いた勘定科目レベルの異常検知

宇宿 他 (2021)

国内でも機械学習を用いた先行事例多数

ラベリングプロセスとデータの特徴

- 粉飾事例の収集方法：
 - オープンなデータセット無
 - 金融庁の課徴金制度「行政上の措置として違反者に対して金銭的負担を課す」：命令一覧
 - 今回は訂正有価証券報告書の存在に着目 <https://www.fsa.go.jp/policy/kachoukin/02.html>
- LLMでフィルタリング
 - 訂正有価証券報告書：軽微なミスによる訂正や、財務数値に関係ない訂正を含む
 - 粉飾検知と比べると簡単なタスク（*ベンチマークの入力は訂正報告書ではなくオリジナルの有価証券報告書）
 - 人間によるチェック：Claude 3.7 Sonnetを用いた場合の誤ラベル率は5%未満

Prompt to detect if the reason of the amendment is related to accounting fraud

以下のテキストは訂正有価証券報告書の冒頭部分です。この訂正有価証券報告書が不適切会計、粉飾決算、会計不正に関連しているかどうかを判断してください。

特に「提出理由」の部分に着目し、以下のような言葉や表現がある場合は不正会計の可能性が高いと考えられます： - 不適切会計 - 会計不正 - 不正行為 - 粉飾決算 - 会計処理の誤り - 売上の過大計上 - 費用の過少計上 - 資産の過大評価 - 不適切な収益認識 - 監査法人からの指摘 - 社内調査 - 第三者委員会

訂正の理由が単純な記載ミスや軽微な修正ではなく、会計上の重大な問題を示している場合は「Yes」と回答し、詳細な説明を提供してください。特に財務諸表（貸借対照表、損益計算書、キャッシュフロー計算書など）の数値に変更が生じた事例に注目してください。会計不正とは関係ない場合や財務諸表に変更が生じていない場合は「No」と回答し、その理由を簡潔に説明してください。

以下のJSON形式で回答してください。このJSONは必ず有効なJSON形式である必要があります：

```
“json { "is_accounting_fraud": bool # true or false "explanation": str # 理由を説明する "company_name": str # 会社名 } “
```

回答は必ずこの形式に一致させてください。JSONは必ず上記の形式の“json”タグで囲んで提供してください。

訂正有価証券報告書のテキスト：

train/testにおける事例数（正例・負例）

Split	Non-fraud	Fraud	Total
Train	453	412	865
Test	102	122	224

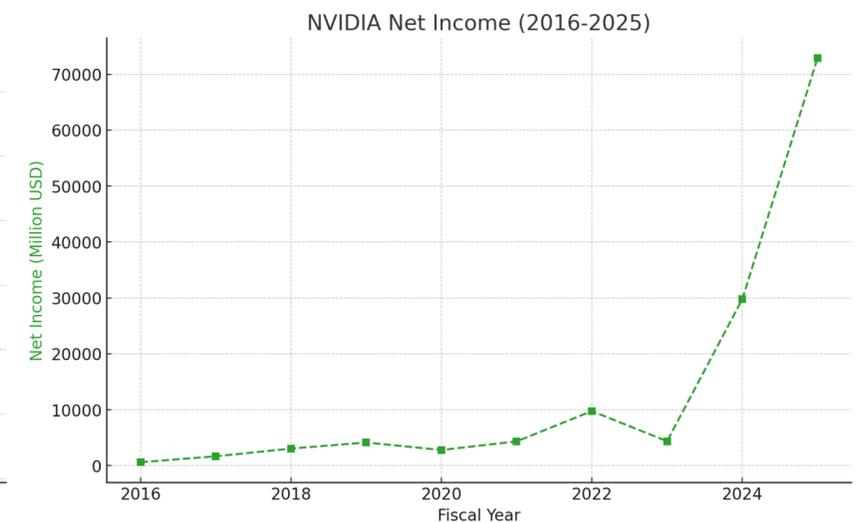
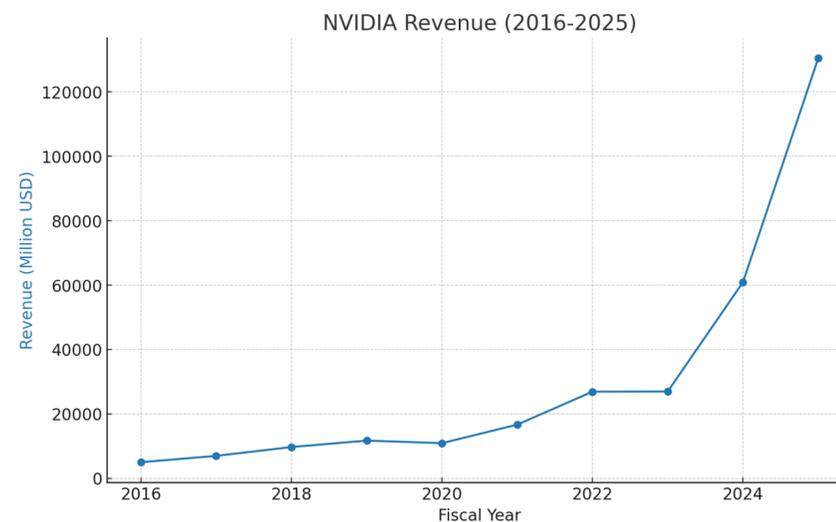
タスク②：業績予測

- 業績予測：個人・機関投資家の投資判断、経営者、経営企画部等の事業計画立案
- 純利益：損益計算書の一番最後、企業活動の最も重要な指標の一つ
- 来期の純利益が増加するか、減少するかのサインを予測するタスク

参考) 米国市場を対象にしたKim他 “Financial statement analysis with large language models”

売上高	18,277,671
売上総利益	4,787,596
営業利益	2,966,857
経常利益	4,538,037
税引前当期純利益	4,538,037
当期純利益	3,845,213

トヨタ自動車の損益計算書より
単位：百万円、2024/4/1～2025/3/31の事業年度



<https://en.wikipedia.org/wiki/Nvidia> の数値をもとに作成。売上予測より純利益予測の方が難しいことが多い。

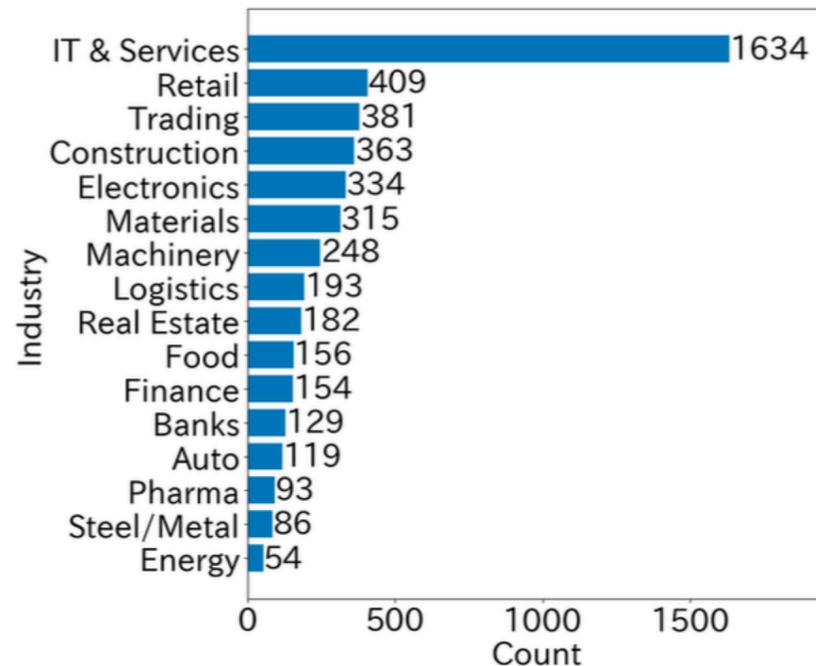
- 9
- オープンなデータセット無。EDINETから業績予測用のベンチマークを自動構築

タスク③：業種分類

- 業種：企業の事業分野や産業の種類を示す区分
 - SICC（証券コード協議会）が有価証券報告書をもとに決定
 - 東証業種別株価指数も人気
- （すでにラベルが存在するのに）なぜ分類したい？
 - 独自の定義に基づいて業種ラベリングしたい：例）運用会社の新商品
 - 再ラベリングしたい：例）M&Aなどで事業が大きく変化
 - 財務諸表の理解能力を計測したい：金融ベンチマークにちょうど良い！

業種タスクの詳細

- EDINETから得られる33業種の情報を用いて、16業種に再構成
- 東証17分類のうち、Energy ResourcesとElectric Power & Gasは事例数が極端に少ないため統合



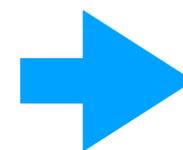
業種別のカウント数

No.	33 Sectors	TOPIX-17 Series
1	Fishery, Agriculture & Forestry	Foods
2	Foods	
3	Mining	Energy Resources
4	Oil and Coal Products	
5	Construction	Construction & Materials
6	Metal Products	
7	Glass and Ceramics Products	
8	Textiles and Apparels	Raw Materials & Chemicals
9	Pulp and Paper	
10	Chemicals	
11	Pharmaceutical	Pharmaceutical
12	Rubber Products	Automobiles & Transportation Equipment
13	Transportation Equipment	
14	Iron and Steel	Steel & Nonferrous Metals
15	Nonferrous Metals	
16	Machinery	Machinery
17	Electric Appliances	Electric Appliances & Precision Instruments
18	Precision Instruments	
19	Other Products	IT & Services, Others
20	Information & Communication	
21	Services	
22	Electric Power and Gas	Electric Power & Gas
23	Land Transportation	Transportation & Logistics
24	Marine Transportation	
25	Air Transportation	
26	Warehousing and Harbor Transportation	
27	Wholesale Trade	Commercial & Wholesale Trade
28	Retail Trade	Retail Trade
29	Banks	Banks
30	Securities and Commodities Futures	Financials (ex Banks)
31	Insurance	
32	Other Financing Business	
33	Real Estate	Real Estate

実験結果

- 設定
 - LLM (Zero-shot, few-shot, SFT) 及びML (logistic regression, etc.) で評価
 - Naive prediction (業績) : 今期の増加・減少が来期も続くと考え同じラベルで予測するベースライン
 - 入力データ : Summary (+ BS,CF,PL) (+Text)の3種類
 - 指標 : ROC-AUC, MCC (粉飾、業績) 正答率 (業種)
- 結果
 - 情報を増やすほど性能が高くなる傾向
 - 改善の余地が残る可能性 : ロジスティック回帰のベースラインを少し上回る程度の性能
 - ドメイン特化の利点 : Llama-3.3-70BよりもLlama-3.2-1B SFTの方が性能が良い

Model	Input Setup	Fraud Detection		Earnings Forecasting		Industry Prediction
		ROC-AUC ↑	MCC ↑	ROC-AUC ↑	MCC ↑	Accuracy ↑
Claude 3.5 Haiku	Summary	0.61 ± 0.01	0.19 ± 0.02	0.41 ± 0.01	-0.02 ± 0.02	0.09 ± 0.00
	Summary+BS+CF+PL	0.60 ± 0.01	0.18 ± 0.03	0.45 ± 0.00	0.00 ± 0.05	0.13 ± 0.00
	Summary+BS+CF+PL+Text	0.67 ± 0.00	0.28 ± 0.02	0.44 ± 0.01	-0.02 ± 0.01	-
Claude 3.5 Sonnet	Summary	0.64 ± 0.01	0.05 ± 0.03	0.54 ± 0.01	0.08 ± 0.01	0.24 ± 0.01
	Summary+BS+CF+PL	0.63 ± 0.03	0.18 ± 0.03	0.55 ± 0.01	0.10 ± 0.02	0.41 ± 0.00
	Summary+BS+CF+PL+Text	0.73 ± 0.02	0.32 ± 0.02	0.52 ± 0.02	0.08 ± 0.02	-
Claude 3.7 Sonnet	Summary	0.59 ± 0.01	0.10 ± 0.02	0.55 ± 0.01	0.06 ± 0.02	0.24 ± 0.01
	Summary+BS+CF+PL	0.58 ± 0.02	0.09 ± 0.04	0.58 ± 0.01	0.13 ± 0.01	0.39 ± 0.01
	Summary+BS+CF+PL+Text	0.67 ± 0.01	0.25 ± 0.02	0.61 ± 0.01	0.16 ± 0.02	-
DeepSeek-V3	Summary	0.61 ± 0.03	0.21 ± 0.02	0.39 ± 0.01	-0.14 ± 0.02	0.12 ± 0.00
	Summary+BS+CF+PL	0.59 ± 0.02	0.12 ± 0.04	0.40 ± 0.01	-0.13 ± 0.01	0.15 ± 0.01
	Summary+BS+CF+PL+Text	0.55 ± 0.03	0.07 ± 0.06	0.39 ± 0.01	-0.15 ± 0.02	-
DeepSeek-R1	Summary	0.54 ± 0.04	0.01 ± 0.08	0.43 ± 0.01	-0.06 ± 0.01	0.11 ± 0.02
	Summary+BS+CF+PL	0.56 ± 0.01	0.09 ± 0.02	0.46 ± 0.00	-0.01 ± 0.01	0.16 ± 0.01
	Summary+BS+CF+PL+Text	0.63 ± 0.01	0.15 ± 0.04	0.44 ± 0.01	-0.05 ± 0.02	-
GPT-4o	Summary	0.59 ± 0.00	0.16 ± 0.02	0.40 ± 0.00	-0.17 ± 0.00	0.14 ± 0.01
	Summary+BS+CF+PL	0.61 ± 0.01	0.19 ± 0.02	0.41 ± 0.00	-0.13 ± 0.01	0.19 ± 0.01
	Summary+BS+CF+PL+Text	0.69 ± 0.01	0.29 ± 0.02	0.41 ± 0.00	-0.14 ± 0.00	-
o4-mini	Summary	0.53 ± 0.00	0.01 ± 0.08	0.51 ± 0.00	0.05 ± 0.02	0.19 ± 0.01
	Summary+BS+CF+PL	0.52 ± 0.01	0.04 ± 0.05	0.54 ± 0.03	0.13 ± 0.04	0.27 ± 0.01
	Summary+BS+CF+PL+Text	0.61 ± 0.01	0.10 ± 0.05	0.58 ± 0.01	0.15 ± 0.01	-
Logistic	Summary	0.68	0.17	0.56	0.05	-
Naive Prediction	-	-	0.44	-0.12	-	



Model	Input Setup	Fraud Detection		Earnings Forecasting		Industry Prediction
		ROC-AUC ↑	MCC ↑	ROC-AUC ↑	MCC ↑	Accuracy ↑
Claude 3.5 Sonnet	Summary	0.64 ± 0.01	0.05 ± 0.03	0.54 ± 0.01	0.08 ± 0.01	0.24 ± 0.01
	Summary+BS+CF+PL	0.63 ± 0.03	0.18 ± 0.03	0.55 ± 0.01	0.10 ± 0.02	0.41 ± 0.00
	Summary+BS+CF+PL+Text	0.73 ± 0.02	0.32 ± 0.02	0.52 ± 0.02	0.08 ± 0.02	-
Claude 3.7 Sonnet	Summary	0.59 ± 0.01	0.10 ± 0.02	0.55 ± 0.01	0.06 ± 0.02	0.24 ± 0.01
	Summary+BS+CF+PL	0.58 ± 0.02	0.09 ± 0.04	0.58 ± 0.01	0.13 ± 0.01	0.39 ± 0.01
	Summary+BS+CF+PL+Text	0.67 ± 0.01	0.25 ± 0.02	0.61 ± 0.01	0.16 ± 0.02	-
Logistic	Summary	0.68	0.17	0.56	0.05	-
Naive Prediction	-	-	-	0.44	-0.12	-

Accounting fraud detection analysis with Claude 3.7 Sonnet

After analyzing the securities report, I've identified several unusual patterns that warrant attention: 1. There's a significant jump in parent company net profit (親会社株主に帰属する当期純利益) from 4,431 million yen in the prior year to 7,468 million yen in the current year (+68.5%), despite operating profit (営業利益) decreasing from 8,693 million yen to 5,592 million yen (-35.7%) and ordinary profit (経常利益) decreasing from 8,135 million yen to 6,773 million yen (-16.7%). This unusual profit pattern requires explanation. 2. The source of this profit increase appears to be from investment securities sales (投資有価証券売却益), which jumped from 20 million yen to 4,078 million yen. This one-time gain is being used to mask deteriorating operational performance. 3. Cash flow patterns show inconsistencies: While the company reports increased profits, operating cash flow decreased from 8,400 million yen to 6,042 million yen. Meanwhile, investing cash flow turned positive (from -5,073 million yen to +2,245 million yen), primarily due to proceeds from investment securities sales (投資有価証券の売却による収入) of 5,551 million yen. 4. The company mentions in the risk section that they sold investment securities to secure liquidity due to COVID-19 concerns, but the scale of these sales suggests they may be liquidating assets to artificially boost profits during a downturn. 5. The company delayed publishing their medium-term management plan twice, citing COVID-19 uncertainty. This could be a legitimate reason, but it also provides a convenient excuse to avoid making specific commitments that might be difficult to achieve given their operational challenges. 6. The company has a high debt ratio (31.9% of total assets are borrowings), which creates pressure to maintain profitability metrics. While these observations don't definitively prove fraud, they suggest the company may be engaging in earnings management by using one-time gains from asset sales to mask deteriorating operational performance. This approach is unsustainable and raises concerns about the company's transparency regarding its true financial condition.

Figure 17: Accounting fraud detection analysis with Claude 3.7 Sonnet. EDINET Code: E00032, DOC ID: S100LQX5, Input: BS+CF+PL+Summary+Text, Ground truth: non-fraud.

データ汚染の有無とその対策

- BS, CF, PL, Summary情報から会社名を予測できるかどうか

	Claude 3.5 Haiku	Claude 3.5 Sonnet	Claude 3.7 Sonnet	DeepSeek-V3	DeepSeek-R1	GPT-4o	o4-mini
Accuracy	0.045	0.050	0.045	0.0	0.005	0.0	0.005

- ダイナミックベンチマーク
 - [LiveBench](#) [White et al., 2025]: 最近の数学コンテスト、ニュース記事、arXiv論文などから質問を収集し、最新の情報を反映。正解に基づいて自動的に回答を採点
 - [AutoBench](#) [Li et al., 2025]: Wikipedia などの情報源から事実に基づいた質問や、Python による計算検証を用いた難易度の高い質問を自動生成
 - [SWE-Bench](#) [Jimenez et al., 2024]: GitHub 上の Python リポジトリに適用。新しい課題や解決策が出るたびに継続的に更新

💡 [EDINET-Bench](#)も事例を毎年更新し続けることができる

まとめ

- サマリー

- 実世界の金融業務（不正検知、業績予測、業種予測）におけるLLMの性能を評価するための、オープンかつ拡張可能な日本語ベンチマークを提案
- 最先端のLLMはベースラインに僅かに上回る程度。モデル改良、エージェント的パイプライン、ドメイン特化の試みの必要性が示唆される

- 課題

- データリーク：有価証券報告書や粉飾の事例等がLLMの訓練セットに含まれているかもしれない 🖱️ 自動更新しやすい設計
- 入力情報が限られているため「本質的難易度」が高い可能性は残る 🖱️ マルチレポート化やウェブアクセスなどの拡張が考えられる

Sakana AI Applied Teamの紹介

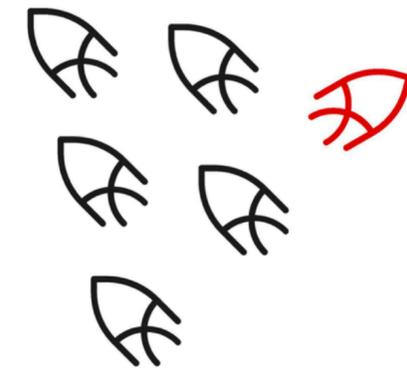
【イベントレポート】 Applied Engineer Open House 2025：金融・防衛の難関課題に挑む、AI社会実装の最前線

August 14, 2025

最先端のAI技術をビジネスへ：Sakana AI、Applied Teamメンバーインタビュー

July 29, 2025

Sakana AI
Applied Engineer
Open House
7th Aug. 2025 @Toranomom
【開催レポート】



sakana.ai

Applied Team
Interview



Applied Research
Engineer
Ryuichi Kanoh



Project
Manager
Takanori Nakagawa

<https://sakana.ai/open-house-2025/>

<https://sakana.ai/applied-team-interview-2025/>

Applied Teamにおける採用情報：<https://sakana.ai/careers/>