Qualitative Insights Meet Quantitative Forecasts with GPT-4o Mini for Taiwan Earnings Prediction

Huei-Wen Teng* Yu-Chuan Shih† Ming-Hsuan Kang‡

February 6, 2026

Abstract

This paper investigates the potential of large language models (LLMs), particularly GPT-4o Mini, in enhancing earnings prediction by transforming structured financial statements into qualitative textual analysis. Using income statements, balance sheets, and cash flow statements from Taiwan-listed firms, we apply GPT-4o Mini with Chain-of-Thought (CoT) reasoning to generate descriptive reports through carefully designed prompts. These textual outputs are then converted into numerical embeddings for predictive modeling. We benchmark GPT-4o Mini's forecasts against classifiers such as Logistic Regression and XGBoost. Our empirical analysis shows that GPT-4o Mini with CoT prompting provides interpretability of numerical values in financial reports. Moreover, augmenting supervised learning methods with text embeddings significantly improves earnings prediction. This study underscores the complementary role of LLMs in financial analysis, bridging qualitative narrative understanding with quantitative forecasting.

Keywords: Earning Prediction; Financial Statement; Large Language Model; Machine Learning

JEL Codes: C30, G17, M41

*Department of Information Management and Finance, National Yang Ming Chiao Tung University. Email: hwteng@nycu.edu.tw

†Department of Information Management and Finance, National Yang Ming Chiao Tung University. Email: ya10742.mg12@nycu.edu.tw

‡Department of Applied Mathematics, National Yang Ming Chiao Tung University. Email: mhkang@math.nctu.edu.tw

1 Introduction

Large Language Models (LLMs) have significantly transformed the landscape of natural language processing (NLP), primarily due to advancements in transformer-based architectures. The Transformer, introduced by Vaswani et al. (2017), employs self-attention mechanisms to process sequential data, overcoming the limitations of recurrent neural networks (RNNs) by efficiently capturing long-range dependencies. This innovation has enabled powerful language comprehension and coherent text generation. Building on this foundation, influential models such as BERT (Devlin et al., 2019), GPT (Achiam et al., 2023), and LLaMA (Touvron et al., 2023) have demonstrated unparalleled ability across diverse NLP tasks, fundamentally reshaping how unstructured text is analyzed and utilized.

The strengths of LLMs extend beyond text generation. They exhibit outstanding linguistic comprehension and contextual reasoning (Brown et al., 2020; Chowdhery et al., 2023), adaptability through zero-shot and few-shot learning (Song et al., 2023), and efficiency in scenarios with limited labeled data (Liu et al., 2023). These capabilities reduce the resources required for model deployment and enable rapid transfer to novel domains. Moreover, LLMs can extract and organize structured information from large text corpora, streamlining tasks such as summarization, data extraction, and retrieval (Ouyang et al., 2022; Le Scao et al., 2023). Collectively, these properties position LLMs as powerful tools for advancing automated analysis and decision-making workflows.

Such strengths have led to diverse applications across NLP domains. In Question Answering (QA), LLMs deliver precise context-aware responses (Lewis et al., 2020). In sentiment analysis, they capture emotional and opinion-driven insights from large-scale text corpora (Yang et al., 2020). For named-entity recognition (NER), they accurately identify and categorize entities within unstructured data (Wang et al., 2023). However, despite these successes, LLMs face challenges in modeling time-series data, where sequential dependencies complicate forecasting (Chang et al., 2025; Yu et al., 2023).

Finance represents a natural domain for applying these advanced techniques. Financial markets are global, multi-layered, and highly dynamic, demanding real-time interpretation

and decision-making under uncertainty (Mishkin, 2007). LLMs can provide adaptability and interpretability suited to this complexity. Prior studies show their promise across financial applications: automating extraction of insights from statements and disclosures (Kim et al., 2024), supplementing numerical forecasting with qualitative analysis for earnings prediction (Yu et al., 2023), enabling early detection of risks (Joshi, 2025; Pele et al., 2024), supporting investment advisory with sentiment-driven recommendations, and enhancing compliance monitoring. These findings underscore the potential of LLMs to bridge human-like reasoning with quantitative rigor in financial analytics (Li et al., 2023a; Zhao et al., 2024).

Within this context, we focus on GPT-4o Mini, a smaller and computationally efficient variant of GPT-4. As a transformer-based model (Vaswani et al., 2017), GPT-4o Mini is trained on diverse text corpora to perform generation, summarization, and question answering (Brown et al., 2020; Achiam et al., 2023). Its smaller size improves scalability, inference speed, and cost-effectiveness, making it well suited for repeated experimentation on large sets of s (Wu et al., 2023; Zhao et al., 2024). These features motivate our choice of GPT-4o Mini as a practical model for evaluating LLM applications in financial forecasting.

This paper makes two contributions to the literature on LLMs in finance. First, we propose a two-stage prompting framework with GPT-4o Mini that simulates the workflow of a financial analyst: multi-period accounting data are converted into qualitative reports using semantic descriptors (e.g., "robust growth," "tight liquidity"), and then extended into explicit forward-looking forecasts of EPS direction $(y_{i,t+1})$, justified through causal reasoning, risks, and opportunities (Emett and Nelson, 2017; Nie et al., 2024; Liu et al., 2023; Zhu et al., 2023; Zhao et al., 2024; Kim et al., 2024; Gupta, 2023; Lee et al., 2024; Yu et al., 2023).

Second, we transform these reports into semantic embeddings using OpenAI's *text-embedding-3-small* model (Khan et al., 2022), apply dimensionality reduction with PCA (Zebari et al., 2020; Musil, 2019), and evaluate their predictive value through classification and regression tasks, comparing embeddings alone and in combination with structured accounting variables. Together, these contributions provide a novel framework to test whether structured prompting and embedding-based modeling enable LLMs to extract decision-relevant semantics from numerical financial statements, bridging qualitative interpretation and quantitative forecasting in

the context of Taiwan earnings prediction.

The remainder of this paper is organized as follows. Section 2 describes the data and feature construction. Section 3 introduces the structured two-stage prompt framework and illustrates examples of generated reports. Section 4 presents the embedding-based predictive modeling design and empirical analysis. Finally, Section 5 concludes with key findings and discusses implications for integrating LLMs into financial forecasting and statistical research.

2 Data

In this section, we describe the dataset and feature construction underlying our empirical framework. Section 2.1 provides an overview of the data source, sample coverage, and the distribution of firms across industry categories. Section 2.2 then details the feature selection and preprocessing procedures that lead to the final set of variables used in prompt construction and subsequent predictive modeling.

2.1 Overview

Our dataset is obtained from the Taiwan Economic Journal (TEJ) and contains quarterly balance sheets, income statements, and cash flow statements for all publicly listed Taiwanese firms. The sample spans 68 quarters from 2008 to 2024 and initially includes 143 accounting items. Table 2.1 reports the seven industry categories and the corresponding proportions of firms in [market capital?] in our sample.

Table 2.1. Industry categories, sub-industries, and [market capitalization proportions] of firms in the sample

Index	Industry Category and Sub-Industries	Proportion
1	Information and Electronics: Semiconductors, Electronic Components, Optoelectronics, Computers and Peripherals, Other Electronics, Telecommunications, Electronic Distribution, Electronic Manufacturing, E-commerce, Digital Cloud, Information Services	46.40%
2	Consumer and Retail: Food Industry, Trading and Department Stores, Home and Living, Sports and Leisure, Tourism and Hospitality, Automotive Industry	17.32%
3	Traditional Manufacturing: Electrical Machinery, Chemical Industry, Textiles and Fibers, Steel Industry, Plastics, Rubber, Electrical Cables, Glass and Ceramics, Paper Manufacturing, Cement Industry	14.35%
4	Finance and Services: Financial Industry, Cultural and Creative Industries, Information Services	7.57%
5	Construction and Infrastructure: Building Materials and Construction, Utilities and Energy, Green Energy and Environmental Protection	5.93%
6	Agriculture and Others: Agricultural Technology, Others	5.45%
7	Transportation and Logistics: Shipping Industry, Automotive Industry	2.97%

2.2 Feature selection and preprocessing

In preliminary analyses, we find that using the full set of 143 accounting items does not lead to satisfactory performance in EPS prediction. Moreover, to avoid exceeding the token input limits of GPT-based models and to improve model parsimony, we follow prior literature and select a subset of 32 accounting variables for our EPS prediction framework. These variables

are chosen to represent four fundamental dimensions of firm performance: (i) liquidity, (ii) capital structure, (iii) operational performance, and (iv) profitability (Kim et al., 2024).

First, liquidity-related variables are included based on their well-documented importance in assessing earnings quality and forecasting future profitability. Prior studies, such as Richards and Laughlin (1980) and Deloof (2003), emphasize that liquidity conditions capture a firm's short-term financial health and operational efficiency, thereby providing relevant information for earnings per share (EPS) prediction. Accordingly, the selected features x_1–x_{11} and x_{29} represent core liquidity measures, including cash and cash equivalents (CCE), current financial assets across valuation categories (FA_FVTPL_C, FA_FVOCI_C, FA_AC_C, HFA_C), contract assets (CA_C), notes and accounts receivable (AR), other receivables (OR), short-term loans to others (STL), inventories (INV), total current assets (TCA), and total receivables (ARtot).

Second, capital structure variables are incorporated based on evidence that balance-sheet composition and leverage contain predictive information for future earnings. Cao and You (2024) show that high-dimensional accounting indicators related to assets and liabilities improve earnings forecasting, while Pabuccu and Barbu (2024) demonstrate that leverage-related variables selected through hybrid feature selection contribute significantly to earnings predictability. These variables characterize a firm's capital base, financing structure, and long-term solvency, which are central to assessing financial risk. Accordingly, the features x_{12}–x_{18}, together with x_{25}–x_{28}, capture the firm's asset, liability, and equity structure, including total non-current assets (TNCA), total assets (TA), short-term borrowings (STB), current and non-current liabilities (TCL, TNCL), total liabilities (TL), total equity (TE), as well as aggregated current and non-current financial assets (CFA, NCFA??) and financial liabilities (CFL, NCFL).

Third, operational performance variables are included based on evidence that core profitability measures provide reliable signals for future earnings. Hunt et al. (2022) show that metrics such as gross profit margins and operating income are strong predictors of subsequent earnings outcomes, underscoring their relevance for EPS modeling. These variables reflect a firm's core business activities and operating efficiency. Accordingly, the operational performance group comprises features x_{19}–x_{22}, including net revenue (NR), cost of goods sold (COGS), gross profit (GP), and operating income (OI).

Fourth, profitability-related variables are included as they directly measure a firm's earnings performance and shareholder returns. Prior studies emphasize that profit-based metrics are central to understanding earnings persistence and firm valuation, and therefore play a critical role in earnings forecasting models (Sloan and Wang, 2024). Accordingly, profitability indicators are captured by features x_{23}, x_{24}, and y, which include net income attributable to owners of the parent (PL_PARENT), basic earnings per share (EPS_BASIC), and earnings per share (EPS), where the latter serves as the target variable.

Finally, we derive three commonly used financial ratios from the raw financial statements to summarize firm-level profitability, leverage, and shareholder returns: Gross Profit Margin (GPM, x_{30}), Debt Ratio (DR, x_{31}), and Return on Equity (ROE, x_{32}). These ratios are widely adopted in both academic research and practice as parsimonious indicators of operating efficiency, financial leverage, and equity performance (Richards and Laughlin, 1980; Gao et al., 2024). GPM reflects the profitability of a firm's core operations by measuring how efficiently revenue is converted into gross profit, and is defined as

$$\text{GPM} = \frac{\text{Gross Profit}}{\text{Net Revenue}}.$$

The DR measures the proportion of a firm's assets financed by liabilities, with higher values indicating greater leverage and financial risk. It is calculated as

$$\text{DR} = \frac{\text{Total Current Liabilities} + \text{Total Non-current Liabilities}}{\text{Total Assets}}.$$

The ROE evaluates a firm's efficiency in generating profits for its shareholders and serves as a key indicator of return on invested equity:

$$\text{ROE} = \frac{\text{Net Income Attributable to Owners of the Parent}}{\text{Total Shareholders' Equity}}.$$

Together, these ratio-based measures complement the level-based accounting variables and enhance the interpretability of firm fundamentals. The complete set of 33 accounting items is summarized in Table 2.2.

Table 2.2. Selected accounting items

Acronym	Original Variable	NaN (%)	Notation
CCE	Cash & Cash Equivalents	0.01	x_1
FA_FVTPL_C	Financial Assets at FVTPL (Current)	0.01	x_2
FA_FVOCI_C	Financial Assets at FVOCI (Current)	2.36	x_3
FA_AC_C	Financial Assets at Amortised Cost (Current)	0.01	x_4
HFA_C	Hedging Financial Assets (Current)	2.36	x_5
CA_C	Contract Assets (Current)	4.16	x_6
AR	Notes and Accounts Receivable	0.07	x_7
OR	Other Receivables	0.01	x_8
STL	Short-term Loans to Others	4.23	x_9
INV	Inventories	3.91	x_{10}
TCA	Total Current Assets	2.97	x_{11}
TNCA	Total Non-current Assets	3.16	x_{12}
TA	Total Assets	0.00	x_{13}
STB	Short-term Borrowings	0.35	x_{14}
TCL	Total Current Liabilities	2.97	x_{15}
TNCL	Total Non-current Liabilities	3.16	x_{16}
TL	Total Liabilities	0.00	x_{17}
TE	Total Equity	0.00	x_{18}
NR	Net Revenue	0.00	x_{19}
COGS	Cost of Goods Sold	2.34	x_{20}

(continued on next page)

Acronym	Original Variable	NaN (%)	Notation
GP	Gross Profit	2.34	x_{21}
OI	Operating Income	2.34	x_{22}
PL_PARENT	Profit (Loss) Attributable to Owners	0.00	x_{23}
EPS_BASIC	Basic Earnings per Share	0.06	x_{24}
CFA	Current Financial Assets (Total)	0.00	x_{25}
NCFA	Non-current Financial Assets (Total)	0.00	x_{26}
CFL	Current Financial Liabilities (Total)	0.00	x_{27}
NCFL	Non-current Financial Liabilities (Total)	0.00	x_{28}
ARtot	Receivables (Total)	0.00	x_{29}
GPM	Gross Profit Margin (Derived)	0.00	x_{30}
DR	Debt Ratio (Derived)	0.00	x_{31}
ROE	Return on Equity (Derived)	0.00	x_{32}
EPS	Earnings per Share	0.00	y

For data preprocessing, we note that EPS volatility differs substantially across industries, with technology-related firms exhibiting greater variability than utilities or financial firms. Despite this heterogeneity, we adopt a unified modeling framework across all industries, rather than estimating sector-specific models, in order to preserve comparability and enhance generalizability. To further control for cross-firm and cross-industry scale differences, we standardize EPS changes using firm-level z-score normalization. This preprocessing step allows the model to focus on relative earnings dynamics within each firm, rather than on absolute magnitude differences driven by firm size, capital structure, or reporting scale. As a result, the standardized inputs improve training stability and facilitate the identification of meaningful temporal patterns in earnings behavior.

2.3 Teng: Aggreated items?

[Teng: Do we need this subsetion??? I found some overlapping variables.] To reduce data dimensionality and improve feature interpretability, we aggregate certain accounting items based on their economic similarity and functional classification. Many financial statement items are highly correlated or represent similar financial constructs (e.g., various types of current financial assets or liabilities). Aggregating such variables helps improve model generalizability, avoids overfitting, and mitigates data sparsity problems for rarely reported items. Specifically, we consolidated individual line items into broader categories such as current assets, non-current assets, current liabilities, non-current liabilities, and shareholders' equity in Table 2.3. This structured grouping not only reduces the complexity of financial input features, but also aligns with standard financial reporting hierarchies.

Table 2.3. Mapping of aggregated accounting items [Teng: I don't understand why this table is necessry? It seems that Current Financial Assets x_{25} is overlapped with FA_FVTPL_C etc.]

Aggregated Item	Original Items
Current Financial Assets	Cash and Cash Equivalents; Financial Assets at FVTPL (Current); Financial Assets at FVOCI (Current); Financial Assets at Amortised Cost (Current); Hedging Financial Assets (Current)

(continued on next page)

(continued from previous page)

Aggregated Item	Original Items
Current Liabilities	Short-term Borrowings; Commercial Paper / Bankers' Acceptances Payable; Financial Liabilities at FVTPL (Current); Hedging Financial Liabilities (Current); Financial Liabilities at Amortised Cost (Current); Contract Liabilities (Current); Notes and Accounts Payable; Other Payables; Current Tax Liabilities; Provisions (Current); Liabilities Associated with Assets Held for Sale; Lease Liabilities (Current); Current Portion of Long-term Borrowings; Preferred Stock Liabilities (Current)
Non-current Financial Assets	Financial Assets at FVTPL (Non-current); Financial Assets at FVOCI (Non-current); Financial Assets at Amortised Cost (Non-current); Hedging Financial Assets (Non-current)
Non-current Liabilities	Financial Liabilities at FVTPL (Non-current); Hedging Financial Liabilities (Non-current); Financial Liabilities at Amortised Cost (Non-current); Contract Liabilities (Non-current); Preferred Stock Liabilities (Non-current); Bonds Payable (Non-current); Bank Loans (Non-current); Other Long-term Borrowings; Lease Liabilities (Non-current); Provisions (Non-current); Deferred Credits; Net Defined Benefit Liability; Deferred Tax Liabilities
Total Receivables	Notes and Accounts Receivable; Other Receivables; Contract Assets (Current)

3 Structured prompt-based report generation

We develop a two-stage prompting framework with GPT-4o Mini to link historical accounting information to forward-looking earnings forecasts. In the first stage, five consecutive quarters of firm-level financial variables are concatenated to form the model inputs, thereby providing one year of contextual financial history. Specifically, the sequence of accounting vectors $(x_{i,t-4}, \ldots, x_{i,t})$ is synthesized into structured qualitative financial reports; see Figure 3.1 for an illustration. This rolling-window design is essential for capturing seasonal effects, earnings cycles, and other temporal dynamics that are relevant for prediction. In the second stage, these narrative reports are extended with explicit predictions of next-period EPS direction $(y_{i,t+1})$, thereby linking interpretability with predictive inference.

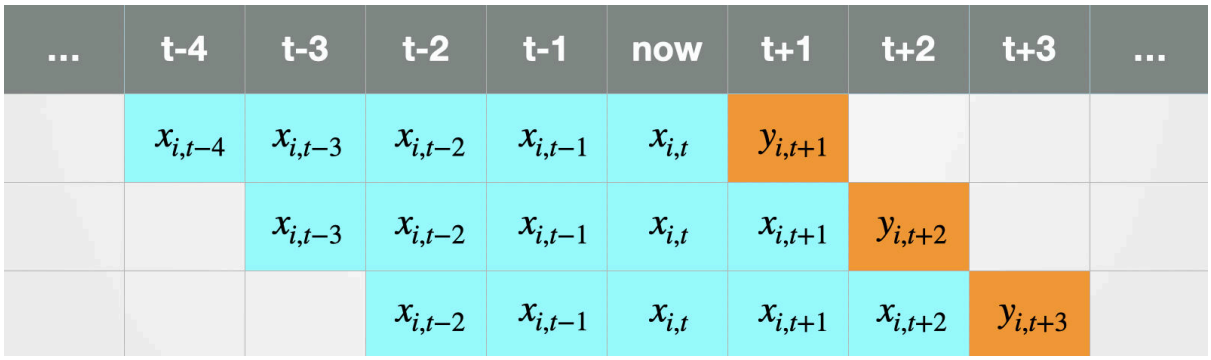


Figure 3.1: Data combination

3.1 First-stage prompt: Generating financial analysis reports

The first stage of our framework transforms multi-period accounting data into structured qualitative reports. Instead of citing raw numerical values, the prompt requires the model to use semantic descriptors (e.g., "robust growth," "tight liquidity") that mimic the interpretive style of human analysts. This design enhances interpretability, generalization across firms and time periods, and downstream usability in predictive modeling.

The prompt is modular, consisting of nine sections designed to simulate the workflow of financial analysts:

1. Company Background and Industry Overview: firm identity, core products, and industry environment (Kleinman, 2001).

2. Revenue and Growth: qualitative evaluation of revenue trajectory (e.g., expansion or stagnation) (Wu et al., 2023).

3. Costs and Profitability: assessment of margins and cost pressures (Huang et al., 2023).

4. Asset and Liability Structure: financial stability, equity base, and leverage (White et al., 2002).

5. Cash Flow: liquidity evaluation across operating, investing, and financing activities (Francis et al., 2005).

6. Risks and Weaknesses: identification of downside factors such as cost pressures or regulatory exposure.

7. Strengths and Opportunities: identification of competitive advantages and growth potential (Yang et al., 2024).

8. Overall Evaluation and Forward-Looking View: integrated assessment with explicit optimism or pessimism.

9. Chain-of-Thought Summary: causal reasoning linking financial drivers to EPS outlook (Wei et al., 2022; Chen et al., 2022).

This design ensures systematic coverage of profitability, leverage, liquidity, risks, and opportunities. It encourages the model to produce evaluative commentary—rather than neutral description—and to synthesize forward-looking judgments. Prior literature supports this modular approach: analyst-style qualitative reports improve interpretability (Kleinman, 2001; Wu et al., 2023; Huang et al., 2023); traditional financial analysis frameworks highlight balance sheet and cash flow assessment (White et al., 2002; Francis et al., 2005); and causal reasoning improves inference in LLMs (Wei et al., 2022; Chen et al., 2022). Below we present the exact prompt design.

Prompt (Stage 1)

Please generate a comprehensive and in-depth financial analysis report based on the following multi-period financial statement data.
The primary goal is to evaluate the company's current financial and operating condition in preparation for forecasting the change in EPS (Earnings Per Share) in the next period.
Avoid quoting any specific numerical values. Instead, describe trends using qualitative expressions

such as:

- "robust growth" or "noticeable decline" for revenue
- "stable profitability" or "profit erosion" for earnings
- "strong cash flow" or "tight liquidity" for liquidity
- "large-scale assets" or "increasing leverage" for the balance sheet

Each section should clearly explain both positive and negative aspects where applicable, and reflect on how these factors may affect future EPS. Also, provide your own analysis and judgment—not just neutral descriptions. Use expressions like "we believe," "it appears likely," or "this suggests that." Please provide your insights—both positive and negative—in each section.

1. Company Background and Industry Overview: Describe the company's size, core products or services, and market positioning. Discuss the industry environment, including competition, regulatory trends, and macroeconomic conditions. Reflect on how these external factors may influence the company's EPS potential.
2. Revenue and Growth: Qualitatively describe the operating revenue trends across periods. Is it showing "modest expansion," "robust growth," or "signs of stagnation"? Explain possible causes and evaluate how top-line trends may impact future profitability.
3. Costs and Profitability: Evaluate gross profit, operating profit, and net income. Use qualitative phrases such as "improving margin," "high cost pressure," or "declining net profit." Offer your view on whether EPS is "steadily increasing," "volatile," or "under pressure."
4. Asset and Liability Structure: Assess total assets, shareholders' equity, and liabilities. Use expressions such as "capital adequacy," "moderate leverage," or "rising debt burden." Discuss whether the financial structure supports or threatens earnings stability.
5. Cash Flow Situation: Evaluate cash flows from operating, investing, and financing activities. Use terms like "ample liquidity," "unstable financing," or "reinvestment capacity." Explain how liquidity may affect the company's flexibility in sustaining EPS.
6. Risks and Weaknesses: Identify any emerging risks—e.g., margin pressure, customer concentration, regulatory exposure. Comment on how these risks may impact earnings or introduce volatility to EPS.
7. Strengths and Opportunities: Highlight competitive advantages, operational efficiencies, or strategic growth opportunities. Explain how these strengths may contribute to future EPS resilience or growth.
8. Overall Evaluation and Forward-Looking View This is a crucial section—please provide detailed analysis and your personal opinion. Use your expertise to highlight key findings, confidence level, potential risks, and the company's advantages and disadvantages. Summarize clearly whether you are optimistic or pessimistic about the company's financial future—not neutral. Conclude with a statement such as:
    - "These trends point toward a potentially improving EPS outlook."
    - "Despite some strengths, downside risks to EPS remain significant."
9. Chain of Thought Summary: Provide at least 2–3 bullet points that capture your reasoning path (cause $\rightarrow$ effect), such as:
    - Revenue rebound $\rightarrow$ Cost absorption $\rightarrow$ Higher profit $\rightarrow$ EPS growth
    - Margin pressure $\rightarrow$ Profit erosion $\rightarrow$ EPS stagnation
    - High leverage $\rightarrow$ Interest burden $\rightarrow$ EPS under pressure

We also provide an illustrative report generated by GPT-4o Mini for a semiconductor firm.

LLM Output (Illustrative Example)

Financial Analysis Report

1. Company Background and Industry Overview The company, a leading player in the semiconductor industry, has established a significant market presence through its innovative technology and extensive product offerings. It operates in a highly competitive environment characterized by rapid technological advancements and stringent regulatory standards. The semiconductor sector is currently experiencing robust demand driven by the proliferation of digital devices and emerging technologies such as AI and IoT. However, macroeconomic factors, including supply chain disruptions and geopolitical tensions, could pose challenges. We believe these external pressures may introduce volatility to earnings, potentially impacting EPS growth in the upcoming periods.

2. Revenue and Growth The company's revenue trajectory indicates noticeable growth over recent periods, reflecting strong demand for its products. This robust expansion can be attributed to increased market penetration and strategic partnerships. However, there are signs of potential stagnation as competition intensifies and pricing pressures mount. If these trends continue, they could hinder future revenue growth, impacting overall profitability and subsequently EPS.

3. Costs and Profitability While gross profit margins have shown signs of improvement, operating costs have also escalated due to higher input prices and increased R&D expenditures. This high cost pressure has led to a mixed performance in net profit, with some periods indicating profit erosion. Consequently, EPS appears to be under pressure as the company navigates these challenges. We believe that sustained focus on cost management will be crucial for stabilizing earnings in the near term.

4. Asset and Liability Structure The company's balance sheet reflects a solid asset base with substantial shareholders' equity, indicating capital adequacy. However, there is an increasing leverage trend that raises concerns about financial stability. The rising debt burden could lead to higher interest expenses, which might constrain future profitability and EPS growth. A careful assessment of capital structure will be essential to mitigate risks associated with leverage.

5. Cash Flow Situation Cash flow from operating activities remains strong, providing ample liquidity for operational needs and potential reinvestment opportunities. However, cash flow from investing activities shows signs of instability due to fluctuating capital expenditures related to technology upgrades. This suggests that while liquidity is currently sufficient, future investments may need careful management to ensure they do not adversely affect cash reserves or operational flexibility.

6. Risks and Weaknesses Emerging risks include margin pressure from rising raw material costs and potential regulatory scrutiny as the industry evolves. Additionally, customer concentration poses a risk if key clients face challenges or shift their purchasing strategies. These factors could introduce volatility to EPS, making it imperative for the company to diversify its customer base and enhance operational resilience.

7. Strengths and Opportunities The company's competitive advantages lie in its advanced technology capabilities and strong brand reputation within the semiconductor space. Operational efficiencies achieved through innovation present opportunities for margin expansion in the future. We believe that capitalizing on these strengths will be vital for sustaining EPS growth amidst competitive pressures.

8. Overall Evaluation and Forward-Looking View In summary, while the company demonstrates several strengths—such as robust revenue growth and strong cash flow—there are notable risks that could impede earnings stability. Our analysis suggests that despite a favorable market position, downside risks related to cost pressures and leverage remain significant concerns for future EPS performance. Therefore, we maintain a cautiously optimistic outlook on the company's financial future. Conclusion: These trends point toward a potentially improving EPS outlook if the company can effectively manage its cost structure while leveraging its competitive advantages.

9. Chain of Thought Summary

    - Robust revenue growth → High cost pressure → Profit erosion → EPS under pressure.
    - Strong cash flow → Ample liquidity → Flexibility for reinvestment → Potential for EPS recovery.
    - Increasing leverage → Rising debt burden → Interest expenses → Risk to EPS stability.

3.2 Second-stage Prompt: Producing prediction reports

In the second stage, the qualitative report is extended with an explicit forecast of next-period EPS direction $(y_{i,t+1})$. The prompt requires the model to articulate causal drivers, risks, and opportunities, thereby shifting the task from descriptive interpretation to predictive inference. Specifically, the model receives the first-stage report and is instructed to take a clear stance—predicting an increase, decrease, or stability in EPS (Cheong and Thomas, 2018)—while structuring the response into five sections: Prediction Summary, Key Drivers, Risks and Weaknesses, Confidence Level, and Conclusion. For example, the model concluded that EPS would likely remain stable, citing strong revenue and cash flow as stabilizing forces, but also identified cost pressures and rising leverage as offsetting risks, assigning a "Moderate" confidence level based on this balance.

This design serves multiple purposes. First, it prevents vague or neutral language, countering the tendency of LLMs to default to neutrality in high-stakes settings. Second, it mimics the tone and logical flow of equity analyst commentary and forward guidance (Li, 2008), encouraging outputs consistent with real-world earnings forecasts. Third, the prompt requires an emotionally grounded tone (e.g., "cautiously optimistic," "strongly bearish") and causal reasoning, aligning with evidence that investor sentiment, polarity, and confidence expressions carry predictive value in financial communication (Yang et al., 2020; Lakkaraju et al., 2023). This layered design tests whether LLMs can generate judgment-oriented textual representations

from numerical data that encode decision-useful signals not explicitly present in raw financial statements.

Prompt (Stage 2)

Based on the following financial analysis report, forecast whether the company's EPS (Earnings Per Share) will increase, decrease, or remain stable in the next period.
You must express a clear and opinionated view. Avoid being neutral or indecisive. Take a firm stance and support it with reasoning.
However, do not default to a pessimistic or overly cautious tone. If the company shows signs of strength, stability, or strategic improvement, you should clearly and confidently reflect that in your forecast.
Please keep your answer between 250 and 400 words, focusing on rich semantic content, clarity of reasoning, and emotional tone. This output will be converted into a text embedding for machine learning analysis, so structure and consistency are important.
Structure your response using the following five sections:

1. Prediction Summary:Clearly state: Do you expect EPS to increase, decrease, or remain stable?
2. Key Drivers: What are the main reasons behind this prediction? Mention specific financial trends or causal factors.
3. Risks and Weaknesses: Identify potential risks, uncertainties, or structural issues that could affect earnings. Be specific.
4. Confidence Level: State your confidence level (High / Moderate / Low) and explain your reasoning. Confidence can be high if indicators are consistent, or lower if signals are mixed.
5. Conclusion: End with a strong and emotionally grounded sentence that summarizes your outlook. Examples:
   - "We are strongly optimistic about EPS growth given the company's margin recovery and top-line acceleration."
   - "Stable cost control and solid demand suggest EPS will likely remain resilient."
   - "Despite some stabilization, high debt and cost uncertainty make the EPS outlook fragile."
6. Example reasoning paths::
   - Revenue rebound $\rightarrow$ Margin expansion $\rightarrow$ Higher net income $\rightarrow$ EPS increase
   - Stable revenue + cost control $\rightarrow$ Flat earnings $\rightarrow$ EPS steady
   - Margin erosion + rising debt $\rightarrow$ Profit pressure $\rightarrow$ EPS decline

Avoid rephrasing the original report. Instead, synthesize it and form your own judgment based on causal logic.

The following illustrates the model's verbatim response to the second-stage prompt.

LLM Output (Illustrative Example)

1. Prediction Summary: I expect the company's EPS to remain stable in the upcoming period. While there are positive indicators, significant risks and cost pressures suggest that substantial growth is unlikely.
2. Key Drivers: The company's revenue growth trajectory indicates a strong demand for its semiconductor products, driven by technological advancements and increased market penetration.

This growth is a positive sign for EPS stability. Additionally, strong cash flow from operating activities provides the necessary liquidity for operational needs and potential reinvestments. However, the improvement in gross profit margins is countered by rising operating costs, primarily due to increased input prices and higher R&D expenditures. These dynamics create a scenario where revenue growth may not translate into significant EPS increases.

3. Risks and Weaknesses: Several risks could impede earnings performance. The escalating operating costs pose a direct threat to profitability, potentially leading to profit erosion in future periods. Furthermore, the increasing leverage trend raises concerns about financial stability; higher interest expenses could further constrain earnings. Additionally, customer concentration presents a risk if key clients face challenges or alter purchasing strategies, which could introduce volatility to EPS. Regulatory scrutiny in the evolving semiconductor industry also poses an uncertainty that could impact operational efficiency and profitability.

4. Confidence Level: My confidence level in this forecast is Moderate. While the company has demonstrated robust revenue growth and strong cash flow, the risks associated with rising costs and increasing debt create a precarious balance. The mixed performance in net profit further complicates the outlook, suggesting that while stability is achievable, significant growth remains uncertain.

5. Conclusion: In conclusion, while the company benefits from strong market demand and operational efficiencies, the looming threats of cost pressures and leverage challenges indicate that EPS will likely remain stable rather than experience significant growth. The path forward requires diligent cost management and strategic focus on diversifying its customer base to mitigate risks effectively.

Based on our experiemence, the LLM shows strong fidelity to the prompt design (Wang et al., 2024), adhering to the prescribed structure, semantics, and reasoning flow. These findings suggest that well-crafted prompts enable LLMs to produce finance-oriented outputs that are both qualitatively rich and closely aligned with analytical objectives.

4 Embedding-based predictive modeling

In this section, we evaluate the predictive value of LLM-generated reports by embedding them into numerical representations and incorporating them into forecasting models. Subsection 4.1 introduces the semantic vectorization step, where the textual outputs of the two-stage prompts are transformed into dense embedding vectors. Subsection 4.2 outlines the experimental design, including dimensionality reduction, feature construction, and the choice of predictive models. Finally, Subsection 4.3 presents the empirical results, assessing whether embedding-based features, alone or combined with structured accounting variables, improve the accuracy and robustness of earnings prediction.

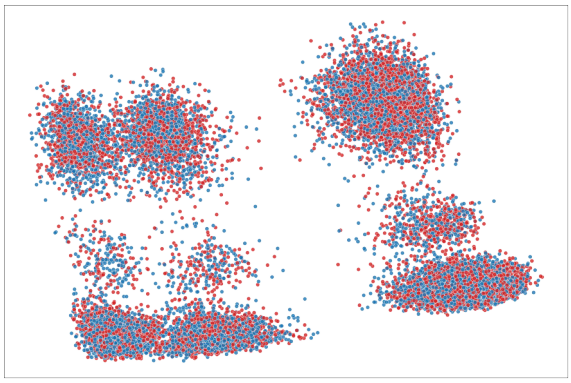
4.1 Semantic vectorization through embedding models

To evaluate the predictive value of LLM-generated reports, we transform the textual outputs into numerical representations suitable for machine learning tasks. Specifically, we employ OpenAI's `text-embedding-3-small` model, released in 2024 as part of the Ada-003 family. This model maps each report into a fixed-length vector of 1536 dimensions, capturing sentiment, causal reasoning, and financial interpretations embedded in the language model's analysis. These vectors serve as compressed latent representations of a firm's financial condition, as interpreted through the lens of the LLM.

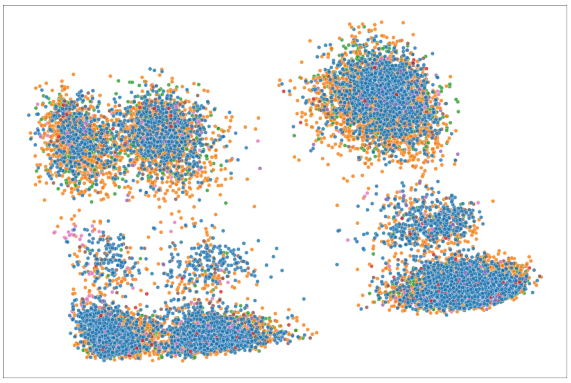
Compared with earlier embedding models, `text-embedding-3-small` exhibits higher semantic fidelity and stronger alignment with domain-specific reasoning, while remaining computationally efficient. Prior benchmark studies show its superior performance in retrieval, clustering, and semantic similarity tasks, making it well-suited for large-scale experimentation in finance. The choice of this model therefore balances domain accuracy with scalability.

In our framework, the embeddings are incorporated as input features for downstream supervising learning tasks. This enables us to test whether the qualitative reasoning captured by the LLM encodes predictive signals that complement structured numerical accounting data. To ensure compatibility and reproducibility, all embeddings are computed without truncation and normalized prior to model training.

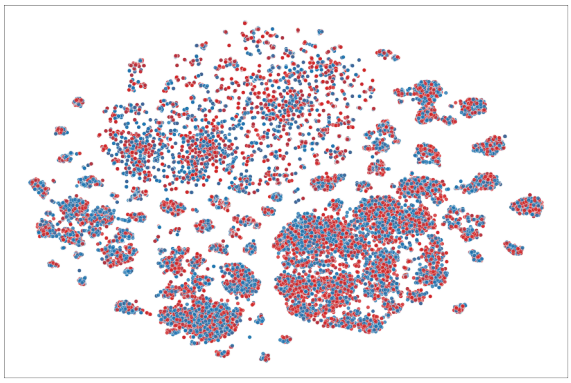
To further examine the semantic structure, we apply three techniques: Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP)—and visualize the resulting two-dimensional projections in Figure 5.1. The plots reveal partial clustering patterns, suggesting that the embeddings capture latent semantic structure. However, class labels, such as EPS direction or industry category, remain intermixed within clusters, indicating limited separability. These findings imply that while embedding-based representations encode meaningful semantics, they are insufficient on their own for clear classification. More expressive predictive models, such as XGBoost, are required to fully exploit the information embedded in the LLM-generated texts.



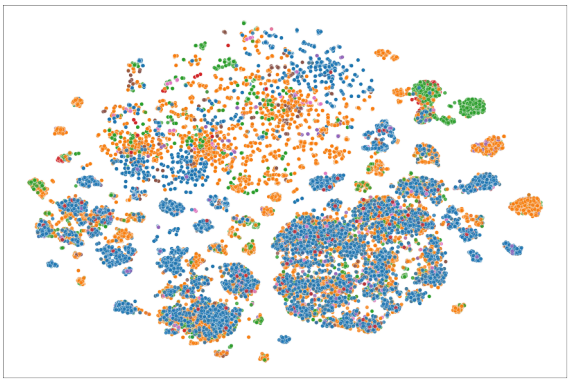
(a) PCA projection colored by EPS direction



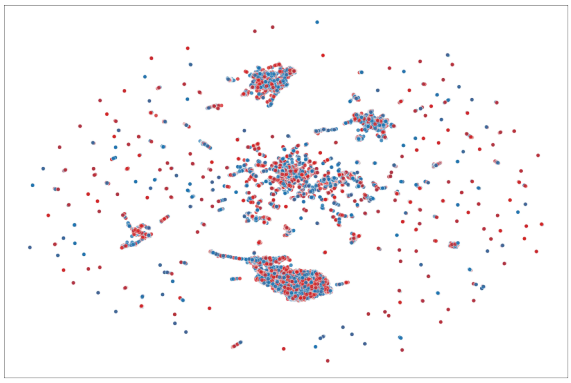
(b) PCA projection colored by industry category



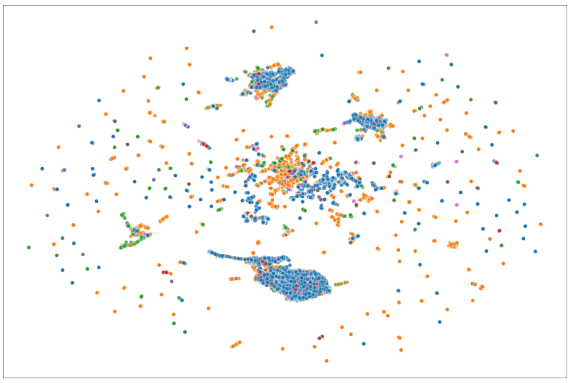
(c) t-SNE projection colored by EPS direction



(d) t-SNE projection colored by industry category



(e) UMAP projection colored by EPS direction



(f) UMAP projection colored by industry category

Figure 4.1: Two-dimensional projections of text embeddings using PCA (top row), t-SNE (middle row), and UMAP (bottom row). Left panels ((a), (c), (e)) are colored by EPS direction: Class 0 (upward) and Class 1 (downward). Right panels ((b), (d), (f)) are colored by industry category: Information Electronics, Consumer & Retail, Traditional Manufacturing, Finance & Services, Construction and Infrastructure, Agriculture & Others, and Transportation and Logistics. While visually discernible clusters emerge under each projection method, EPS directions and industry categories are largely intermixed, indicating only partial separability in the semantic embedding space.

Table 4.1. Features definition

Acronym	Descriptions
X_1	Financial data (33 accounting items in the past consecutive five periods)
X_2	Text embedding vector (1536 dimensions)
$X_1 \cup X_2$	Financial data with text embedding vector
PCA$(X_2, n\%)$	PCA-transformed X_2 preserving $n\%$ of variance explained.

4.2 Design of experiment

To evaluate whether the text-embedding vector generated from LLM-based financial analysis is an effective for prediction, we compare whether the embedding-based features, X_2, outperforms when using only traditional structured accounting variables, X_1, in terms pf prediction. To address the curse of dimensionality, we apply PCA to the raw 1536-dimensional text embedding vectors (Jolliffe and Cadima, 2016). This step reduces the dimensionality of the embeddings while retaining the majority of their variance, and serves as a diagnostic tool to test whether lower-dimensional semantic representations can retain predictive power.

We compare three types of feature representations in the empirical analysis:

(i) structured financial statement variables, denoted by X_1;

(ii) the original high-dimensional text embedding vectors generated from LLM-based reports, denoted by X_2; and

(iii) their dimension-reduced counterparts, $\mathrm{PCA}(X_2, n\%)$, obtained via principal component analysis.

Here, $\mathrm{PCA}(X_2, n\%)$ represents the PCA-transformed embedding vectors that preserve $n\%$ of the total variance explained. These feature representations are evaluated both individually and in combination to examine their predictive contribution and complementarity. Table 4.1 summarizes the definitions of all feature sets considered in the empirical analysis. The resulting hybrid representations integrate semantic information extracted from language models with structured financial variables, allowing for a systematic comparison of alternative feature engineering strategies in EPS forecasting.

[Mentioning that we are working with 10-fold CV.] The final step involves training predictive models on multiple constructed feature sets to forecast changes in corporate earnings. To enhance model robustness, we preprocess numerical financial statement features by applying z-score normalization within each variable (Wu et al., 2022). This ensures that differences in accounting item scales do not unduly influence the learning process (Eutsler and Lang, 2015). In addition, we explicitly encode missing values using binary missing indicators, which help the model recognize patterns related to the presence or absence of certain accounting entries—an important consideration in real-world ing data (Greenland and Finkle, 1995; Poulos and Valle, 2018; Che et al., 2018; Emmanuel et al., 2021). This approach allows the model to learn from the signal potentially carried by missingness itself, rather than imputing values that may introduce bias. We use Logistic Regression and XGBoost as our main supervising learning tasks, evaluate each configuration using confusion matrices, ROC curves, and decile accuracy analyses.

While binary classification offers a useful baseline for detecting directional changes in EPS, it does not reflect the magnitude of these shifts. In financial analysis, it is often more important to distinguish between moderate fluctuations and extreme earnings movements. We therefore consider a three-class setting, aiming to identify firms with significant upward or downward EPS changes, as well as those with relatively stable performance. For classification tasks where the goal is to predict the directional change of EPS. We evaluate models using accuracy, F1 score, and the area under the ROC curve (AUC).

4.3 Empirical results

For the three-class EPS classification tasks, we consider two ways to define the three classes. The first one is to use a 25% and 75% quantile split, and the second one is the to use a 10% and 90% split. We use class 0 to represent the lowest quartile (significant decrease), class 1 the middle half (stable), and class 2 the highest quartile (significant increase). Here, AUC_0 is the AUC by considering the first group of EPS as the benchmark, and merging the other two groups as one group.

Using Logistic Regression and XGBoost, we compare the three-class EPS classification in

Table 4.2. First all, it is shown that XGBoost consistently outperforms logistic regression in terms of AUC_1, AUC_0, and AUC_2, across all feature sets. Particularly for feature set $X_1 \cup X_2$, XGBoost achieves the best performance with test AUC in the testing dataset. This indicates that including X_2 may posess strong capability in identifying companies with substantial EPS growth.

Table 4.2. Three-class classification results by feature set, split, and model (quantile)

Feature	F1		AUC_1		AUC_0		AUC_2	
	LR	XGB	LR	XGB	LR	XGB	LR	XGB
Training dataset								
X_1	0.6376	0.8036	0.8087	0.9365	0.8404	0.9271	0.8427	0.9300
X_2	0.4977	0.7457	0.6820	0.8984	0.7009	0.9045	0.6995	0.9147
PCA(X_2, 90%)	0.4304	0.7142	0.6100	0.8730	0.6240	0.8801	0.6128	0.8870
PCA(X_2, 80%)	0.3939	0.6615	0.5758	0.8338	0.5836	0.8408	0.5735	0.8531
$X_1 \cup X_2$	0.6698	0.8317	0.8335	0.9497	0.8617	0.9438	0.8626	0.9465
$X_1 \cup$ PCA(X_2, 90%)	0.6576	0.8206	0.8248	0.9443	0.8544	0.9367	0.8553	0.9397
$X_1 \cup$ PCA(X_2, 80%)	0.6495	0.8173	0.8189	0.9430	0.8497	0.9352	0.8502	0.9380
Testing dataset								
X_1	0.6085	0.7600	0.7445	0.7614	0.7559	0.7645	0.7606	0.8083
X_2	0.4543	0.5685	0.5280	0.5155	0.5173	0.5170	0.5228	0.5065
PCA(X_2, 90%)	0.4309	0.5828	0.5247	0.5065	0.5214	0.4983	0.5147	0.5075
PCA(X_2, 80%)	0.4386	0.5503	0.5328	0.5129	0.5420	0.5250	0.5148	0.5008
$X_1 \cup X_2$	0.6206	0.7570	0.7424	0.7677	0.7507	0.7714	0.7584	0.8208
$X_1 \cup$ PCA(X_2, 90%)	0.6140	0.7623	0.7410	0.7698	0.7509	0.7763	0.7561	0.8190
$X_1 \cup$ PCA(X_2, 80%)	0.6101	0.7625	0.7412	0.7694	0.7527	0.7760	0.7553	0.8185

As a robust check in defining three-class EPS classification, Table 4.3 replicate the same experiment except that the definition of the three classes is now defined by the one standard deviation split. Indeed, we find similar pattern. XGBoost consistently outperforms logistic

regression across all feature sets and evaluation metrics. Particularly for feature set $X_1 \cup X_2$, which combines financial variables and text-based embeddings, XGBoost achieves the best performance with test F1 scores and AUC, and AUC_0 means the first group of EPS compares with other two groups. This indicates strong capability in identifying companies with substantial EPS growth.

Table 4.3. Three-class classification results by feature set, split, and model (z-score)

Feature	F1		AUC_1		AUC_0		AUC_2	
	LR	XGB	LR	XGB	LR	XGB	LR	XGB
Training dataset								
X_1	0.6149	0.6978	0.7744	0.8687	0.8023	0.8677	0.8089	0.8848
X_2	0.4251	0.6181	0.6076	0.8006	0.6180	0.7844	0.6143	0.8236
PCA(X_2, 90%)	0.3908	0.6048	0.5706	0.7810	0.5842	0.7785	0.5734	0.8009
PCA(X_2, 80%)	0.3712	0.5543	0.5533	0.7386	0.5665	0.7294	0.5542	0.7554
$X_1 \cup X_2$	0.6221	0.7242	0.7838	0.8866	0.8109	0.8849	0.8161	0.9029
$X_1 \cup$ PCA(X_2, 90%)	0.6135	0.7135	0.7731	0.8807	0.8034	0.8799	0.8072	0.8980
$X_1 \cup$ PCA(X_2, 80%)	0.6125	0.7091	0.7720	0.8789	0.8027	0.8783	0.8065	0.8964
Testing dataset								
X_1	0.5889	0.6327	0.7594	0.8106	0.7857	0.8165	0.7855	0.8492
X_2	0.3310	0.3560	0.5104	0.5172	0.5227	0.5360	0.5056	0.5111
PCA(X_2, 90%)	0.3287	0.3760	0.5152	0.5075	0.5240	0.5255	0.5157	0.4995
PCA(X_2, 80%)	0.3231	0.3539	0.5164	0.5053	0.5279	0.5236	0.5175	0.4985
$X_1 \cup X_2$	0.5846	0.6270	0.7542	0.8152	0.7826	0.8204	0.7829	0.8594
$X_1 \cup$ PCA(X_2, 90%)	0.5813	0.6303	0.7541	0.8172	0.7831	0.8218	0.7836	0.8617
$X_1 \cup$ PCA(X_2, 80%)	0.5815	0.6277	0.7541	0.8177	0.7831	0.8222	0.7836	0.8621

5 Conclusion

This study explores whether LLMs, when prompted effectively, can extract semantically meaningful insights from structured financial statements to assist forecasting EPS. By designing a two-stage prompt framework, we guide the LLM to first generate qualitative financial analysis reports based solely on multi-period numerical accounting items, and then form directional prediction report for EPS forecasts. These outputs are transformed into high-dimensional embeddings, which are used to predict EPS using Logistic Regression and XGBoost, alongside traditional financial variables.

Our findings show that text embeddings alone (X_2) are not sufficient for strong predictive performance, particularly in linear models like logistic regression. However, when combined with the selected accounting items (X_1), the resulting hybrid representations ($X_1 \cup X_2$) significantly enhances model performance when using XGBoost in the three-class EPS classification task.

While PCA is adopted in this study as a linear and reproducible dimensionality reduction technique for embedding-based predictive modeling, we emphasize that alternative nonlinear methods such as t-SNE and UMAP are employed exclusively for visualization purposes. These visualization tools are used to explore whether the LLM-generated embedding representations exhibit latent structure with respect to earnings outcomes or industry classifications, rather than to enhance predictive performance. A systematic comparison between PCA and nonlinear dimensionality reduction techniques in downstream prediction tasks may yield further insights into predictive accuracy. We therefore leave such extensions for future research.

Finally, while this study focuses on GPT-4o Mini as a representative general-purpose LLM, the proposed framework is intentionally designed to be model-agnostic and extensible. Benchmarking against other general-purpose GAI models (e.g., Gemini, Claude), finance-specific LLMs (e.g., BloombergGPT, FinBERT), or alternative text-embedding models would further strengthen the external validity and international relevance of our findings (Wu et al., 2023; Yang et al., 2020; Li et al., 2023b). At the same time, rapid advances in foundation models—driven by sustained improvements in computational efficiency and scaling—suggest that

the relative performance of specific LLMs may evolve quickly over time (Kaplan et al., 2020). Accordingly, we position our contribution not as a model-comparison exercise, but as a reusable and transparent analytical pipeline that can be readily applied to future LLMs and embedding models as they emerge (Achiam et al., 2023). We therefore leave systematic cross-model benchmarking as an important direction for future research.

References

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. (2023). GPT-4 technical report. `https://arxiv.org/abs/2303.08774`. Accessed: 2025-06-23.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, 1877–1901.

Cao, K. and H. You (2024). Fundamental analysis via machine learning. *Financial Analysts Journal 80*(2), 74–98.

Chang, C., W.-Y. Wang, W.-C. Peng, and T.-F. Chen (2025). LLM4TS: Aligning pre-trained LLMs as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology 16*(3), 1–20.

Che, Z., S. Purushotham, K. Cho, D. Sontag, and Y. Liu (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports 8*(1), 6085.

Chen, Z., S. Li, C. Smiley, Z. Ma, S. Shah, and W. Y. Wang (2022). Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. `https://arxiv.org/abs/2210.03849`. Accessed: 2022-10-07.

Cheong, F. S. and J. Thomas (2018). Management of reported and forecast EPS, investor responses, and research implications. *Management Science 64*(9), 4277–4301.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research 24*(240), 1–113.

Deloof, M. (2003). Does working capital management affect profitability of Belgian firms? *Journal of Business Finance & Accounting 30*(3-4), 573–588.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Emett, S. A. and M. W. Nelson (2017). Reporting accounting changes and their multi-period effects. *Accounting, Organizations and Society 57*, 52–72.

Emmanuel, T., T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona (2021). A survey on missing data in machine learning. *Journal of Big Data 8*, 1–37.

Eutsler, J. and B. Lang (2015). Rating scales in accounting research: The impact of scale points and labels. *Behavioral Research in Accounting 27*(2), 35–51.

Francis, J., R. LaFond, P. Olsson, and K. Schipper (2005). The market pricing of accruals quality. *Journal of Accounting and Economics 39*(2), 295–327.

Gao, H., G. Kou, H. Liang, H. Zhang, X. Chao, C.-C. Li, and Y. Dong (2024). Machine learning in business and finance: A literature review and research opportunities. *Financial Innovation 10*(1), 86.

Greenland, S. and W. D. Finkle (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology 142*(12), 1255–1264.

Gupta, U. (2023). GPT-InvestAR: Enhancing stock investment strategies through annual report analysis with large language models. `https://arxiv.org/abs/2309.03079`. Accessed: 2023-09-06.

Huang, A. H., H. Wang, and Y. Yang (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research 40*(2), 806–841.

Hunt, J. O., J. N. Myers, and L. A. Myers (2022). Improving earnings predictions and abnormal returns with machine learning. *Accounting Horizons 36*(1), 131–149.

Jolliffe, I. T. and J. Cadima (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of The Royal Society A: Mathematical, Physical and Engineering Sciences 374*(2065), 20150202.

Joshi, S. (2025). Review of gen AI models for financial risk management. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology 11*(1), 709–723.

Kaplan, J., S. McCandlish, T. Henighan, T. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Khan, W., M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing 13*(7), 3433–3456.

Kim, A., M. Muhn, and V. Nikolaev (2024). Financial statement analysis with large language models. `https://arxiv.org/abs/2407.17866`. Accessed: 2024-07-25.

Kleinman, G. (2001). Financial statement analysis and security valuation. *Issues in Accounting Education 16*(3), 513.

Lakkaraju, K., S. K. R. Vuruma, V. Pallagani, B. Muppasani, and B. Srivastava (2023). Can LLMs be good financial advisors?: An initial study in personal decision making for optimized outcomes. `https://arxiv.org/abs/2307.07422`. Accessed: 2023-07-08.

Le Scao, T., A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. (2023). Bloom: A 176b-parameter open-access multilingual language model. https://inria.hal.science/hal-03850124/. Accessed: 2023-11-20.

Lee, H., Y. Choi, and Y. Kwon (2024). Quantifying qualitative insights: Leveraging LLMs to market predict. https://arxiv.org/abs/2411.08404. Accessed: 2024-11-13.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems 33*, 9459–9474.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics 45*(2-3), 221–247.

Li, Y., S. Wang, H. Ding, and H. Chen (2023a). Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382.

Li, Y., S. Wang, H. Ding, and H. Chen (2023b). Large language models in finance: A survey. *ACM International Conference on AI in Finance*.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys 55*(9), 1–35.

Liu, Y., Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li (2023). Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. https://arxiv.org/abs/2308.05374. Accessed: 2023-08-10.

Mishkin, F. (2007). *Money, Banking and Financial Markets*. Paris, France: New Horizons.

Musil, T. (2019). Examining structure of word embeddings with pca. In *Text, Speech, and Dialogue*, pp. 211–223. Springer International Publishing.

Nie, Y., Y. Kong, X. Dong, J. M. Mulvey, H. V. Poor, Q. Wen, and S. Zohren (2024). A survey of large language models for financial applications: Progress, prospects and challenges. https://arxiv.org/abs/2406.11903. Accessed: 2024-06-15.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35*, 27730–27744.

Pabuccu, H. and A. Barbu (2024). Feature selection with annealing for forecasting financial time series. *Financial Innovation 10*(1), 87.

Pele, D. T., V. Bolovaneanu, M.-B. Lin, R. Ren, A. T. Ginavar, B. Spilak, A. V. Andrei, M. Toma, S. Lessmann, and W. K. Härdle (2024). In the beginning was the word: LLM-VaR and LLM-ES. Avaiable at SSRN: https://ssrn.com/abstract=5104383. Accessed: 2025-03-11.

Poulos, J. and R. Valle (2018). Missing data imputation for supervised learning. *Applied Artificial Intelligence 32*(2), 186–196.

Richards, V. D. and E. J. Laughlin (1980). A cash conversion cycle approach to liquidity analysis. *Financial Management 9*(1), 32–38.

Sloan, R. G. and A. Y. Wang (2024). Predictable EPS growth and the performance of value investing. *Review of Accounting Studies 29*(1), 1–46.

Song, C. H., J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su (2023). LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2998–3009.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. (2023). LLaMA: Open and efficient foundation language models. `https://arxiv.org/abs/2302.13971`. Accessed: 2025-06-23.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Wang, L., X. Chen, X. Deng, H. Wen, M. You, W. Liu, Q. Li, and J. Li (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digital Medicine 7*(1), 41.

Wang, S., X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang (2023). GPT-NER: Named entity recognition via large language models. `https://arxiv.org/abs/2304.10428`. Accessed: 2023-04-20.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35*, 24824–24837.

White, G. I., A. C. Sondhi, and D. Fried (2002). *The Analysis and Use of Financial Statements* (3rd ed.). New York, NY: John Wiley & Sons.

Wu, D., X. Ma, and D. L. Olson (2022). Financial distress prediction using integrated z-score and multilayer perceptron neural networks. *Decision Support Systems 159*, 113814.

Wu, S., H. Liu, R. Kapoor, R. Jain, J. Liao, T. Bansal, et al. (2023). BloombergGPT: A large language model for finance. `https://arxiv.org/abs/2303.17564`. Accessed: 2023-03-30.

Yang, X., S. Zang, Y. Ren, D. Peng, and Z. Wen (2024). Evaluating large language models on financial report summarization: An empirical study. `https://arxiv.org/abs/2411.06852`. Accessed: 2024-11-11.

Yang, Y., M. C. S. Uy, and A. Huang (2020). FinBERT: A pretrained language model for financial communications. `https://arxiv.org/abs/2006.08097`. 2020-06-15.

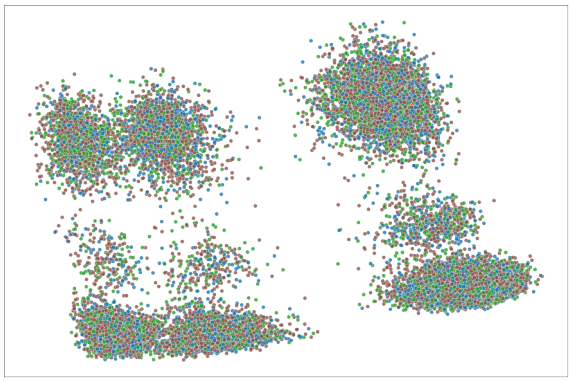
Yu, X., Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu (2023). Temporal data meets LLM–explainable financial time series forecasting. `https://arxiv.org/abs/2306.11025`. Accessed: 2023-06-19.

Zebari, R., A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends 1*(1), 56–70.

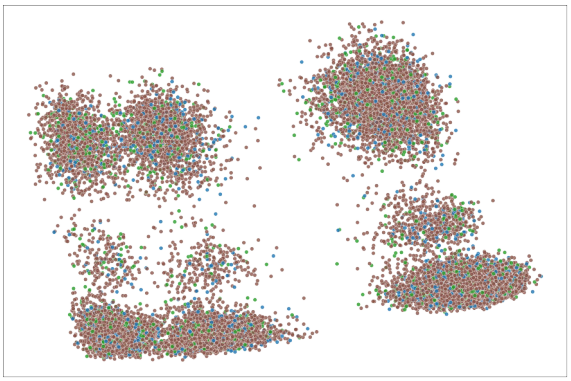
Zhao, H., Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, et al. (2024). Revolutionizing finance with LLMs: An overview of applications and insights. `https://arxiv.org/abs/2401.11641`. Accessed: 2024-01-22.

Zhu, K., J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, Y. Zhang, N. Gong, et al. (2023). Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pp. 57–68.

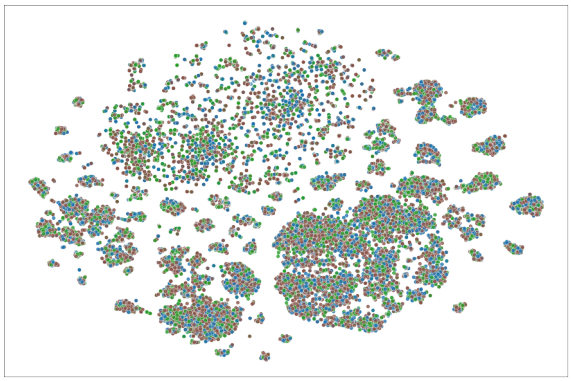
Appendix: Supplementary materials



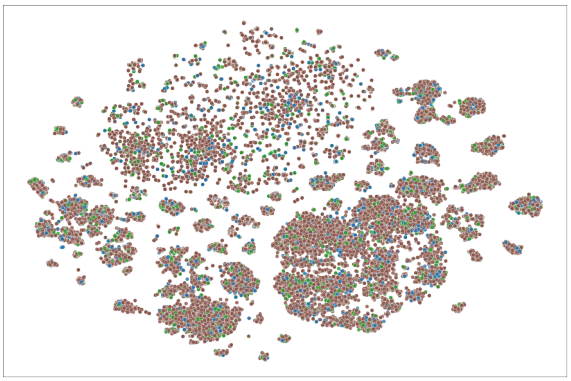
(a) PCA projection colored by EPS quantile



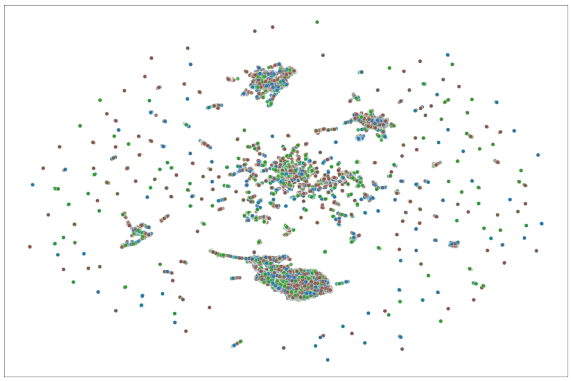
(b) PCA projection colored by EPS zscore



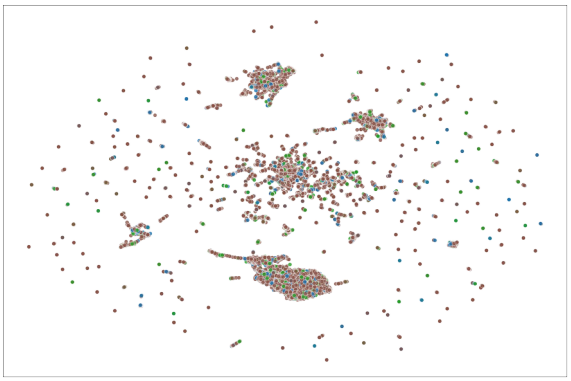
(c) t-SNE projection colored by EPS quantile



(d) t-SNE projection colored by EPS zscore



(e) UMAP projection colored by EPS quantile



(f) UMAP projection colored by EPS zscore

Figure 5.1: using PCA, t-SNE, and UMAP, colored by three-class EPS classification, where blue indicates the lowest (decrease) class , brown represents the middle class , and green denotes the highest (increase) class.