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

Issa Sugiura[1,2], **Takashi Ishida**[1], **Taro Makino**[1], **Chieko Tazuke**[1], **Takanori Nakagawa**[1], **Kosuke Nakago**[1] and **David Ha**[1]
[1]Sakana AI, [2]Kyoto University

Financial analysis presents complex challenges that could leverage large language model (LLM) capabilities. However, the scarcity of challenging financial datasets, particularly for Japanese financial data, impedes academic innovation in financial analytics. As LLMs advance, this lack of accessible research resources increasingly hinders their development and evaluation in this specialized domain. To address this gap, we introduce EDINET-Bench, an open-source Japanese financial benchmark designed to evaluate the performance of LLMs on challenging financial tasks including accounting fraud detection, earnings forecasting, and industry prediction. EDINET-Bench is constructed by downloading annual reports from the past 10 years from Japan's Electronic Disclosure for Investors' NETwork (EDINET) and automatically assigning labels corresponding to each evaluation task. Our experiments reveal that even state-of-the-art LLMs struggle, performing only slightly better than logistic regression in binary classification for fraud detection and earnings forecasting. These results highlight significant challenges in applying LLMs to real-world financial applications and underscore the need for domain-specific adaptation. Our dataset, benchmark construction code, and evaluation code is publicly available to facilitate future research in finance with LLMs.

https://huggingface.co/datasets/SakanaAI/EDINET-Bench
https://github.com/SakanaAI/edinet2dataset
https://github.com/SakanaAI/EDINET-Bench

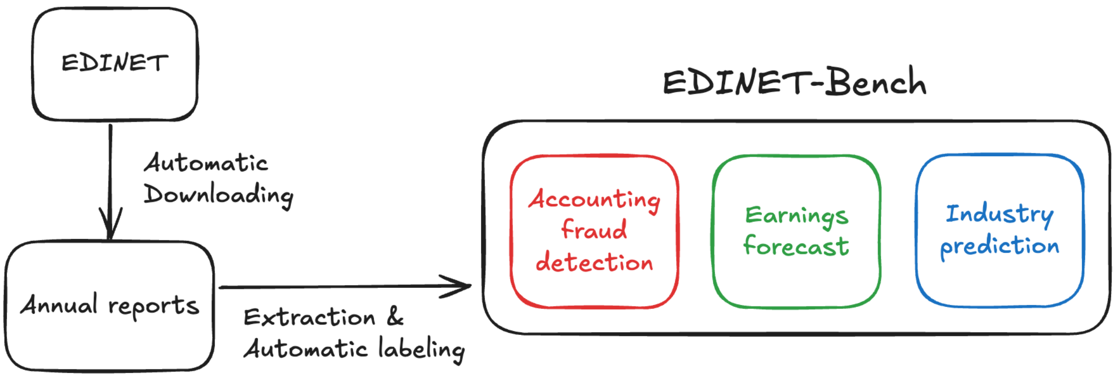


Figure 1 | Overview of EDINET-Bench.

1. Introduction

Large Language Models (LLMs) have recently demonstrated remarkable capabilities across various domains, including general knowledge, mathematics, and coding (Anthropic, 2024; DeepSeek-AI, 2025; Google, 2024; OpenAI, 2024). LLMs are also being explored for applications in the financial sector, where they have shown strong performance in various benchmark tasks (Wu et al., 2023; Xie et al., 2023; Yang et al., 2023). However, most existing benchmarks focus on relatively simple tasks such as information extraction and question answering (Chen et al., 2021, 2022b; Islam et al., 2023).

To make a significant societal impact in the financial domain, LLMs must excel in expert-level decision-making tasks such as accounting fraud detection and earnings forecasting (Beneish, 1999; Kondo et al., 2019; Monahan, 2018; Penman and Sougiannis, 1998; Song et al., 2016). In addition, current financial benchmarks predominantly focus on English-speaking markets (US and Europe) or Chinese-language content. Japan, the fourth largest economy globally by nominal GDP (in 2023; World Bank (2025)), plays a crucial role in the global economy, yet remains underserved and underrepresented by AI technology developments. Japanese financial benchmarks could significantly advance digital transformation efforts throughout the country.

To address these gaps, we introduce EDINET-Bench, an open-source Japanese financial benchmark designed to evaluate LLMs on complex financial tasks. We construct EDINET-Bench using Japan's Electronic Disclosure for Investors' NETwork (EDINET)[1], an electronic system that digitizes and publicly discloses securities reports to improve market efficiency and information accessibility. EDINET-Bench features challenging tasks including accounting fraud detection and earnings forecasting, which require sophisticated financial understanding. Our benchmark is distinctive for its automated construction and exclusive Japanese language content. As the dataset construction is automatic, we can easily incorporate future annual reports.

Using EDINET-Bench, we evaluate the performance of frontier LLMs in a zero-shot setting. Our results indicate that even state-of-the-art models struggle, performing only slightly better than logistic regression in binary classification for fraud detection and earnings forecasting. These results highlight current LLM limitations in complex financial analysis and underscore the need for dedicated research to address these challenges. By open-sourcing our benchmark dataset and toolkit, we enable researchers to tackle challenging financial tasks using AI and facilitate continuous development of the benchmark.

2. Background and related work

Financial benchmarks Many financial benchmarks have been proposed in order to evaluate the financial knowledge and reasoning abilities of LLMs (Chen et al., 2021; Guo et al., 2025; Islam et al., 2023; Xie et al., 2024). For example, FinQA (Chen et al., 2021) and ConvFinQA (Chen et al., 2022b) are closed numerical reasoning question answering tasks related to financial analysis. Similarly, FinanceBench (Islam et al., 2023) is an open-book financial question answering benchmark dataset, and FinBen (Xie et al., 2024) is a benchmark-suit that contains 24 tasks. FAMMA (Xue et al., 2025) uses university textbooks and the CFA exam to construct a multimodal question answering benchmark. Several benchmarks exist in the Japanese financial domain (see Nakagawa et al. (2025) for a recent survey), including Japanese-lm-fin-harness (Hirano, 2024) and JCPA dataset (Masuda et al., 2023), which focus on Japanese financial exam questions. Most of these datasets rely on financial experts to design high-quality questions, and therefore cannot be easily extended to include up-to-date questions after the publication. On the other hand, our framework can easily incorporate future financial documents, and automatically annotates the labels for each task.

Accounting fraud detection Accounting fraud detection involves identifying deliberate misstatements, manipulations, or discrepancies in financial reports. These reports are critical for numerous stakeholders, as investors base allocation decisions on them, while job seekers evaluate potential employers through their financial disclosures (Sou, 2018). Despite preliminary audits, fraudulent activities are frequently discovered only after publication, necessitating subsequent corrections (Japan Institute of Certified Public Accountants, 2024). Accounting fraud detection has

[1] https://disclosure2.edinet-fsa.go.jp/

Table 1 | Sample of TSV records. Some columns are omitted for ease of reading.

Element Name	Relative Year	Unit	Value
提出回数 (Number of Submissions)	提出日時点 (As of Submission Date)	–	1
売上高 (Sales)	前期 (Previous year)	円 (Yen)	10000000
売上高 (Sales)	今期 (Current year)	円 (Yen)	20000000

been studied for many years, with pioneering work by Beneish (1999). After that, traditional machine learning approaches have been applied to accounting fraud detection (Dechow et al., 2011; Kondo et al., 2019; Perols, 2011; Song et al., 2016; West and Bhattacharya, 2016), but effectively processing textual information alongside numerical data in annual reports using LLMs remains underexplored.

Earnings forecasting While earnings forecasting typically involves predicting the numeric magnitude of future earnings (Monahan, 2018; Penman and Sougiannis, 1998), a large body of work instead focuses on predicting the sign of the change in earnings (Chen et al., 2022a; Ou and Penman, 1989), which remains a difficult task for professionals. Kim et al. (2024) investigated if GPT-4, when provided with financial statements in which company names and fiscal years are anonymized, and without using any textual information, will outperform human analysts in predicting the direction of earnings for the following year. However, it is based on proprietary data. On the other hand, we will publish both the evaluation dataset and evaluation code to facilitate future research in this area.

Industry prediction Industry prediction is a multi-class classification task that aims to predict the industry category based on security reports. While listed firms already are tagged with industry labels, e.g., by Securities Identification Code Committee (SICC), portfolio managers may want to adopt their own industry definitions, e.g., see Kimura and Nakagawa (2022) for a data-driven approach, or anticipate reclassifications as firms' business evolve or mergers reshape their operations. Evaluating whether LLMs can predict industries based on financial information such as balance sheets and profit/loss statements serves as an effective task for measuring LLMs' financial domain knowledge. Dolphin et al. (2023); Van Der Heijden (2022) used machine learning for this application.

EDINET Electronic Disclosure for Investors' NETwork (EDINET) is a platform managed by the Financial Services Agency (FSA) of Japan that provides access to disclosure documents such as securities reports. This platform offers both a web interface and an EDINET API to access reports, allowing users to download annual, semi-annual, quarterly, and amended reports for the past ten years. In addition to PDFs, TSV files are also available, containing structured data where each row represents a record of individual attributes that exist in the annual report PDF. Table 1 shows examples of TSV records. Detailed information for each disclosure item can be extracted and analyzed by parsing these attributes. EDINET operates as the equivalent of EDGAR[2] for the United States. We use EDINET as the primary data source for our benchmark.

3. Construction of EDINET-Bench

In this section, we describe the construction method of EDINET-Bench. Overview of our construction pipeline is shown in Figure 1. Generally, EDINET-Bench is constructed by downloading Japanese listed companies' annual reports from EDINET and automatically assigning labels for each task.

[2] https://www.sec.gov/edgar/search/

Table 2 | Number of annual reports per fiscal year.

Start of Fiscal Year	# Reports
2014	3,638
2015	4,047
2016	4,065
2017	4,111
2018	4,120
2019	4,146
2020	4,206
2021	4,242
2022	4,263
2023	4,267
2024	586
Total	41,691

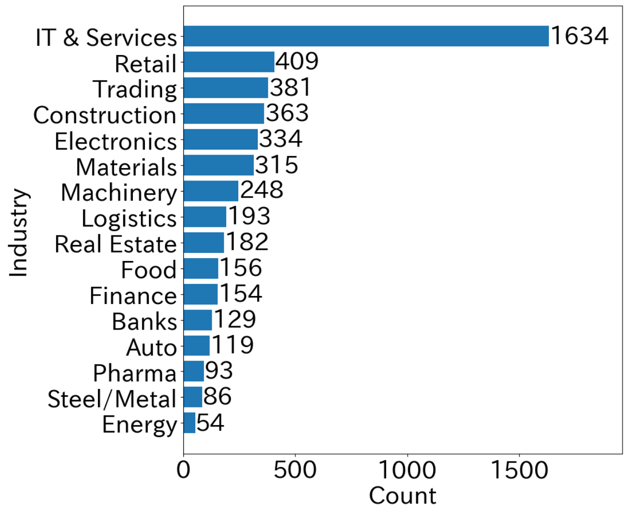


Figure 2 | Number of companies per industry.

3.1. edinet2dataset

We first create edinet2dataset, a tool for downloading and parsing financial documents from EDINET. This tool includes a download function for securities reports using the EDINET API and a parsing function to extract financial information from the downloaded report files in TSV format. The parsing process leverages `Polars`[3] to enable high-speed processing. edinet2dataset is inspired by the edgar-crawler[4] from the Edger-Corpus (Loukas et al., 2021) and shares similar functionality. However, edinet2dataset is capable of extracting more detailed information, such as the current year's sales. The information extracted by edinet2dataset can be broadly categorized as follows: Meta: Metadata such as company name. Summary: Key financial indicators. BS: Balance sheet statement. PL: Profit and loss statement. CF: Cash flow statement. Text: Other textual information in the report, such as company history, business risk explanations, and auditor information. Appendix B shows an example of parsed items from an annual report using edinet2dataset.

3.2. EDINET-Corpus

By using edinet2dataset, we collect all available annual reports and amended annual reports from EDINET for the period from April 2014 to April 2025, which includes approximately 40,000 documents (4,000 listed companies × 10 years). The number of annual reports obtained from EDINET over the past ten years for each year, as well as the number of companies per industry, are shown in Table 2 and Figure 2. We call this dataset the EDINET-Corpus, and it will be publicly available for reproducibility. The EDINET-Corpus serves as the primary data source for constructing EDINET-Bench, as described in the next section.

3.3. EDINET-Bench

Using the aforementioned tool and corpus, we construct three challenging financial benchmark tasks: accounting fraud detection, earnings forecasting, and industry prediction.

Accounting fraud detection We construct a binary classification dataset for accounting fraud detection. We collect both fraudulent and non-fraudulent reports in order to build this dataset. For the fraudulent reports, we first downloaded the past ten years of amended annual reports from

[3] `https://github.com/pola-rs/polars`
[4] `https://github.com/lefterisloukas/edgar-crawler`

Table 3 | Class distribution in accounting fraud detection task.

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

Table 4 | Class distribution in earnings forecasting task.

Split	Decrease	Increase	Total
Train	254	295	549
Test	157	294	451

Table 5 | Class distribution in industry prediction task.

Industry	Count
Banking	35
Electronics & Precision Instruments	34
Automobiles & Transportation	34
Transportation & Logistics	34
Electricity, Gas & Energy Resources	34
Real Estate	33
Machinery	33
Steel & Nonferrous Metals	33
Materials & Chemicals	32
Finance (Excluding Banking)	31
Food	31
Construction & Materials	29
Trading Companies & Wholesale	28
Information & Communication Services	28
Pharmaceuticals	24
Retail	23
Total	476

EDINET, obtaining a total of 6,712 reports. An amended annual report is a document that discloses amendments to the original annual report. These amendments can be related to fraudulent activities or to non-fraudulent issues, such as misreporting the ownership percentage and share count of major shareholders, under-reporting executive compensation, or failing to disclose critical information. To identify amendments linked to fraudulent activities, we extract text from amended annual reports in PDF format using pdfminer[5], a tool for PDF text extraction. We then analyze the extracted text using Claude 3.7 Sonnet, prompting the model to determine whether amendments are associated with fraud. The specific prompt used for this classification is detailed in Appendix C. As a result, 668 reports were identified as fraudulent and are labeled as fraudulent. Note that our prompt leaves room for unintentional errors, such as misstatements, to be included. According to our manual check, our fraudulent set consists mostly of intentional manipulations, but a small proportion of the reports are unintentional errors. For the non-fraudulent reports, we randomly sample 700 companies from all annual reports submitted over the past ten years, excluding those identified as fraudulent. From each selected company, we randomly choose one annual report from the reports of the company that have been downloaded. We then split the fraudulent and non-fraudulent samples into training and test sets, ensuring that data from the same company does not overlap between the two sets. This resulted in a dataset with 865 samples for training and 224 samples for testing, totaling 1,089 instances. Due to document parsing errors, the final dataset was reduced to 534 fraudulent and 555 non-fraudulent samples, smaller than the initial 668 and 700 samples, respectively. The class distribution of the final dataset is shown in Table 3.

Earnings forecasting We construct a binary classification dataset for earnings forecasting to predict whether a company's earnings will increase or decrease in the following fiscal year, given its current annual report[6]. To create this dataset, we randomly select 1,000 companies. For each company, we generate pairs of consecutive annual reports spanning two fiscal years, such as (2015, 2016), (2016, 2017), . . . , (2023, 2024). From these pairs, we randomly choose one pair for each company to include in the dataset (e.g., (2016, 2017)). For each sampled pair, we extract the value of "Profit Attributable to Owners of the Parent Company"[7] from the corresponding TSV files. We then compare the profit value between the two years: if the current year's profit exceeded the previous year's, the

[5] https://pypi.org/project/pdfminer/

[6] Note that with our edinet2dataset, it is relatively easy to modify the task depending on the user's needs, e.g., change to a regression task or use other metrics.

[7] This value represents the portion of the consolidated net profit that is attributable to the shareholders of the parent company.

instance is labeled as an increase; otherwise, it is labeled as a decrease. Additionally, to evaluate baseline performance, we collect labels for earnings changes between two and one year prior as a naive baseline model. The dataset is split into train and test sets based on the fiscal year of the previous report in each pair. If the fiscal year started on or before January 1, 2020, it is included in the train set; otherwise, it is assigned to the test set. This resulted in 549 samples for training and 451 for testing, totaling 1,000 instances. The class distribution for each split is shown in Table 4.

Industry prediction We construct a multi-class classification dataset for industry prediction to predict a company's industry based on its annual report. We use the 33-class industry information provided by EDINET for each company, also known as TOPIX-33 and is widely recognized among Japanese investors. The industry definitions follow the classification determined by the SICC, which is based on the Japan Standard Industrial Classification published by the Ministry of Internal Affairs and Communications (MIC). However, the granularity of 33 industries is considered too fine for our purposes. To enhance interpretability and simplify the prediction task, we consolidate similar industries into 16 broader categories[8]. We categorize all companies available in the EDINET-Corpus into 16 industry groups and randomly sample approximately 35 companies from each category. For each sampled company, we use reports with fiscal year starting from 2023. This process yielded a dataset of 496 samples with the class distribution shown in Table 5. In this task, since the minimum number of companies per industry among the 16 industries was only 23, which is relatively small, we did not split the data into training and test sets.

4. Evaluation

We evaluate the zero-shot performance of current state-of-the-art LLMs, along with classical baselines, on EDINET-Bench.

4.1. Evaluation Setup

Baseline models We evaluate seven frontier LLMs in a zero-shot setting, including GPT-4o (gpt-4o-2024-11-20) and o4-mini (o4-mini-2025-04-16) from OpenAI, Claude 3.5 Haiku (claude-3-5-haiku-20241022) and Claude 3.5 Sonnet (claude-3-5-sonnet-20241022) and Claude 3.7 Sonnet (claude-3-7-sonnet-20250219) from Anthropic, and DeepSeek-V3[9] and DeepSeek-R1[10] from DeepSeek. For inference, the temperature parameter is set to 0.0 and the maximum token length for generation is set to 4096. For GPT-4o, the random seed is fixed at 0. We use the test split for evaluation. We use the test split for evaluation and perform three runs for LLMs, as predictions with these models are not fully deterministic even with the temperature set to 0.0. We also use following baseline models:

- **Logistic regression**: We use this model for accounting fraud detection and earnings forecasting tasks. We train this model on the training split and evaluate it on the test split. We use scikit-learn (Pedregosa et al., 2011) for the implementation, employing its default hyperparameter settings. To avoid excessive input attributes, only the Summary section of the annual report is used as the input feature.
- **Naive prediction**: This model is used solely for earnings forecasting, predicting this year's change will follow last year's trend.

Prompt We use the following system prompt: "You are a financial analyst." For each task, we use the instruction prompt shown in Appendix A. For the binary classification tasks of accounting

[8] In TOPIX-17, the number of cases in the industries of "Energy Resources" and "Electricity & Gas," which are loosely related domains, was smaller compared to other industries. Therefore, these two were merged to create 16 industries.

[9] https://openrouter.ai/deepseek/deepseek-chat

[10] https://openrouter.ai/deepseek/deepseek-r1

fraud detection and earnings forecasting, we prompt the models to output a prediction (0 or 1), a probability value (between 0 to 1), and the reasoning behind their judgment based on information from the securities reports. For the multi-class industry prediction task, we prompt the models to output the industry class and their reasoning for the classification.

Data processing We extract the following items from the TSV files using edinet2dataset for prediction: Balance Sheet (BS), Cash Flow (CF), Profit and Loss (PL), Consolidated Summary (Summary), and Text Block (Text). These items are concatenated, appended to the end of the instruction prompt, and fed into the models in a zero-shot manner. To assess how information quantity affects LLM performance, we tested three configurations: Summary only, BS+CF+PL+Summary, and all items (BS+CF+PL+Summary+Text). Note that BS, CF, PL, and Summary consist of tabular data and do not contain company names or specific year information, whereas Text includes unstructured textual content that contains both company names and year information.

Cost and processing time analysis Each annual report in EDINET-Bench contains approximately 30,000 tokens spanning all sections. The average output length is around 500 tokens. For Claude 3.7 Sonnet, which costs $3 per million tokens for input and $15 per million tokens for output, the cost per report is approximately $0.1. Regarding processing time, Claude 3.7 Sonnet can handle each case in about 10 seconds.

Evaluation metrics For accounting fraud detection and earnings forecasting, we use ROC-AUC (Receiver Operating Characteristic Area Under the Curve) and MCC (Matthews Correlation Coefficient) as evaluation metrics, and for industry prediction, we use accuracy.

4.2. Results

Table 6 shows the performance of each model on each task in EDINET-Bench.

Accounting fraud detection In the accounting fraud detection task, all LLMs faced challenges, with even the state-of-the-art Claude 3.5 Sonnet only slightly surpassing logistic regression in terms of ROC-AUC. Notably, most models demonstrated substantial improvements when text information was incorporated, indicating effective utilization of textual data for fraud detection.

Logistic regression, despite its simplicity, achieved surprisingly strong results, only slightly behind the best model/setup (Claude 3.5 that uses Text) in fraud detection. This highlights the substantial room for LLM improvement.

We take a closer look at the feature importance of logistic regression, shown in Table 7. This reveals that total comprehensive income and total assets were influential factors in non-fraud predictions, suggesting a bias towards predicting larger companies as non-fraud and smaller ones as fraud. Furthermore, the confusion matrix of Claude 3.7 Sonnet, shown in Figure 4, illustrates how incorporating textual information improved Sonnet's accuracy in classifying both fraud and non-fraud cases.

Earnings forecasting For the earnings forecasting task, many LLMs struggled to deliver precise predictions. Even Claude 3.7 Sonnet achieved only a ROC-AUC of 0.61, aligning with the low forecasting performance reported by Xie et al. (2024). This limitation likely stems from the challenge of predicting next-year earnings based solely on the current securities report. Unlike the fraud detection task, incorporating text information did not enhance performance in earnings forecasting, suggesting a weaker reliance on textual factors.

Table 6 | Performance of each model on each task in EDINET-Bench. For LLMs, the scores represent the mean ± standard deviation over three runs. **Bold** indicates the best score. "–" indicates the score that could not be calculated due to inappropriate settings. For industry prediction, results including text input are not displayed since the text section of the annual report often contains industry-related information. As a note, when text was incorporated, many models achieved accuracy exceeding 0.75.

		Fraud Detection		Earnings Forecasting		Industry Prediction
Model	**Input Setup**	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	–

Industry prediction In the industry prediction task, all models demonstrated predictive performance significantly higher than random guessing. Furthermore, across all models, increasing the input information from Summary to Summary+BS+CF+PL resulted in improved performance. Notably, Claude 3.5 Sonnet achieved state-of-the-art performance with an accuracy score of 0.41 when provided with Summary+BS+CF+PL. The confusion matrix of Claude 3.5 Sonnet for the industry prediction task is shown in Figure 5. We can observe that increasing the input information improves its performance.

These results indicate that industry prediction is a relatively simpler task compared to accounting fraud detection and earnings forecasting, with predictions often possible using only securities reports. This may be attributed to the distinctive characteristics present in the BS, CF, PL, and Summary statements across different industries, such as how assets are primarily composed of "cash and deposits", "loans", and "securities" in specific proportions depending on the industry.

5. Contamination

Since the securities reports included in EDINET-Bench are publicly available on the Internet, there is a potential for contamination. As a simple check, we conduct company name prediction and year-wise performance analysis.

Company name prediction To confirm whether the LLM already knew the content of the securities reports used in the evaluation dataset, we measure the performance of a task where the model is

Table 7 | Feature importance of logistic regression on accounting fraud detection, sorted by absolute value of feature importance scores.

Feature	Importance	Abs Importance
Comprehensive Income (CurrentYear)	-1.165	1.165
Total Assets (Prior4Year)	-1.118	1.118
Total Assets (Prior3Year)	-1.073	1.073
Comprehensive Income (Prior3Year)	1.066	1.066
⋮	⋮	⋮
Price-to-Earnings Ratio (Prior2Year)	0.019	0.019
Price-to-Earnings Ratio (Prior4Year)	-0.014	0.014
Return on Equity (Prior1Year)	-0.008	0.008
Price-to-Earnings Ratio (Prior1Year)	0.001	0.001

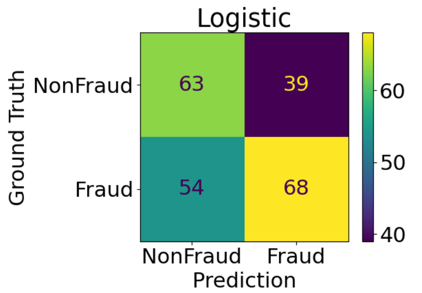
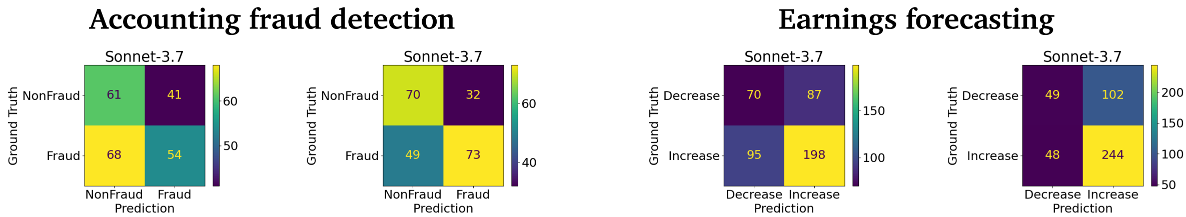


Figure 3 | Confusion matrix of logistic regression on accounting fraud detection.



a BS+CF+PL+ Summary

b BS+CF+PL+ Summary+Text

c BS+CF+PL+ Summary

d BS+CF+PL+ Summary+Text

Figure 4 | Confusion matrix of Claude 3.7 Sonnet on accounting fraud detection and earnings forecasting.

given tabular data (BS, CF, PL, Summary) from the test split of the accounting fraud detection dataset and asked to predict the company names. The prompt is shown in Figure 10. For evaluation, we allowed minor company name variations such as including or omitting "Kabushiki Kaisha" by relying on the LLM-as-a-judge approach. For all models, the accuracy is below 0.05, indicating the difficulty for models to associate tabular data with company names.

	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

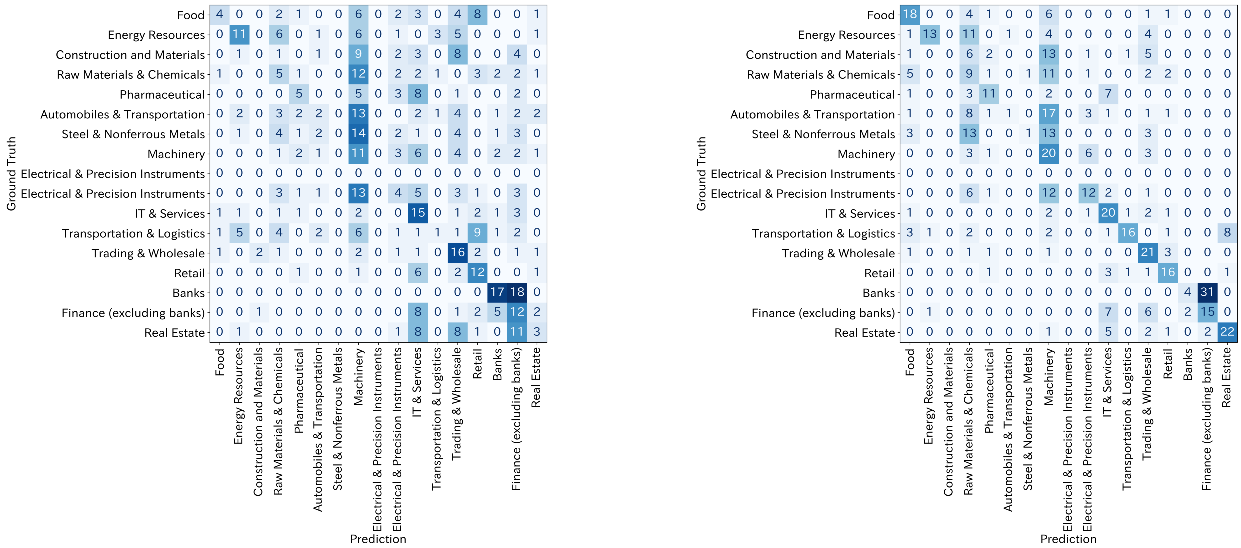
Table 8 | Performance of each model on company name prediction. BS+CF+PL+Summary is used as input. The comparison between the ground truth company names and the predicted names was conducted using GPT-4o as a judge to account for inconsistencies in name representations.

Year-wise performance analysis We divide the test split of accounting fraud detection by fiscal start year, and plot the scores for each year. If contamination were present, we would expect higher performance on older data, but as shown in Figure 6, we did not observe the anticipated trend of declining performance on more recent data.

6. Limitations

Our approach to constructing EDINET-Bench has several limitations:

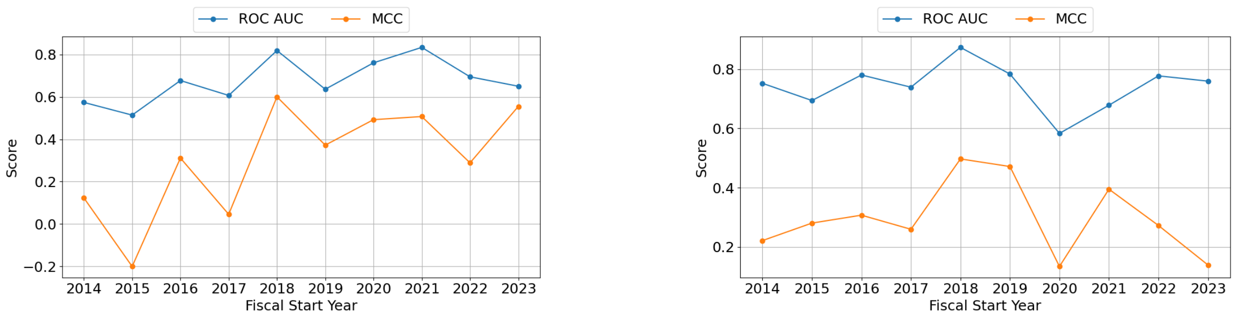
1. **Intrinsic difficulty**: Among the tasks in our benchmark, the fraud detection and earnings forecasting tasks may be intrinsically challenging with a performance upper bound, as the LLM relies solely on information from a single annual report for its predictions. According to interviews we conducted with auditors, when detecting fraud, auditors first identify suspicious areas based on statistical information from sources like annual reports. They then access confidential internal information and other detailed data required to calculate the statistics



a Summary

b Summary + BS + CF + PL

Figure 5 | Confusion matrix of Claude 3.5 Sonnet on industry prediction.



a GPT 4o

b Claude 3.5 Sonnet

Figure 6 | Performance per fiscal first year on accounting fraud detection.

presented in the annual report to verify their accuracy. Therefore, future research directions could explore the development of benchmark task designs that enable the model to utilize information beyond the annual report with novel agentic pipelines. Additionally, benchmark evaluations could be designed based on a *rubric evaluation*, where model responses are evaluated according to multiple criteria tailored to each specific question (Arora et al., 2025; Starace et al., 2025).

2. **Mislabeling**: When constructing the benchmark dataset for the accounting fraud detection task, we assume that only cases explicitly reported as fraudulent are labeled as such, while all others are considered non-fraudulent. However, there may be undiscovered fraud cases that remain unreported, introducing potential label noise into the dataset. Additionally, our fraud examples are constructed by having the LLM read the contents of the amended reports and determine whether they are related to fraudulent activities. Due to the hallucination problem inherent in LLMs (Huang et al., 2025) and lack of instruction following abilities, there is a risk that some cases may be incorrectly identified as fraudulent. To evaluate the reliability of these classifications, we manually checked the amended reports labeled as fraud cases. We found that most of them involved corrections typically associated with fraudulent behavior, such as fictitious sales. However, we also observed a few cases involving corrections unrelated to

financial statements, such as revisions to the number of board members, which do not directly impact the BS, CF, or PL and therefore do not fit our definition of accounting fraud in Appendix 16.

3. **Contamination**: Our benchmark dataset is constructed using EDINET. Since EDINET's published documents are accessible on the Internet, some reports in EDINET-Bench may have been included in training datasets of LLMs. In such cases, the models might already have knowledge of the correct labels (Oren et al., 2024). However, since our benchmark can be automatically constructed, we can update our benchmark with future annual reports that are not part of the LLMs' pre-training data, thereby mitigating contamination concerns. Also see Sec. 5 for our preliminary contamination analysis.
4. **Parsing inconsistency**: Although our tool, edinet2dataset, is capable of extracting detailed financial information, such as total assets for each fiscal year from TSV files, variations in terminology and formatting across different securities reports can lead to parsing inconsistencies. As a result, LLMs may interpret the absence of parsed data not as a parsing error, but as if the original securities report lacks the information entirely, potentially leading to false positives in accounting fraud detection. Therefore, it is necessary to specify this aspect in the prompt to avoid misjudgment.

7. Conclusion

In this work, we introduced EDINET-Bench, an open-source Japanese financial benchmark designed to evaluate the performance of LLMs on challenging financial tasks including accounting fraud detection, earnings forecasting, and industry prediction. We leveraged Japan's Electronic Disclosure for Investors' NETwork (EDINET) to automatically construct EDINET-Bench and can easily incorporate future reports using our toolkit, edinet2dataset. Our experiments reveal that even state-of-the-art LLMs perform only slightly better than logistic regression in binary classification for fraud detection and earnings forecasting. These results highlight current LLM limitations in complex financial analysis and underscore the need for dedicated research to address these challenges, including novel agentic pipelines. We hope our benchmark and toolkit will accelerate advancements in applying large language models to complex financial analysis tasks.

Ethical statement and broader impact

We constructed EDINET-Bench using only publicly available data from EDINET. By releasing EDINET-Bench and edinet2dataset, we aim to contribute significantly to the development of LLMs in real-world financial domain applications. However, since the data represents real companies, there is a risk of misuse, such as damaging a company's reputation. This benchmark is intended solely for advancing LLM applications in finance and must not be used to target or harm any real companies included in the dataset. In some cases, we noticed fairness issues such as LLM referencing the reputation of auditors (e.g., "*Their financial statements have been audited by a reputable firm ([auditor name])*") as part of the decision-making process, which may exacerbate bias against newer or smaller auditors. This demonstrates that alignment and fairness considerations benefit from being examined within each specific domain (in our case, finance). Furthermore, continued research into and mitigation of these societal issues is vital for the responsible advancement of the community. Finally, there is a possibility that this dataset could be misused to make fraud more difficult for LLMs to detect, and we strongly discourage any such attempts, especially in real-world applications.

Acknowledgments

We would like to thank Shimpei Fukagai, Shengran Hu, and Robert Tjarko Lange for helpful discussions.

References

Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health, 2025. URL https://arxiv.org/abs/2505.08775.

Messod D. Beneish. The detection of earnings manipulation. *Financial Analysts Journal*, 55(5):24–36, 1999. doi: 10.2469/faj.v55.n5.2296. URL https://doi.org/10.2469/faj.v55.n5.2296.

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*, 60(2):467–515, 2022a.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *EMNLP*, pages 3697–3711, November 2021. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300/.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *EMNLP*, pages 6279–6292, December 2022b. doi: 10.18653/v1/2022.emnlp-main.421. URL https://aclanthology.org/2022.emnlp-main.421/.

Patricia M Dechow, Weili Ge, Chad R Larson, and Richard G Sloan. Predicting material accounting misstatements. *Contemporary accounting research*, 28(1):17–82, 2011.

DeepSeek-AI. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Rian Dolphin, Barry Smyth, and Ruihai Dong. A machine learning approach to industry classification in financial markets. In *Artificial Intelligence and Cognitive Science*, pages 81–94. Springer Nature Switzerland, 2023. ISBN 978-3-031-26438-2.

Google. Our next-generation model: Gemini 1.5, 2024. URL https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models. In *NAACL*, pages 6258–6292, April 2025. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.318/.

Masanori Hirano. Construction of a Japanese Financial Benchmark for Large Language Models. In *Joint Workshop of FinNLP, KDF, and ECONLP*, pages 1–9, 2024. doi: 10.2139/ssrn.4769124. URL https://aclanthology.org/2024.finnlp-1.1.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering, 2023. URL https://arxiv.

org/abs/2311.11944.

Japan Institute of Certified Public Accountants. Trends in accounting fraud among listed companies, 2024. URL https://jicpa.or.jp/specialized_field/files/2-3-10-2-20240716.pdf.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models, 2024. URL https://arxiv.org/abs/2407.17866. Paper withdrawn by authors on February 20, 2025.

Yuya Kimura and Kei Nakagawa. Industry momentum strategy based on text mining in the japanese stock market. In *IIAI-AAI*, pages 420–423, 2022. doi: 10.1109/IIAIAAI55812.2022.00089.

Satoshi Kondo, Daisuke Miyakawa, Kengo Shiraki, Miki Suga, and Teppei Usuki. Using machine learning to detect and forecast accounting fraud, 2019.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. EDGAR-CORPUS: Billions of tokens make the world go round. In *ECONLP*, pages 13–18, November 2021. doi: 10.18653/v1/2021.econlp-1.2. URL https://aclanthology.org/2021.econlp-1.2/.

Tatsuki Masuda, Kei Nakagawa, and Takahiro Hoshino. Can ChatGPT pass the JCPA exam? Challenge for the short-answer method test on auditing. In *JSAI*, 2023.

Steven J. Monahan. Financial Statement Analysis and Earnings Forecasting. volume 12, pages 105–215, July 2018. doi: 10.1561/1400000036. URL https://ideas.repec.org/a/now/fntacc/1400000036.html.

Kei Nakagawa, Masanori Hirano, and Kaito Takano. Survey and perspectives on large language models in Japanese financial sector. 2025. URL https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/1268/.

OpenAI. Gpt-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *ICLR*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

Jane A Ou and Stephen H Penman. Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics*, 11(4):295–329, 1989.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Stephen H Penman and Theodore Sougiannis. A comparison of dividend, cash flow, and earnings approaches to equity valuation. *Contemporary accounting research*, 15(3):343–383, 1998.

Johan Perols. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, 30(2):19–50, 2011.

Mingzi Song, Naoto Oshiro, and Akinobu Shuto. Predicting accounting fraud: Evidence from japan. *The Japanese Accounting Review*, 6(2016):17–63, 2016.

Akiko Sou. Research on early detection of inappropriate accounting in response to recent changes in economic environment, 2018. URL https://www.fsa.go.jp/frtc/seika/discussion/2017/DP2017-6.pdf.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating ai's ability to replicate ai research, 2025. URL https://arxiv.org/abs/2504.01848.

Hans Van Der Heijden. Predicting industry sectors from financial statements: An illustration of machine learning in accounting research. *The British Accounting Review*, 54(5):101096, 2022.

Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57:47–66, 2016. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2015.09.005. URL https://www.sciencedirect.com/science/article/pii/S0167404815001261.

World Bank. GDP (current US$), 2025. URL https://data.worldbank.org/indicator/NY.GDP.MKTP.CD. Accessed on May 15, 2025.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. PIXIU: A large language model, instruction data and evaluation benchmark for finance, 2023. URL https://arxiv.org/abs/2306.05443.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. FinBen: A holistic financial benchmark for large language models. In *NeurIPS*, volume 37, pages 95716–95743, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/adb1d9fa8be4576d28703b396b82ba1b-Paper-Datasets_and_Benchmarks_Track.pdf.

Siqiao Xue, Xiaojing Li, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. FAMMA: A benchmark for financial domain multilingual multimodal question answering, 2025. URL https://arxiv.org/abs/2410.04526.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. *IJCAI*, 2023.

A. Instruction prompt for each task

Figures 7, 8, 9, and 10 show the prompts used in our experiments.

Prompt for accounting fraud detection

Please analyze the following information extracted from a Japanese company's securities report for any signs of fraudulent activities. Please note that some data may be missing and represented as "-" due to parsing errors. The report has been verified by a certified public accountant, and the numerical values are consistent and correct from a calculation perspective. Therefore, please focus your analysis on non-numerical inconsistencies or logical red flags that could suggest fraud. Respond in the following format: JSON: "'json { "reasoning": "string", "prob": float (between 0 and 1, probability that the report is fraudulent), "prediction": int (0: No fraud, 1: fraud) } "' The extracted securities report is as follows:

Figure 7 | Prompt for accounting fraud detection.

Prompt for earnings forecasting

Please predict whether the "親会社株主に帰属する当期純利益" (Net income attributable to owners of the parent) in the next fiscal year's securities report will increase compared to the current fiscal year, based on the information available in the current year's securities report. - The input is extracted from a Japanese company's securities report. - Some information may be missing and represented as "-" due to parsing errors. - Some attributes are missing and the total does not equal the sum of the parts. Respond in the following format: JSON: "‘json { "reasoning": "string", "prob": float (between 0 and 1, probability that the profit will increase), "prediction": int (0: Decrease, 1: Increase) } "‘ The current year's extracted securities report is as follows:

Figure 8 | Prompt for earnings forecasting.

Prompt for industry prediciton

Based on the following financial report, classify the company into one of the Japanese industry categories. You must classify the company into exactly one of these categories: ["食品", "電気・ガス・エネルギー資源", "建設・資材", "素材・化学", "医薬品", "自動車・輸送機", "鉄鋼・非鉄", "機械", "電機・精密", "情報通信・サービスその他", "運輸・物流", "商社・卸売", "小売", "銀行", "金融(除く銀行)", "不動産"] - The input is extracted from a Japanese company's securities report. - Some information may be missing due to parsing errors. Respond in the following format: JSON: "‘json { "reasoning": str (reasoning for the classification), "prediction": str (predicted industry category in Japanese) } "‘ The current year's extracted securities report is as follows:

Figure 9 | Prompt for industry prediciton.

B. Example of parsed annual reports

The results of extracting BS, CF, PL, Summary, and Text items from the annual report (EDINET ID: E39409, DOC ID: S100VGBV) using edinet2dataset are shown in Figure 11, 12, 13, 14, and 15.

C. Accounting fraud detection

Figure 16 shows the prompt used to assess whether amendment reasons are related to accounting fraud violations. Figure 17 and 18 shows the confusion matrices for each model. Figure 19 and 20 show examples of model responses.

D. Earnings forecasting

Figure 21 and 22 shows the confusion matrices for each model. Figure 23 and 24 shows examples of model response.

E. Industry prediction

Table 9 shows the list of the 33 and 17 class labels we use for industry prediction. Note that as we explained in Section 3.3, we further merged "Energy Resources" and "Electricity & Gas" into a single class in our benchmark due to severe class imbalance.

One dataset construction issue is that SICC will reassign the industry classification when a company's business focus shifts significantly, e.g., when the sales in the new sector increases or when mergers or business transfers cause significant changes. Since EDINET only provides the *latest* industry classification, we may potentially have label noise when using previous reports. Our implementation excludes companies that changed their industry label in the past based on information from the Japan Exchange Group[11].

Figures 25 shows the confusion matrices. Figure 26 and 27 show examples of model response.

[11] https://www.jpx.co.jp/sicc/sectors/index.html

Table 9 | Mapping of SICC 33-sector classification to the TOPIX-17 Series. Note that we further merged “Energy Resources” and “Electricity & Gas” into a single class in our benchmark due to severe class imbalance.

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

Prompt for company name prediciton

Please predict the name of company of the securities report, based on the information available in the current year's securities report. - The input is extracted from a Japanese company's securities report. - Some information may be missing and represented as "-" due to parsing errors. - Some attributes are missing and the total does not equal the sum of the parts. Respond in the following format: JSON: "'json { "reasoning": "string", "prediction": "string" } "' The current year's extracted securities report is as follows:

Figure 10 | Prompt for company name prediciton.

BS

{"現金及び預金": {"CurrentYear": "1935112000"}, "現金及び現金同等物": {"Prior1Year": "1009777000", "CurrentYear": "1871411000"}, "売掛金": {"CurrentYear": "273866000"}, "仕掛品": {"CurrentYear": "5321000"}, "その他": {"CurrentYear": "36103000"}, "貸倒引当金": {"CurrentYear": "-205000"}, "流動資産": {"CurrentYear": "2240284000"}, "建物及び構築物（純額）": {"CurrentYear": "56996000"}, "機械装置及び運搬具（純額）": {"CurrentYear": "5487000"}, "土地": {"CurrentYear": "25695000"}, "減価償却累計額": {"CurrentYear": "-28267000"}, "有形固定資産": {"CurrentYear": "103401000"}, "無形固定資産": {"CurrentYear": "100789000"}, "投資有価証券": {"CurrentYear": "154996000"}, "その他、投資その他の資産": {"CurrentYear": "1784000"}, "投資その他の資産": {"CurrentYear": "263960000"}, "固定資産": {"CurrentYear": "468151000"}, "総資産": {"CurrentYear": "2708435000"}, "買掛金": {"CurrentYear": "18385000"}, "短期借入金": {"CurrentYear": "60000000"}, "１年内返済予定の長期借入金": {"CurrentYear": "263334000"}, "未払法人税等": {"CurrentYear": "33519000"}, "流動負債": {"CurrentYear": "545541000"}, "長期借入金": {"CurrentYear": "435074000"}, "非流動負債": {"CurrentYear": "453872000"}, "負債": {"CurrentYear": "999413000"}, "資本金": {"CurrentYear": "629867000"}, "資本剰余金": {"CurrentYear": "2003537000"}, "利益剰余金": {"CurrentYear": "-924382000"}, "株主資本": {"CurrentYear": "1709021000"}, "純資産": {"Prior1Year": "636232000", "CurrentYear": "1709021000"}, "負債及び純資産": {"CurrentYear": "2708435000"}}

Figure 11 | A balance sheet extracted from a securities report using edinet2dataset.

CF

{"当期利益": {"CurrentYear": "13054000"}, "税引前当期純利益": {"CurrentYear": "18083000"}, "減価償却費及び償却費": {"CurrentYear": "25106000"}, "貸倒引当金の増減額（△は減少）": {"CurrentYear": "23000"}, "受取利息及び受取配当金": {"CurrentYear": "-156000"}, "支払利息": {"CurrentYear": "7331000"}, "売上債権の増減額（△は増加）": {"CurrentYear": "-42526000"}, "棚卸資産の増減額（△は増加）": {"CurrentYear": "-2231000"}, "仕入債務の増減額（△は減少）": {"CurrentYear": "5663000"}, "その他": {"CurrentYear": "-1570000"}, "利息及び配当金の受取額": {"CurrentYear": "156000"}, "利息の支払額": {"CurrentYear": "-7331000"}, "営業キャッシュフロー": {"CurrentYear": "27817000"}, "投資有価証券の取得による支出": {"CurrentYear": "-129996000"}, "投資キャッシュフロー": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "-378161000"}, "長期借入れによる収入": {"CurrentYear": "190000000"}, "長期借入金の返済による支出": {"CurrentYear": "-14812000"}, "財務キャッシュフロー": {"CurrentYear": "1211978000"}, "現金及び現金同等物の増減額": {"CurrentYear": "861634000"}, "現金及び現金同等物": {"Prior1Year": "1009777000", "CurrentYear": "1871411000"}}

Figure 12 | A cash flow statement extracted from a securities report using edinet2dataset.

PL

{"売上高": {"CurrentYear": "1594038000"}, "売上原価": {"CurrentYear": "562526000"}, "売上総利益又は売上総損失（△)": {"CurrentYear": "1031512000"}, "販売費及び一般管理費": {"CurrentYear": "984255000"}, "営業利益": {"CurrentYear": "47256000"}, "受取利息": {"CurrentYear": "156000"}, "その他": {"CurrentYear": "0"}, "営業外収益": {"CurrentYear": "1671000"}, "支払利息": {"CurrentYear": "7331000"}, "営業外費用": {"CurrentYear": "30274000"}, "経常利益": {"CurrentYear": "18653000"}, "特別損失合計": {"CurrentYear": "570000"}, "税引前利益": {"CurrentYear": "18083000"}, "法人税、住民税及び事業税": {"CurrentYear": "7350000"}, "法人税等調整額": {"CurrentYear": "-2322000"}, "法人所得税費用": {"CurrentYear": "5028000"}, "当期利益": {"CurrentYear": "13054000"}, "親会社株主に帰属する当期純利益": {"CurrentYear": "13054000"}}

Figure 13 | A profit and loss statement extracted from a securities report using edinet2dataset.

Summary

{"売上高": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "1594038000"}, "経常利益": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "18653000"}, "親会社株主に帰属する当期純利益": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "13054000"}, "包括利益": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "13054000"}, "純資産額": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "1709021000"}, "総資産額": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "2708435000"}, "１株当たり純資産額": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "333.48"}, "１株当たり当期純利益又は当期純損失": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "2.67"}, "潜在株式調整後１株当たり当期純利益": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "ー"}, "自己資本比率": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "0.631"}, "自己資本利益率、経営指標等": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "0.008"}, "株価収益率": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "244.6"}, "営業活動によるキャッシュ・フロー": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "27817000"}, "投資活動によるキャッシュ・フロー": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "-378161000"}, "財務活動によるキャッシュ・フロー": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "1211978000"}, "現金及び現金同等物の残高": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "1871411000"}, "従業員数": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "121"}, "平均臨時雇用人員": {"Prior4Year": "ー", "Prior3Year": "ー", "Prior2Year": "ー", "Prior1Year": "ー", "CurrentYear": "4"}}

Figure 14 | A summary extracted from a securities report using edinet2dataset.

Text

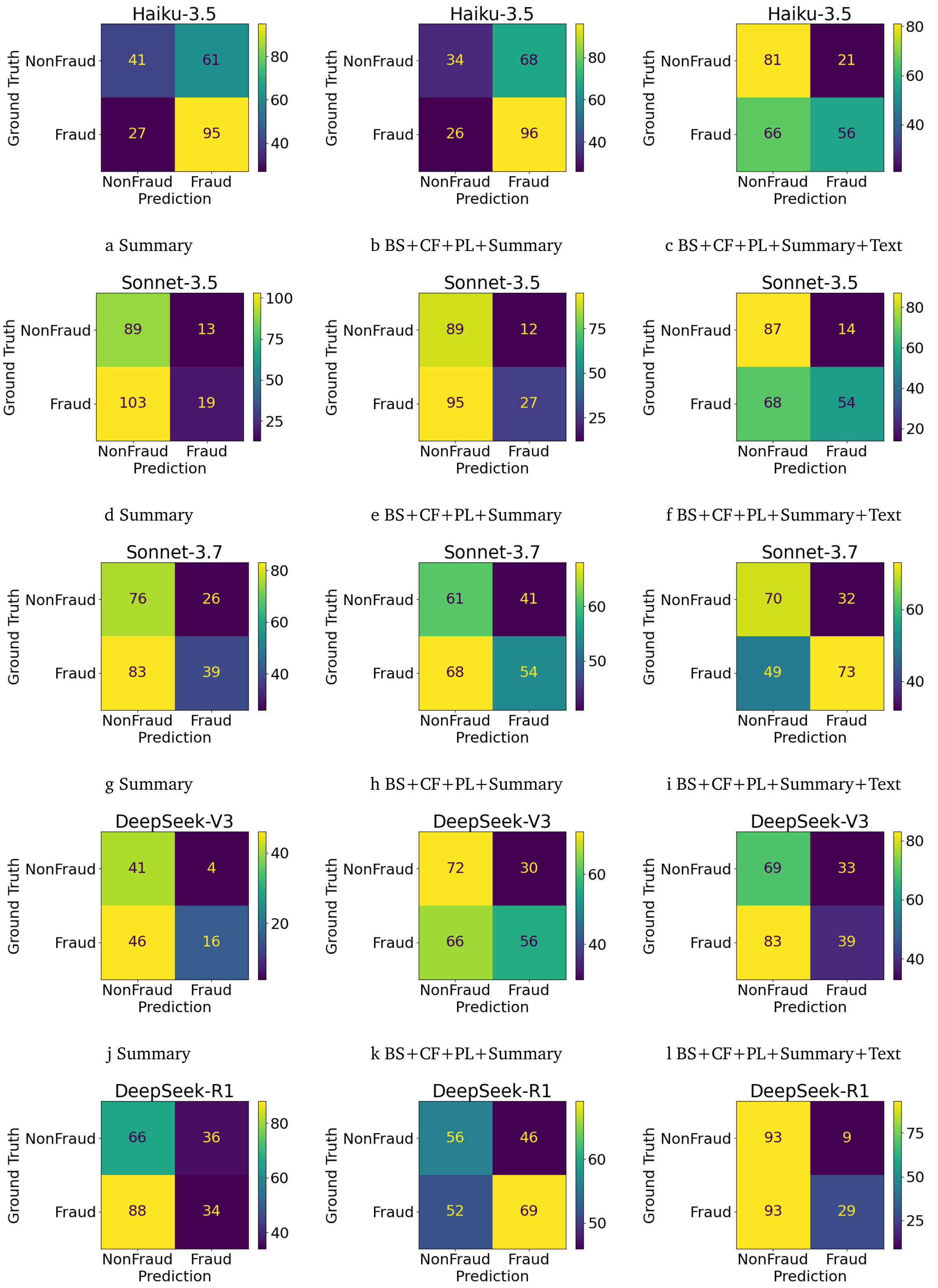
{"沿革": {"FilingDate": " ２【沿革】年月概要2010年９月静岡県焼津市にて資本金5,000千円で当社設立2011年４月感情解析エンジン「Social Emotion Engine」を開発同エンジン搭載のTwitterアプリ「Feel on!」を提供開始2012年２月東京都江東区青海に本社を移転2014年５月東京都千代田区外神田へ本社を移転2014年10月ビジネスチャットサービス「direct」を提供開始2014年12月チャットボット開発環境「daab SDK」を公開2015年10月「direct」に顧客が協力会社と安全につながるオプションサービス「direct Guest Mode（ダイレクトゲストモード）」を提供開始2016年３月徳島県徳島市に徳島ラボを開設2016年10月働き方改革支援ソリューション「direct Smart Working Solution」を提供開始2016年12月情報セキュリティマネジメントシステム（ISMS）のISO/IEC27001及びISO/IEC27017認証を取得2017年１月チャットボットレンタルサービス「direct bot RENTAL」を提供開始2018年５月大阪府大阪市西区に関西支社を開設2018年７月ユーザーの思考に合わせて進化するFAQソリューション「AI-FAQボット」を提供開始2019年５月東京都千代田区岩本町へ本社を移転2019年６月第三者割当増資により総額2.8億円を資金調達2019年10月福岡県福岡市中央区に九州支社を開設2021年５月第三者割当増資により総額約10億円を資金調達2021年10月関西支社を大阪府大阪市西区に移転2022年４月タスク管理、スケジュール管理、掲示板を搭載した「direct Apps」を提供開始2023年３月「direct Apps」に日程調整アプリ「トリスケ」を追加で提供開始2023年６月「タグショット／タグアルバム」を提供開始2024年３月東京証券取引所グロース市場に株式を上場及び公募増資（資本金537,184千円）2024年４月オーバーアロットメントによる売出しに係る第三者割当増資（資本金629,867千円）2024年11月株式会社システム・エムズの株式の取得（子会社化） "},
"事業の内容": {"FilingDate": " ３【事業の内容】当社グループは、当社及び連結子会社１社（株式会社システム・エムズ）の計２社で構成しており、DXソリューション事業を展開しております。当社グループは、「アイデアとテクノロジーで人々を笑顔にする!」をミッションに掲げ、「ココロ踊るチャレンジと心からの感謝にあふれる会社であり続ける」をビジョンとしております。また、役職員向けの行動指針（バリュー）として、「顧客志向」「チャレンジ」「スピード」「チームワーク」「尊重・信頼」「プロフェッショナル」の６つを定義し、徹底した顧客志向のもと、業務課題を深く理解し、課題解決を実現するためにデジタルサービスの開発・提供をしております。社名である「L is B（エルイズビー）」は、働く人々の人生の彩り・潤いになるサービスを提供したいという想いから、Life is Beautifulの頭文字を由来としてい [...]
連結子会社の数及び主要な連結子会社の名称": {"CurrentYear": "(1) 連結子会社の数　1社連結子会社の名称株式会社システム・エムズ株式会社システム・エムズについては、当連結会計年度において新たに株式を取得したため、連結の範囲に含めております。" }}

Figure 15 | Text items extracted from a securities report using edinet2dataset, partially omitted due to space constraints.

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

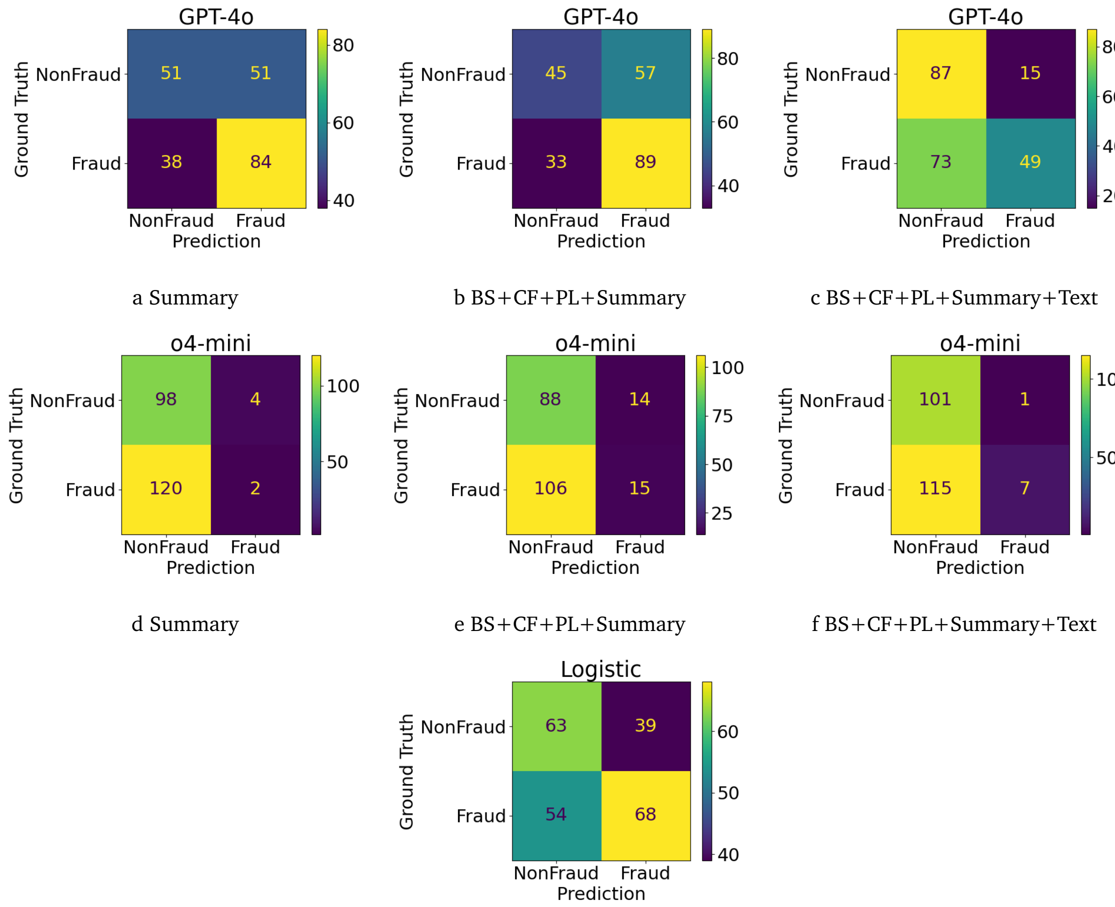
以下のテキストは訂正有価証券報告書の冒頭部分です。この訂正有価証券報告書が不適切会計、粉飾決算、会計不正に関連しているかどうかを判断してください。
特に「提出理由」の部分に着目し、以下のような言葉や表現がある場合は不正会計の可能性が高いと考えられます： - 不適切会計 - 会計不正 - 不正行為 - 粉飾決算 - 会計処理の誤り - 売上の過大計上 - 費用の過少計上 - 資産の過大評価 - 不適切な収益認識 - 監査法人からの指摘 - 社内調査 - 第三者委員会
訂正の理由が単純な記載ミスや軽微な修正ではなく、会計上の重大な問題を示している場合は「Yes」と回答し、詳細な説明を提供してください。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:
(The following text is the beginning of a corrected securities report. Please determine whether this corrected securities report is related to inappropriate accounting, window dressing (fraudulent accounting), or accounting fraud. In particular, focus on the “Reason for Submission” section. If it contains any of the following words or expressions, it is highly likely to indicate fraudulent accounting: - Inappropriate accounting - Accounting fraud - Fraudulent acts - Window dressing - Errors in accounting treatment - Overstatement of sales - Understatement of expenses - Overvaluation of assets - Improper revenue recognition - Indications from the audit firm - Internal investigation - Third-party committee If the reason for the correction is not a simple clerical mistake or minor revision but indicates a serious accounting issue, respond with “Yes” and provide a detailed explanation. Pay special attention to cases where figures in financial statements (such as the balance sheet, income statement, or cash flow statement) have changed. If it is unrelated to accounting fraud or there are no changes to the financial statements, respond with “No” and briefly explain the reason. Please respond in the following JSON format. This JSON must be valid: ‘‘‘json { "is_accounting_fraud": bool # true or false "explanation": str # Provide an explanation of the reason "company_name": str # Company name } ‘‘‘ Make sure your response matches this exact format. Wrap your JSON response with the json tag as shown above. Text of the corrected securities report:)

Figure 16 | Prompt to detect if the reason of the amendment is related to accounting fraud.



a Summary
b BS+CF+PL+Summary
c BS+CF+PL+Summary+Text
d Summary
e BS+CF+PL+Summary
f BS+CF+PL+Summary+Text
g Summary
h BS+CF+PL+Summary
i BS+CF+PL+Summary+Text
j Summary
k BS+CF+PL+Summary
l BS+CF+PL+Summary+Text
m Summary
n BS+CF+PL+Summary
o BS+CF+PL+Summary+Text

Figure 17 | Confusion matrix on accounting fraud detection with Anthropic and DeepSeek models.



a Summary
b BS+CF+PL+Summary
c BS+CF+PL+Summary+Text
d Summary
e BS+CF+PL+Summary
f BS+CF+PL+Summary+Text
g Summary

Figure 18 | Confusion matrix on accounting fraud detection with OpenAI and logistic models.

Accounting fraud detection analysis with Claude 3.7 Sonnet

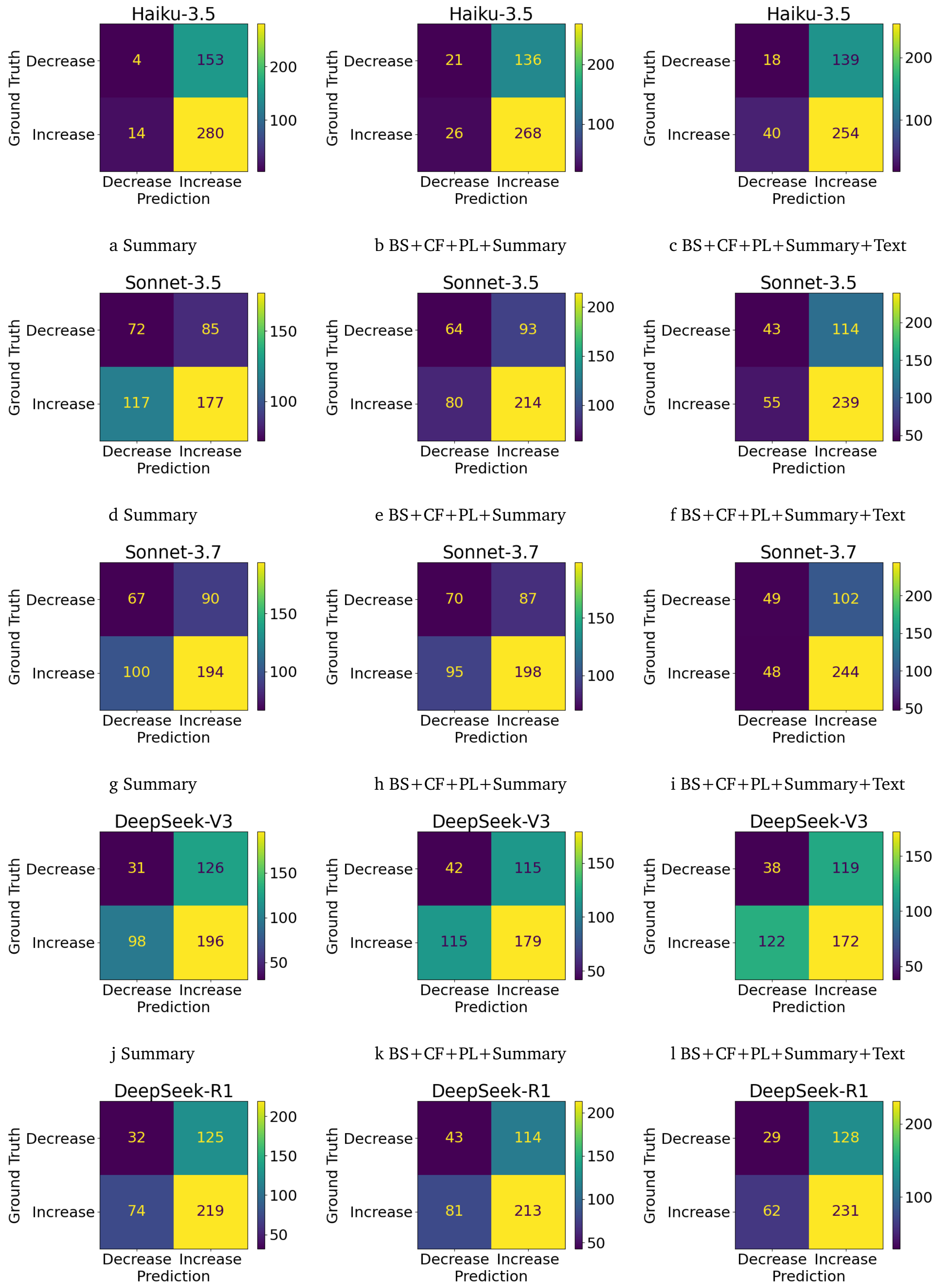
After analyzing the securities report, I've identified several unusual patterns that warrant attention: 1. There's a significant jump in parent company net profit (親会社株主に帰属する当期純利益) from 4,431 million yen in the prior year to 7,468 million yen in the current year (+68.5%), despite operating profit (営業利益) 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 19 | Accounting fraud detection analysis using Claude 3.7 Sonnet based on extracted data (BS, CF, PL, Summary, Text) from the securities report (EDINET Code: E00032, DOC ID: S100LQX5). Ground truth: Non-fraud.

Accounting fraud detection analysis with GPT-4o

The securities report appears to be consistent in terms of numerical data and calculations, as stated. However, there are several non-numerical inconsistencies and logical red flags that could suggest potential fraudulent activities or at least warrant further investigation: 1. **Unusual Increase in Parent Company Net Income**: The parent company’s net income ('親会社株主に帰属する当期純利益') increased significantly in the current year (¥7,468 million) compared to the prior year (¥4,431 million), despite a decline in sales ('売上高') and operating profit ('営業利益'). This discrepancy is partially explained by a large gain on the sale of investment securities (¥4,078 million), but the reliance on non-operating income for profit growth could indicate a lack of sustainable operational performance. 2. **Significant Changes in Comprehensive Income**: Comprehensive income ('包括利益') shows a dramatic increase from a near-zero figure in the prior year (¥-60 million) to ¥18,543 million in the current year. This is primarily driven by gains in other comprehensive income items, such as securities valuation differences and retirement benefit adjustments. While these changes may be legitimate, the magnitude of the shift warrants scrutiny. 3. **Cash Flow from Investing Activities**: The cash flow from investing activities turned positive (¥2,245 million) in the current year, primarily due to the sale of investment securities. This reliance on asset sales for cash flow could indicate liquidity issues or an attempt to boost financial metrics artificially. 4. **Unusual Accounting Practices**: The report mentions the use of 'land revaluation' and 'deferred hedge accounting,' which are complex accounting treatments. While these are not inherently fraudulent, they can be used to manipulate financial results if not applied correctly. 5. **Inconsistent Employee Trends**: The number of temporary employees ('平均臨時雇用人員') has decreased significantly over the years, from 255 in the prior four years to 215 in the current year, despite relatively stable total employee numbers. This could indicate cost-cutting measures or operational downsizing, which are not clearly explained in the report. 6. **Related Party Transactions**: The report mentions significant transactions with related parties, such as subsidiaries and affiliates. While these are disclosed, the terms and conditions of these transactions are not detailed, leaving room for potential manipulation. 7. **Complex Corporate Structure**: The company has a large number of subsidiaries and affiliates, including many overseas entities. This complexity can make it easier to obscure financial irregularities. 8. **Unusual Gains and Losses**: The report includes significant one-time gains (e.g., investment securities sales) and losses (e.g., asset impairments). While these may be legitimate, the timing and magnitude of these items could be used to smooth earnings or meet financial targets. 9. **Lack of Transparency in Certain Areas**: Some sections of the report, such as the details of the '2025 Vision' and the use of funds from asset sales, lack sufficient detail to fully assess their impact on the company’s financial health. While these issues do not conclusively indicate fraud, they highlight areas of concern that should be investigated further.

Figure 20 | Accounting fraud detection analysis using GPT-4o based on extracted data (BS, CF, PL, Summary, Text) from the securities report (EDINET Code: E00032, DOC ID: S100LQX5). Ground truth: Non-fraud.



a Summary b BS+CF+PL+Summary c BS+CF+PL+Summary+Text

d Summary e BS+CF+PL+Summary f BS+CF+PL+Summary+Text

g Summary h BS+CF+PL+Summary i BS+CF+PL+Summary+Text

j Summary k BS+CF+PL+Summary l BS+CF+PL+Summary+Text

m Summary n BS+CF+PL+Summary o BS+CF+PL+Summary+Text

Figure 21 | Confusion matrix on earnings forecasting with Anthropic and DeepSeek models.

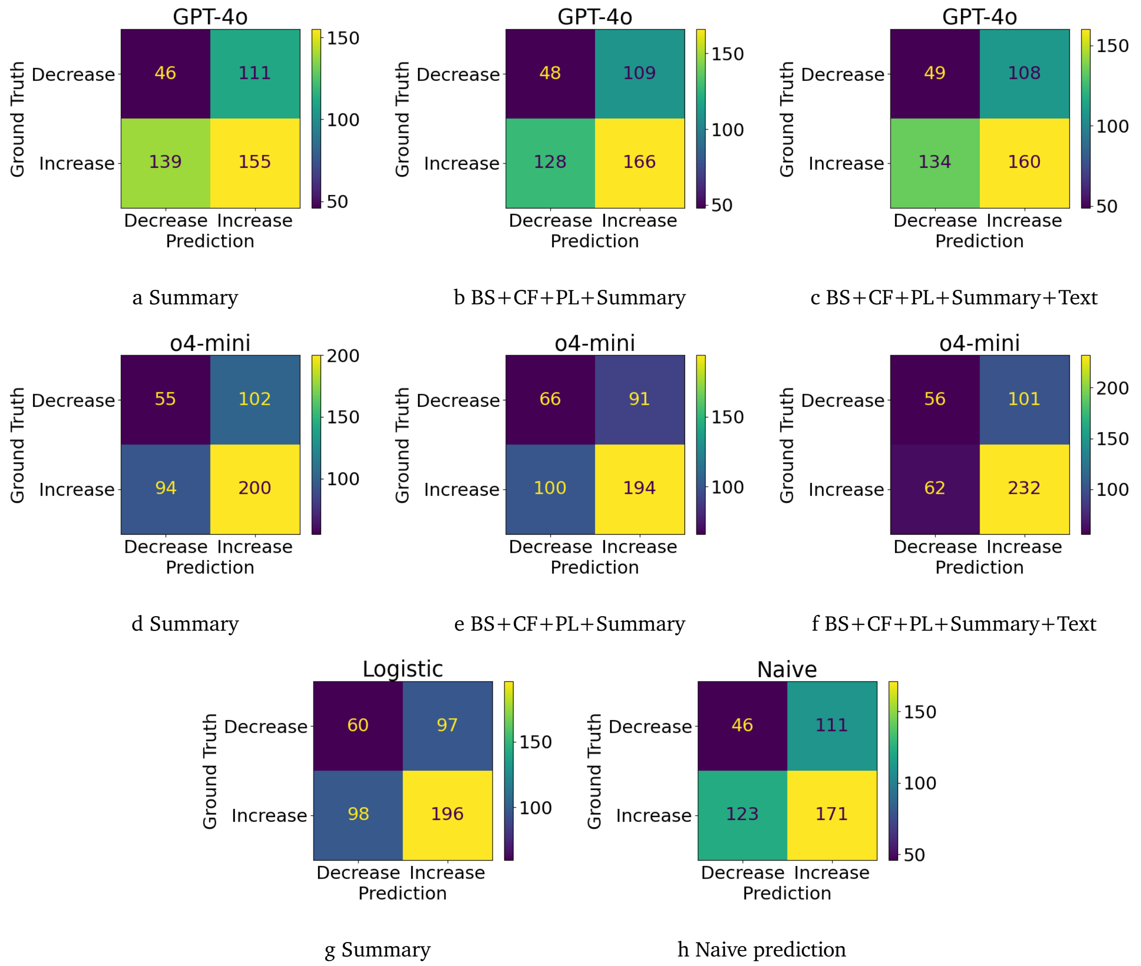


Figure 22 | Confusion matrix on earnings forecasting with OpenAI, logistic, and naive prediction models.

Earnings forecasting analysis with Claude 3.7 Sonnet

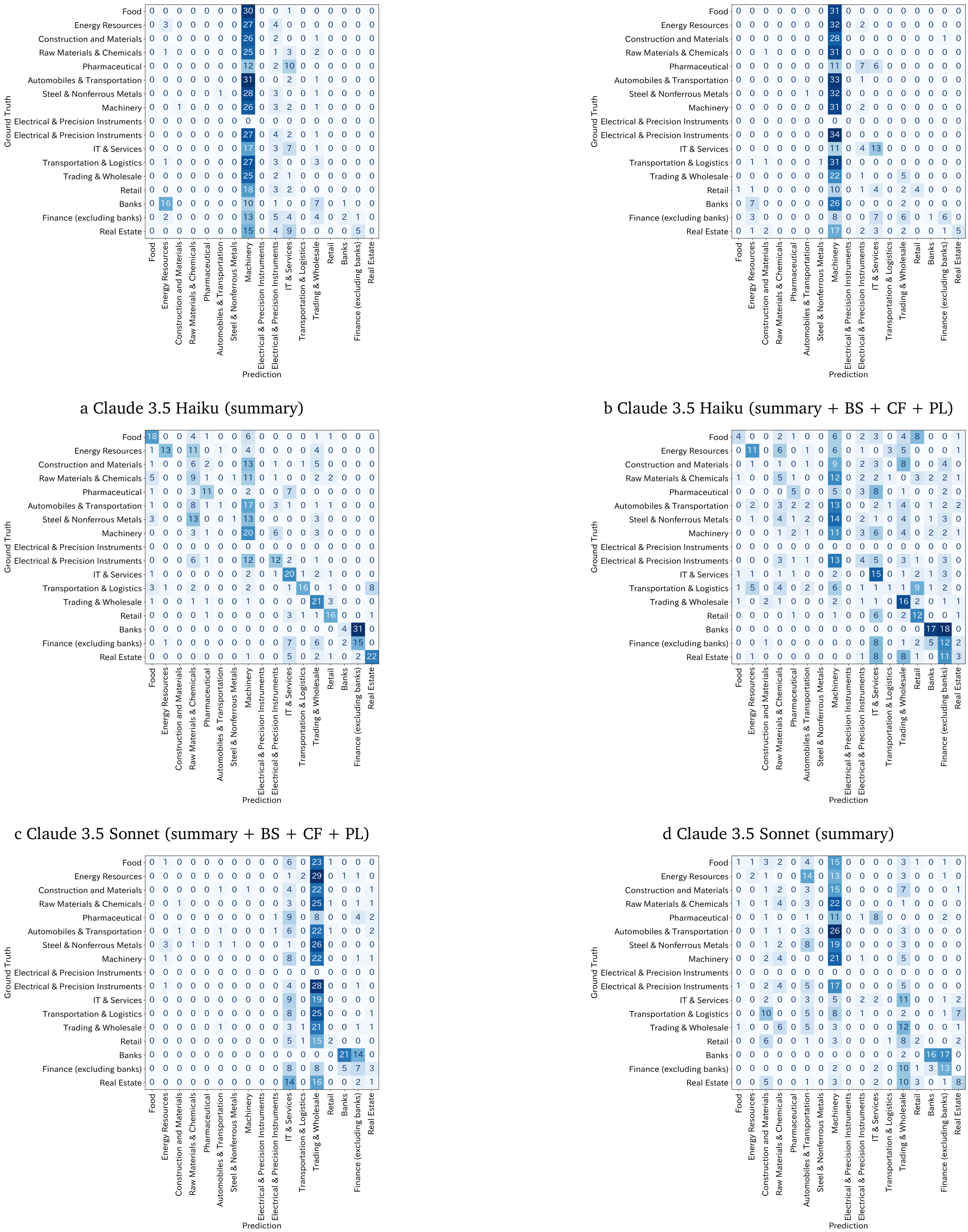
Based on the financial data, I can observe several positive trends that suggest continued profit growth: 1) The company has shown consistent growth in net income over the past 4 years (160,483,000 → 176,451,000 → 217,213,000 → 285,490,000), with a significant 31.4% increase in the most recent year. 2) Sales have been steadily increasing, with a 1.4% growth in the current year. 3) Operating profit increased by 17.6% (327,476,000 → 385,065,000) and ordinary profit increased by 20.0% (330,891,000 → 397,228,000) in the current year. 4) The company has a healthy financial position with increasing net assets (1,974,469,000 → 2,335,993,000) and a strong cash position (2,045,720,000). 5) The company is in the IT solutions industry, which continues to see demand for digital transformation, cloud services, and other technology solutions. 6) The company has been expanding its workforce (605 → 637 employees) to support growth. However, there are some potential concerns: 1) The growth rate of sales has slowed (1.4% vs. previous years' higher growth). 2) Operating cash flow decreased significantly (273,993,000 → 157,958,000). 3) The company mentions potential impacts from COVID-19 on the next fiscal year's performance. 4) There are increasing investments and expenses related to business expansion. Despite these concerns, the overall trend of profitability improvement, the company's strong financial position, and the continued demand in the IT industry suggest that profit growth is likely to continue, albeit possibly at a more moderate pace than the current year's exceptional 31.4% growth.

Figure 23 | Earnings forecasting analysis using Claude 3.7 Sonnet based on extracted data (BS, CF, PL, Summary, Text) from the securities report (EDINET Code: E35466, Doc ID: S100LP1H). Ground truth: Increase.

Earnings forecasting analysis with GPT-4o

The company's financial performance has shown consistent growth in key metrics over the past few years. Revenue ('売上高') has increased steadily from 4,964,794,000 yen three years ago to 6,233,729,000 yen in the current year. Operating profit ('営業利益') and ordinary profit ('経常利益') have also grown significantly, with ordinary profit rising from 241,174,000 yen three years ago to 397,228,000 yen in the current year. Net income attributable to owners of the parent ('親会社株主に帰属する当期純利益') has similarly increased from 160,483,000 yen three years ago to 285,490,000 yen in the current year. Additionally, the company has a strong financial position, with increasing net assets ('純資産額') and a high equity ratio ('自己資本比率'). While there are some risks, such as potential economic fluctuations and challenges in securing human resources, the company's strategic focus on IT solutions, including emerging technologies like AI and cloud services, positions it well for continued growth. Based on these factors, it is highly probable that the net income attributable to owners of the parent will increase in the next fiscal year.

Figure 24 | Earnings forecasting analysis using GPT-4o based on extracted data (BS, CF, PL, Summary, Text) from the securities report (EDINET Code: E35466, Doc ID: S100LP1H). Ground truth: Increase.



a Claude 3.5 Haiku (summary)

b Claude 3.5 Haiku (summary + BS + CF + PL)

c Claude 3.5 Sonnet (summary + BS + CF + PL)

d Claude 3.5 Sonnet (summary)

e GPT-4o (summary)

f GPT-4o (summary + BS + CF + PL)

Figure 25 | Confusion matrix of industry prediction.

Industry prediction analysis with Claude 3.7 Sonnet

Based on the financial data, this appears to be a financial services company, specifically not a bank. Key indicators include: 1) Very large total assets (3.39 trillion yen) compared to revenue (708.5 billion yen), indicating a capital-intensive financial business. 2) Significant borrowings and lending activities - large short-term borrowings (558.1 billion yen), long-term borrowings (976.1 billion yen), and bond issuances (393.5 billion yen). 3) Negative operating cash flow (-112.1 billion yen) with positive financing cash flow (131.8 billion yen), typical of lending businesses. 4) Low self-capital ratio (12.7%), common in financial institutions. 5) The balance sheet structure shows characteristics of a financial services company with large assets and liabilities related to financing activities. The company appears to be involved in lending or leasing operations based on its financial structure, making it most likely a non-bank financial institution.

Figure 26 | Industry prediction analysis using Claude 3.7 Sonnet based on extracted data (BS, CF, PL, Summary) from the securities report (EDINET Code: E05438, Doc ID: S100TQ74). Ground truth: Non-bank financial institution.

Industry prediction analysis with GPT-4o

The financial report indicates a company with significant revenue and profit figures, as well as substantial assets and liabilities. The company also has a high level of cash flow activity, including borrowing and bond issuance, which is characteristic of a trading company. Additionally, the presence of investment securities and activities related to dividends and interest income further supports the classification as a trading company. These characteristics align with the '商社・卸売' (trading and wholesale) industry category in Japan.

Figure 27 | Industry prediction analysis using GPT-4o based on extracted data (BS, CF, PL, Summary) from the securities report (EDINET Code: E05438, Doc ID: S100TQ74). Ground truth: Non-bank financial institution.