View From Board: Financial Statement Analysis with Large Language Models

January 6, 2026

Abstract

The integration of Large Language Models (LLMs) into financial statement analysis (FSA) presents a transformative opportunity to enhance predictive accuracy and decision-making. This study evaluates the effectiveness of LLMs, particularly GPT-4o, in predicting future earnings based on structured financial data and narrative disclosures. We explore the impact of Chain-of-Thought (CoT) prompting and Management Discussion and Analysis (MD&A) integration on LLM performance. Using a dataset from Compustat spanning 1968 to 2021, we compare LLM-based predictions against traditional benchmarks, including logistic regression, feedforward neural networks (FNNs), and analyst forecasts. Our findings reveal that CoT-enhanced LLMs outperform naive models and early analyst predictions, with further gains achieved by incorporating MD&A insights. However, despite these improvements, LLMs still underperform specialized deep learning models in numerical prediction tasks. Moreover, LLMs exhibit challenges in anticipating macroeconomic shocks, limiting their robustness during economic downturns. Our study highlights the potential of LLMs in FSA while emphasizing the need for hybrid approaches that integrate structured and unstructured data for superior predictive performance.

Keywords: Large Language Models, ChatGPT, Man- agement Discussion and Analysis, Financial Statement Analysis, Chain-of-Thought

1 Introduction

The increasing complexity and globalization of financial markets have elevated the importance of tools capable of analyzing vast quantities of financial data efficiently and accurately. Among such tools, Large Language Models (LLMs), such as ChatGPT-4o and its derivatives, have emerged as transformative technologies in the domain of financial statement analysis (FSA). LLMs offer the ability to analyze, interpret, and reason over both tex-

tual and numeric data, which has traditionally been the domain of financial analysts and narrowly focused machine learning (ML) models. Despite this promise, the integration of LLMs into FSA remains an area of active exploration, particularly in understanding their comparative advantage, limitations, and practical applications.

The development of language models from early models like BERT to the latest iterations such as GPT-4o and Deepseek-v3 has significantly influenced their applications in finance. BERT based models, as shown by Devlin (2018), provided robust contextual embeddings, making them suitable for tasks like sentiment analysis and news classification. FinBERT, proposed by Huang et al. (2023), a domain-specific adaptation, enhanced performance in financial sentiment tasks. However, these earlier models were limited in understanding long-term dependencies and insensitivity to numerical data. The advent of transformer-based models like GPT-3 (Brown et al. 2020) introduced autoregressive capabilities, enabling more complex reasoning over financial disclosures. GPT-4 and similar advanced models improved on this by integrating multimodal capabilities, as demonstrated by Kim et al. (2024a).

LLMs have initially demonstrated its potential in Economics and Finance field. Cao et al. (2024) examined the use of LLMs in crisis prediction, finding that models like GPT-4 could identify early warning signals from textual data in financial reports and news articles. Kong et al. (2024) applied LLMs to forecast stock price movements based on ESG (Environmental, Social, Governance) disclosures, showcasing their ability to synthesize diverse narrative themes. Additional research, such as the work of Bybee (2023), explored how LLMs process macroeconomic predictions from news headlines, demonstrating their ability to align with expert survey results. Similarly, Lopez-Lira and Tang (2023) highlighted GPT's effectiveness in explaining short-term stock returns, while Hansen and Kazinnik (2023) demonstrated LLMs' capabilities in interpreting central bank announcements and predicting subsequent macroeconomic shocks.

However, in fields like financial statement analysis, significant questions remain unanswered. Can LLM be used to do the financial statement analysis like human? Can it provide a unique perspective that is difficult for human analysts to discover?

Financial statement analysis (FSA) is a cornerstone of decision-making in financial markets. It provides investors, analysts, and other stakeholders with critical insights into a firm's financial health, operational efficiency, and future prospects. Traditional FSA involves detailed scrutiny of balance sheets, income statements, and cash flow statements, complemented by contextual interpretation of narrative disclosures such as MD&A sections. These analyses guide key decisions, including investment strategies, credit evaluations, and regulatory compliance. However, there remains a notable gap in the literature regarding the inclusion of the Management Discussion and Analysis (MD&A) section in FSA. The MD&A section is a critical component of financial disclosures, providing management's perspective on the company's financial results, operations, risks, and forward-looking strategies. Unlike purely numeric disclosures, the MD&A offers a nuanced, qualitative layer of information that contextualizes financial performance and helps stakeholders assess the implications of reported numbers. Prior research, such as Hale and Wetmiller (2024) and Bybee (2023), has demonstrated the value of narrative analysis in enhancing predictive accuracy. However, human analysts often struggle with information overload, biases, and inconsistencies in interpreting qualitative narratives. Moreover, narrowly focused machine learning and deep learning models, while effective in analysing numerical data, lack the flexibility and generalizability required for comprehensive FSA. Against this backdrop, LLMs present an unprecedented opportunity to revolutionize FSA by combining the interpretive skills of human analysts with the computational power of ML models. However, the potential of LLMs to extract insights specifically from MD&A sections has not been fully explored.

Building on these insights, our research seeks to answer the following questions:

(1) How about LLM's prediction performance when compared with Human Analysts,

traditional models and machine learning models?

(2) Can LLMs' performance be further enhanced by incorporating contextual information from narratives such as the Management Discussion and Analysis (MD&A) section?

(3) Is there any way to improve LLM performance in FSA

Our paper builds on the insights of existing literature, including the work of Kim et al. (2024a,b), who investigated the efficacy of LLM in analyzing standardized financial statements and narrative sections of corporate disclosures. These studies demonstrated that LLMs can rival or even surpass human analysts and specialized ML models in predicting directional changes in earnings and stock returns. The foundational research by Kim et al. (2024a) demonstrated that GPT-4, when provided with standardized financial statements, can achieve prediction accuracies comparable to specialized ML models and superior to human analysts in certain scenarios. Notably, the study highlighted the effectiveness of CoT prompting in guiding LLMs to emulate human-like reasoning in analyzing numeric data. In parallel, Kim et al. (2024b) emphasized the role of attention mechanisms in processing the narrative content of annual reports, showing how specific topics, such as liquidity and capital resources, significantly influence investor attention. However, the study also acknowledged limitations, including the exclusion of narrative data and the potential for LLMs to inadvertently leverage external knowledge, thereby introducing biases.

Based on their work, we propose to integrate qualitative insights from MD&A into the Chain-of-Thought (CoT) prompting framework to provide a richer context for earnings predictions. We design a CoT framework that combines numerical and narrative reasoning, enabling the LLM to analyze financial trends alongside the contextual explanations of the management. Our study adopts a two-pronged approach: (1) preprocessing and anonymizing financial statements (balance sheet and income sheet) and narrative data and (2) designing and testing a novel CoT prompting framework. We preprocess financial statements by removing all firm-specific identifiers, replacing names and dates with placeholders

(e.g., "Company A," "Year t"). The financial data is then standardized using a consistent template to eliminate format variations. MD&A sections are similarly anonymized and parsed into thematic segments (e.g., performance highlights, risk factors) to facilitate targeted analysis. We extend traditional CoT prompting to incorporate narrative reasoning. This paper aims to make a key contribution. We aim to introduce a novel CoT prompting technique that integrates numeric and narrative reasoning for financial analysis. Currently, applications focus on numerical data but omit MD&A, which is indispensable for financial statements analysis. Our aim is to provide a robust framework for anonymizing and standardizing financial data, ensuring unbiased LLM performance, and offer empirical evidence on the efficacy of narrative-enhanced LLM analysis in predicting EPS, outperforming both traditional methods and standalone numeric analysis. We believe that our work will advance the understanding of how LLMs can emulate human-like reasoning in complex decision-making tasks, contributing to the broader literature on artificial intelligence in finance.

In the subsequent sections, we detail our methodology, experimental design, and findings, illustrating how LLMs, when guided by advanced CoT techniques, can serve as a powerful tool for financial statement analysis.

2 Methodology and Data

2.1 LLM and Chatgpt

A language model is a statistical framework trained on extensive text corpora to predict the probability distribution of word sequences. Consider a word sequence $W = w_1, w_2, \ldots, w_n$, where w_i represents the i-th word. The goal is to calculate $P(W)$, expressed as:

$$P(W) = P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2) \ldots P(w_n \mid w_1, w_2, \ldots, w_{n-1})$$

Here, $P(w_i \mid w_1, w_2, \ldots, w_{i-1})$ captures the likelihood of w_i given its preceding words. Over the past few decades, language model architectures have undergone significant evolution from the basic n-gram models, which represented word sequences as Markov processes to neural network-based models, which are capable of capturing long-term dependencies in sequential data. However, in 2017, the introduction of the transformer architecture, proposed by Vaswani (2017), revolutionized language modeling, surpassing the performance of neural networks in tasks such as machine translation. Transformers employ self-attention mechanisms to model parallel relationships between words, facilitating efficient training on large-scale datasets. Prominent transformer-based models include GPT (Generative Pre-trained Transformer) (Wang et al. 2020), which is decoder-only framework, BERT (Bidirectional Encoder Representations from Transformers) Devlin (2018), which is encoder-only framework, and T5 (Text-to-Text Transfer Transformer) (Raffel et al. 2020), which leverages both encoder and decoder structures. These models have achieved state-of-the-art results on various natural language processing (NLP) tasks through transfer learning.

Among all large language models, ChatGPT is undoubtedly the most representative of them all. ChatGPT is an artificial intelligence (AI) chatbot developed by OpenAI, designed to generate human-like conversational responses using natural language processing. It can answer questions and compose various forms of written content, such as articles, social media posts, essays, code, and emails. As a type of generative AI, ChatGPT allows users to input prompts and receive AI-generated text, images, or videos that resemble human creations. It functions similarly to automated chat services found on customer service websites, enabling users to ask questions or seek clarification on its responses. The acronym "GPT" stands for "Generative Pre-trained Transformer," indicating the model's method of processing requests and generating replies. ChatGPT is trained using reinforcement learning from human feedback and reward models that rank optimal responses, enhancing its performance through machine learning techniques. ChatGPT operates through its

Generative Pre-trained Transformer, utilizing specialized algorithms to identify patterns within data sequences. Initially, it employed the GPT-3 large language model, a neural network machine learning model and the third generation of Generative Pre-trained Transformer. This transformer model draws from extensive data to generate responses. Building on the capabilities of GPT-3, OpenAI introduced GPT-4, which marked a significant advancement in generative AI technology. GPT-4 improved upon its predecessor by incorporating a larger and more refined dataset, resulting in more accurate, coherent, and context-aware responses. It also introduced enhanced reasoning abilities, allowing for better problem-solving and the ability to handle more complex queries.

One of the key upgrades in GPT-4 was its multimodal capabilities, enabling it to process not just text but also images, making it versatile across a wider range of applications. Furthermore, the model was optimized for better scalability and efficiency, making it more accessible to developers and organizations. These enhancements allowed GPT-4 to power applications requiring deeper contextual understanding, such as advanced customer support, educational tools, and creative content generation. Our paper is using the latest version chatgpt-4o-turbo to train our models.

2.2 Chain-of-Thought

Modern large language models (LLMs), such as GPT-4 and its successors, have demonstrated impressive capabilities in retrieving data from structured tables and performing simple calculations. However, they often lack the nuanced judgment required for tasks that involve human-like reasoning and complex decision-making, such as financial statement analysis. This limitation has driven the development of Chain-of-Thought (CoT) prompting, a method designed to enhance the reasoning and problem solving abilities of LLMs (Wei et al., 2022). The CoT approach enables LLMs to emulate the structured thought processes of human analysts by breaking down complex tasks into sequential, in-

terpretable steps. When applied to financial analysis, CoT prompting allows the model to act as a financial analyst tasked with evaluating a firm's financial health and predicting future earnings performance.

The implementation of CoT prompting begins by instructing the model to identify key financial metrics and their changes over time. This step involves prompting the model to detect notable variations in items such as revenue, net income, operating expenses, and liquidity metrics. The model is not restricted to predefined thresholds but is encouraged to focus on material changes that are likely to influence the firm's financial standing. For instance, if revenue has increased by 15% while operating costs have risen by 20%, the model is expected to highlight these trends and consider their implications on profitability. Unlike traditional approaches that require explicit coding for each calculation, the CoT prompt encourages the model to reason through its task by first identifying relevant metrics and then calculating financial ratios where applicable.

Subsequently, the model is guided to compute key financial ratios such as the current ratio, return on assets, and debt-to-equity ratio. These ratios provide insights into the firm's operational efficiency, liquidity, and financial leverage. To enhance the interpretability of the analysis, the CoT prompt requires the model to articulate the formula for each ratio before performing the computation. For example, the model might state: "The current ratio is calculated as current assets divided by current liabilities. Using the provided data, the current ratio is 1.5." This step ensures transparency and mirrors how financial analysts document their calculations for verification and reporting purposes.

After computing the ratios, the model is instructed to interpret the results within the broader context of the firm's financial performance. This involves linking the computed values to economic implications, such as a declining current ratio signaling potential liquidity challenges or an increasing debt-to-equity ratio indicating higher financial risk. The CoT prompt explicitly directs the model to consider external factors, such as macroeco-

nomic trends or industry-specific conditions, that might influence these interpretations. For instance, in the case of a retail firm, the model might analyze whether seasonal demand fluctuations explain variations in revenue and inventory turnover.

The next phase of the CoT prompting framework focuses on predicting future earnings trends. Based on the insights generated from the previous steps, the model is tasked with determining whether earnings are likely to increase, decrease, or remain stable in the next reporting period. This prediction is not limited to numeric changes but also includes a qualitative rationale that explains the underlying drivers. For example, the model might conclude: "Earnings are expected to decline due to a 10% drop in revenue and an 8% increase in raw material costs, driven by ongoing supply chain disruptions." This step enhances the utility of the prediction by providing actionable insights that align with the reasoning process of professional financial analysts.

In addition to the directional prediction of earnings, the CoT framework incorporates the magnitude of change. The model categorizes the predicted change into one of three levels: large, moderate, or small. For example, a "moderate" decline might be defined as a reduction in earnings between 5% and 10%, with the model explicitly justifying this categorization based on the observed trends. To further enhance reliability, the CoT prompt includes a confidence score that quantifies the model's certainty in its predictions. These scores range from 0 (random guess) to 1 (high confidence), providing users with an additional layer of information for decision-making.

The CoT framework also incorporates a rationale statement that summarizes the model's findings and explains how the computed values and interpretations support its predictions. This summary is designed to replicate the narrative reasoning process that human analysts use to communicate their conclusions to stakeholders. For example, the model might generate the following explanation: "The firm's earnings are projected to decline moderately due to rising costs and declining market share in its key segments. This conclusion is supported

by a 20% increase in operating expenses and a 15% drop in revenue in the last quarter." Here are examples to show prompts with and without CoT.

Example without CoT: "You are a financial analyst. Analyze the financial statements of Company XYZ and predict its earnings for next year. Use the income statement and balance sheet provided."

Example with CoT: "You are a financial analyst. Analyze the financial statements of Company XYZ step-by-step to predict its earnings for next year. Start by: (1) Identifying key metrics (e.g., revenue growth, net profit margin, debt-to-equity ratio) from the income statement and balance sheet. (2) Analyzing trends and comparing them to industry benchmarks. (3) Considering risks and opportunities based on the financial data. (4) Summarizing your reasoning and providing the earnings prediction."

2.3 Data

We refer Kim et al. (2024a) to use the full Compustat annual financial dataset. Our dataset is from fiscal years 1968 to 2021, reserving 2022 data to predict 2023 earnings. This allows us to evaluate the model's performance beyond GPT's training window, which ends in April 2023. The step is to ensure no exposure to 2023 earnings data released in March 2024. Following prior literature, we include only observations with non-missing total assets, year-end assets above $1 million, stock prices exceeding $1 per share, and a December 31 fiscal year-end.

We exclude cash flow statements because of its complexity of Standardization and limited predictive value. Cash flow statements include non-cash items like depreciation, amortization, and working capital adjustments, which vary widely across industries and companies. Standardizing these elements can be challenging and may introduce noise into the analysis. Their predictive value for next-year earnings might not be as direct as that of balance sheets or income statements. Besides, using cash flow statements increase risks of

identifiability and search bias. Cash flow statements often include specific line items unique to a company's operations (e.g., "cash flows from a specific investment"), which might inadvertently reveal the company's identity. This undermines anonymization and could lead to direct identification. It might allow the model to recognize and match patterns directly to specific companies rather than making generalized predictions. This could make the model rely on memorization (searching) rather than reasoning or analyzing relationships within the data.

We reconstruct balance sheets and income statements for each firm-year. All identifying information, such as firm names and financial statement dates, is omitted to ensure uniformity. Each observation includes two years of balance sheet data and three years of income statement data, consistent with US GAAP requirements. Examples of balance sheet and income statement are provided in Table 1 and **??**.

Table 1: Balance Sheet Example

Account Items	t	t-1
Cash and Short-Term Investments	33.67	33.93
Receivables	5.15	4.99
Inventories	22.63	20.67
Other current assets	4.25	3.00
Current Assets	65.70	62.59
Property, Plant, and Equipment (Net)	100.27	59.83
Other investments	113.63	98.30
Intangible assets	4.22	3.76
Other assets	2.73	2.16
Total Asset	286.55	226.64
Debt in current liabilities	17.31	12.10
Account payable	23.90	18.60
Income taxes payable	0.000	0.000
Other current liabilities	35.15	27.57
Current liabilities	76.36	58.27
Long-term debt	94.57	59.70
Deferred taxes and investment tax credit	11.78	8.52
Other liabilities	47.51	41.01
Total Liabilities	230.22	167.50
Stockholders' equity total	42.37	47.52
Noncontrolling interest	9.97	7.85

Table 2: Income Statement Example

Account Items	t	t-1	t-2
Sales (net)	11.76	7.00	4.05
Cost of Goods Sold	9.54	5.40	3.12
Gross Profit	2.22	1.60	0.92
Selling, General and Administrative Expenses	2.48	1.43	0.93
Research and Development Expenses	1.38	0.83	0.71
Operating Expenses	3.86	2.27	1.64
Operating Income	-1.63	-0.67	-0.72
Nonoperating income (excluding interest income)	-0.58	-0.08	-0.16
Pretax income	-2.21	-0.75	-0.88
Income taxes (current)	0.032	0.026	0.013
Income from continuous operations	0.032	0.026	0.013
Net Income	-1.96	-0.68	-0.89
EBITDA	0.095	0.38	-0.22
Basic Earnings per Share	-0.79	-0.31	-0.46

2.4 Model Prompts

After processing, we come up with three different prompts to train models. Prompts without CoT, Prompts with CoT (No MD&A), and Prompts with CoT (MD&A)

Prompts without CoT: "You are a financial analyst. Perform a detailed financial statement analysis for the company based on uploaded balance sheets, income statements and MD&A, tailoring your approach to its industry. "

Prompts with CoT (No MD&A): "You are a financial analyst. Perform a detailed financial statement analysis for the company, tailoring your approach to its industry. Follow

these structured steps to ensure a comprehensive and actionable analysis:

1. Trend Analysis: Perform a comprehensive review of trends across key financial metrics, considering industry-specific factors with following steps:

Step 1: Revenue Trend:

Extract revenue values over the given period (yearly) and calculate the percentage change year-over-year (YoY). Identify whether revenue is increasing, decreasing, or stable.

Step 2: Cost of Goods Sold (COGS) Trend:

Extract COGS figures for the same periods and calculate percentage changes.

Compare COGS trends with revenue trends. Is COGS growing faster, slower, or in line with revenue? Discuss implications for gross margins and cost efficiency.

Step 3: Gross Profit Trend:

Compute gross profit for each period (Gross Profit = Revenue - COGS). Calculate the gross margin percentage (Gross Margin = Gross Profit / Revenue). Evaluate whether gross profit is improving or declining over time. Discuss any divergence between gross profit and revenue trends, highlighting its impact on margins.

Step 4: Net Income Trend:

Analyze net income figures over the same periods. Identify if net income aligns with trends in revenue and gross profit. Examine external influences such as tax changes, interest rates, or one-time charges.

Step 5: Summarize Observations:

Highlight the most critical trends in revenue, costs, and profitability. Identify anomalies or areas needing attention.

2. Ratio Analysis: Perform an in-depth ratio analysis tailored to the company's industry.

Profitability Ratios:

Operating Margin: Operating Income / Sales. Net Profit Margin: Net Income / Sales.

Return on Assets (ROA): Net Income / Total Assets. Return on Equity (ROE): Net Income / Shareholders' Equity.

Liquidity Ratios: Current Ratio: Current Assets / Current Liabilities. Quick Ratio: (Current Assets - Inventory) / Current Liabilities.

Efficiency Ratios:

Asset Turnover Ratio: Sales / Total Assets. Inventory Turnover Ratio: COGS / Average Inventory. Receivables Turnover Ratio: Net Credit Sales / Average Accounts Receivable.

Leverage Ratios:

Debt-to-Equity Ratio: Total Liabilities / Shareholders' Equity. Interest Coverage Ratio: EBIT / Interest Expense.

Step-by-Step Instructions:

Write the formula for each ratio. Perform step-by-step calculations with the given data. Provide the results and an economic interpretation of each ratio. Compare ratios to prior periods and discuss significant changes.

3. Rationale for Prediction: Use findings from Trend Analysis, Ratio Analysis, and insights from the MD&A to predict future performance.

Step 1: Integrate Quantitative and Qualitative Findings: Combine trends, ratios, and MD&A insights to form a cohesive prediction.

Step 2: Predict Future EPS: State whether EPS is likely to increase, decrease, or remain stable. Justify your prediction with specific data points, ratios, and management commentary.

Step 3: Highlight Risks and Opportunities: Discuss potential risks identified in the MD&A or financial statements (e.g., rising costs, declining asset turnover). Highlight growth opportunities or operational efficiencies that could drive future performance.

5. Confidence Score: Assess the reliability of your prediction by assigning a confidence

score between 0 and 1 (0 = No confidence, 1 = Absolute confidence).

Step 1: Base your score on: The consistency and alignment of trends, ratios, and MD&A insights. Transparency and reliability of the information provided in the MD&A. External factors such as market conditions or industry risks.

Step 2: Justify the Confidence Score: Provide a brief explanation of what supports or reduces your confidence.

6. Final Summary:

Step 1: Consolidate the Findings: Summarize the key trends, ratio results, MD&A insights, and their implications. Recap your EPS prediction and confidence score.

Step 2: Provide Actionable Insights: Highlight the company's strengths, weaknesses, opportunities, and risks. Offer recommendations for improving performance or mitigating risks. Deliver the analysis in a structured and clear format, presenting it as if you were delivering a professional report to stakeholders.

Prompts with CoT (MD&A): "You are a financial analyst. Perform a detailed financial statement analysis for the company, tailoring your approach to its industry. Follow these structured steps to ensure a comprehensive and actionable analysis:

1. Trend Analysis: Perform a comprehensive review of trends across key financial metrics, considering industry-specific factors with following steps:

Step 1: Revenue Trend:

Extract revenue values over the given period (yearly) and calculate the percentage change year-over-year (YoY). Identify whether revenue is increasing, decreasing, or stable.

Step 2: Cost of Goods Sold (COGS) Trend:

Extract COGS figures for the same periods and calculate percentage changes.

Compare COGS trends with revenue trends. Is COGS growing faster, slower, or in line with revenue? Discuss implications for gross margins and cost efficiency.

Step 3: Gross Profit Trend:

Compute gross profit for each period (Gross Profit = Revenue - COGS). Calculate the gross margin percentage (Gross Margin = Gross Profit / Revenue). Evaluate whether gross profit is improving or declining over time. Discuss any divergence between gross profit and revenue trends, highlighting its impact on margins.

Step 4: Net Income Trend:

Analyze net income figures over the same periods. Identify if net income aligns with trends in revenue and gross profit. Examine external influences such as tax changes, interest rates, or one-time charges.

Step 5: Summarize Observations:

Highlight the most critical trends in revenue, costs, and profitability. Identify anomalies or areas needing attention.

2. Ratio Analysis: Perform an in-depth ratio analysis tailored to the company's industry.

Profitability Ratios:

Operating Margin: Operating Income / Sales. Net Profit Margin: Net Income / Sales. Return on Assets (ROA): Net Income / Total Assets. Return on Equity (ROE): Net Income / Shareholders' Equity.

Liquidity Ratios: Current Ratio: Current Assets / Current Liabilities. Quick Ratio: (Current Assets - Inventory) / Current Liabilities.

Efficiency Ratios:

Asset Turnover Ratio: Sales / Total Assets. Inventory Turnover Ratio: COGS / Average Inventory. Receivables Turnover Ratio: Net Credit Sales / Average Accounts Receivable.

Leverage Ratios:

Debt-to-Equity Ratio: Total Liabilities / Shareholders' Equity. Interest Coverage Ratio: EBIT / Interest Expense.

Step-by-Step Instructions:

Write the formula for each ratio. Perform step-by-step calculations with the given data. Provide the results and an economic interpretation of each ratio. Compare ratios to prior periods and discuss significant changes.

3. Management Discussion and Analysis (MD&A): Discuss the qualitative insights provided in the MD&A section and evaluate their implications:

Step 1: Summarize Key Insights: Identify the major themes in the MD&A, such as management's perspective on financial performance, future strategies, and risk factors.Highlight forward-looking statements about growth plans, cost management, or market positioning.

Step 2: Compare MD&A Insights with Financial Results: Assess whether the trends and ratios observed in the financial statements align with management's narrative. Example: If management highlights cost-saving initiatives, verify if this is reflected in improving operating margins.

Step 3: Identify Consistencies and Conflicts: Discuss areas where the MD&A supports the quantitative findings (e.g., strong revenue growth supported by expansion efforts). Highlight any discrepancies, such as management's optimism contrasting with declining profitability or rising debt levels.

Step 4: Discuss Transparency: Evaluate whether the MD&A provides adequate disclosure of risks and opportunities. Identify any critical missing information that could affect the reliability of the analysis.

4. Rationale for Prediction: Use findings from Trend Analysis, Ratio Analysis, and insights from the MD&A to predict future performance.

Step 1: Integrate Quantitative and Qualitative Findings: Combine trends, ratios, and MD&A insights to form a cohesive prediction.

Step 2: Predict Future EPS: State whether EPS is likely to increase, decrease, or remain stable.Justify your prediction with specific data points, ratios, and management

commentary.

Step 3: Highlight Risks and Opportunities: Discuss potential risks identified in the MD&A or financial statements (e.g., rising costs, declining asset turnover). Highlight growth opportunities or operational efficiencies that could drive future performance.

5. Confidence Score: Assess the reliability of your prediction by assigning a confidence score between 0 and 1 (0 = No confidence, 1 = Absolute confidence).

Step 1: Base your score on: The consistency and alignment of trends, ratios, and MD&A insights. Transparency and reliability of the information provided in the MD&A. External factors such as market conditions or industry risks. Step 2: Justify the Confidence Score: Provide a brief explanation of what supports or reduces your confidence. Example: "Confidence is moderate (0.7) due to improving profitability but offset by weak asset utilization."

6. Final Summary: Step 1: Consolidate the Findings: Summarize the key trends, ratio results, MD&A insights, and their implications. Recap your EPS prediction and confidence score.

Step 2: Provide Actionable Insights: Highlight the company's strengths, weaknesses, opportunities, and risks. Offer recommendations for improving performance or mitigating risks. Deliver the analysis in a structured and clear format, presenting it as if you were delivering a professional report to stakeholders.

3 Empirical Results

3.1 Benchmark and Evaluation Metrics

In this section, we evaluate the performance of the large language model in the analysis of financial statements aimed at predicting the direction of future earnings. All prediction models have a binary target variable that indicates an increase or decrease in EPS in the

following year.

We refer Kim et al. (2024a) to use a naive model as the benchmark. The naive model is designed as follows: we assume that the directional change in earnings will stay the same. In particular, if EPS has increased (decreased) in year t compared to year $t-1$, the naive prediction for year $t+1$ is also "increase" "decrease". Beisdes, we extract the analysts' forecast issued one month after the previous earnings release (Analyst 1m), three months after previous earnings release (Analyst 3m), and six months after previous earnings release (Analyst 6m) from the dataset of Kim et al. (2024a).

Moreover, we follow Ou and Penman (1989), Hunt et al. (2022), Gu et al. (2020) to use some classical machine learning and deep learning methods. We perform two different prediction exercises: logistic regression and Feedforward Neural Networks (FFNs). In both cases, we focus on 46 financial variables obtained to predict future earnings but exclude the price related variables. (eg. price-to-earnings ratio) because stock price is not financial statement information. We use a rolling five-year training window. That is, we estimate (train) the model using data from years $t-5$ to $t-1$, and apply the trained model to the year t data to generate forecasts. By doing so, we ensure that the models do not learn from the test data during the training phase. Since our sample spans from fiscal year 1962 to 2021, we train 56 distinct models for each prediction method. For both methods, the trained model yields a probability value instead of a binary variable as its output. We classify observations with a probability value higher than 0.5 as an increase (and a decrease otherwise). The deatils of parameters are shown in Appendix.

We report two common metrics to evaluate the quality of the prediction method: accuracy and F1-score. Accuracy is the percentage of correctly predicted cases scaled by the total number of predictions made. F1-score is the harmonic mean of precision and recall. Precision measures the proportion of true positive predictions in the total positive predictions, while recall measures the proportion of true positive predictions out of all actual

positives. Those two formulas are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{2}$$

3.2 Main Results

Table 3 compares models' prediction accuracy and F1 scores. The naive model achieves an lowest accuracy of 49.11% and an F1-score of 53.02%. These results reinforce the widely accepted notion that predicting earnings changes remains a highly challenging task.

Turning to GPT's predictions, we observe the following: Using a simple prompt instructing GPT to analyze financial statements and predict the direction of future earnings yields an accuracy of 52.85% and an F1-score of 54.61%. While this performance is better than the naive model, it remains on par with analyst forecasts made one month after the earnings release (Kim et al. 2024a). However, incorporating chain-of-thought (CoT) reasoning into the GPT prompt leads to an improvement, with accuracy rising to 57.35% and the F1-score reaching 56.37%. This marks a notable increase compared to analyst predictions at the one-month mark, suggesting that structured reasoning helps GPT extract more relevant insights from financial statements.

Further enhancements occur when MD&A analysis is included in the GPT prompt. This modification boosts accuracy to 60.31% and the F1-score to 60.74%, highlighting the importance of qualitative disclosures in improving earnings predictions. However, despite these gains, GPT's best-performing configuration (with CoT and MD&A) still does not outperform the feedforward neural network (FNN), which achieves the accuracy at 60.25% and the best F1-score at 61.62%. This suggests that while large language models can provide meaningful insights from unstructured data, deep learning architectures specifically

trained for numerical prediction remain more effective. This aligns with broader findings in machine learning research, where LLMs often struggle with numerical reasoning and structured data processing, as they are primarily optimized for text-based tasks rather than direct numerical computation. Similar results are observed in Kim et al. (2024a),

Moreover, when comparing GPT with traditional methods such as logistic regression, we see that the latter achieves a comparable F1-score of 57.23%, despite being a much simpler model. Even the six-month analyst consensus forecast, with an accuracy of 56.68% and an F1-score of 56.85%, is not far behind CoT-based GPT. These findings suggest that although GPT benefits from CoT and qualitative data integration, its advantage over established forecasting methods remains modest.

Our results show that while GPT demonstrates notable improvements over naive predictions and basic analyst forecasts, its best performance still lags behind FNN and is only marginally better than logistic regression and multi-month analyst consensus forecasts. This underscores the importance of combining GPT with structured financial models or hybrid approaches to enhance predictive accuracy further.

3.3 Can LLM survive recessions?

In this part, we check if LLM can identify recessions to avoid loss when predicting earnings. Figure 1 shows. We report the overall time trend of ChatGPT's and FFN's prediction accuracy in Figure 1 (detailed annual accuracy and F1-scores are reported in Appendix A). The panel indicates a declining trend in ChatGPT's prediction accuracy over time.

A key observation is that ChatGPT exhibits sharp declines in prediction accuracy during known macroeconomic downturns. Notably, we observe significant drops in 1974, 2008–2009, and 2020, which align with the 1973 oil crisis, the 2008 global financial crisis, and the COVID-19 pandemic, respectively. These results suggest that ChatGPT struggles during periods of heightened economic uncertainty, as it does not anticipate exogenous

Table 3: Comparison of accuracy and F1 scores for different prediction methods.

Method	**Accuracy**	**F1**
Naive Model	49.11%	53.02%
Analyst 1m (Kim et al. 2024a)	52.71%	54.48%
Analyst 3m (Kim et al. 2024a)	55.95%	55.33%
Analyst 6m (Kim et al. 2024a)	56.68%	56.85%
Logistic Regression	52.94%	57.23%
FNN	60.25%	61.62%
GPT (without CoT)	52.85%	54.61%
GPT (with CoT)	57.35%	58.37%
GPT (with CoT and MD&A)	60.31%	60.74%

shocks.

Most importantly, Figure 1 presents the time-series trend of the accuracy difference between ChatGPT and FFN models. The FFN model follows a similar trend. This indicates that neither ChatGPT nor FFN consistently outperforms the other across different economic conditions. The negative trend in prediction accuracy implies that forecasting future earnings using only numerical data has become increasingly challenging.

Interestingly, while ChatGPT reacts to recessions with sharp declines in accuracy, it does not appear to anticipate them beforehand. This suggests that LLMs, in their current form, primarily process historical patterns rather than detecting early warning signals of macroeconomic downturns.

4 Conclusion

Our study explored the potential of Large Language Models (LLMs) in financial statement analysis (FSA), particularly their ability to predict earnings changes. The findings provide valuable insights into the effectiveness of LLMs compared to traditional financial analysis methods, including human analysts, logistic regression, and deep learning models. By integrating Chain-of-Thought (CoT) prompting and Management Discussion and Analysis (MD&A) data, we examined whether LLMs could enhance predictive accuracy and provide more comprehensive financial insights.

The results indicate that LLMs, even in their basic form, outperform naive models and achieve prediction accuracy comparable to early analyst forecasts. However, incorporating structured reasoning through CoT prompts significantly improves performance. Notably, when narrative analysis from MD&A sections is added, the predictive accuracy of GPT-4o surpasses traditional logistic regression models and nearly matches feedforward neural networks (FNNs). These findings suggest that qualitative financial disclosures play a crucial role in enhancing LLM-driven analysis and decision-making.

Despite these advancements, several challenges remain. First, while LLMs can interpret and analyze structured financial data, they still lag behind specialized deep learning models in predictive accuracy. The inherent limitations of LLMs in handling numerical computations and complex financial patterns suggest that hybrid models—combining LLM-driven qualitative insights with robust numerical processing techniques—may offer a more effective solution. Additionally, LLMs struggle to anticipate macroeconomic shocks, as observed in their performance dips during recessions. This limitation underscores the need for models that integrate broader economic indicators and external data sources to enhance forecasting reliability.

Our research highlights the transformative potential of LLMs in financial analysis while

recognizing the necessity for continued refinement. Future research should explore ways to further enhance LLM capabilities, including improved prompt engineering, fine-tuned financial domain adaptation, and hybrid model integration. Additionally, extending the evaluation framework to incorporate real-world investment decision-making scenarios could provide deeper insights into practical applications.

In conclusion, while LLMs represent a promising advancement in financial statement analysis, they are not a standalone replacement for traditional financial models. Instead, their greatest value lies in complementing existing methodologies by offering nuanced qualitative insights alongside numerical analysis. As the field of artificial intelligence in finance continues to evolve, leveraging the strengths of both structured and unstructured data processing will be key to achieving superior financial forecasting and decision-making outcomes.

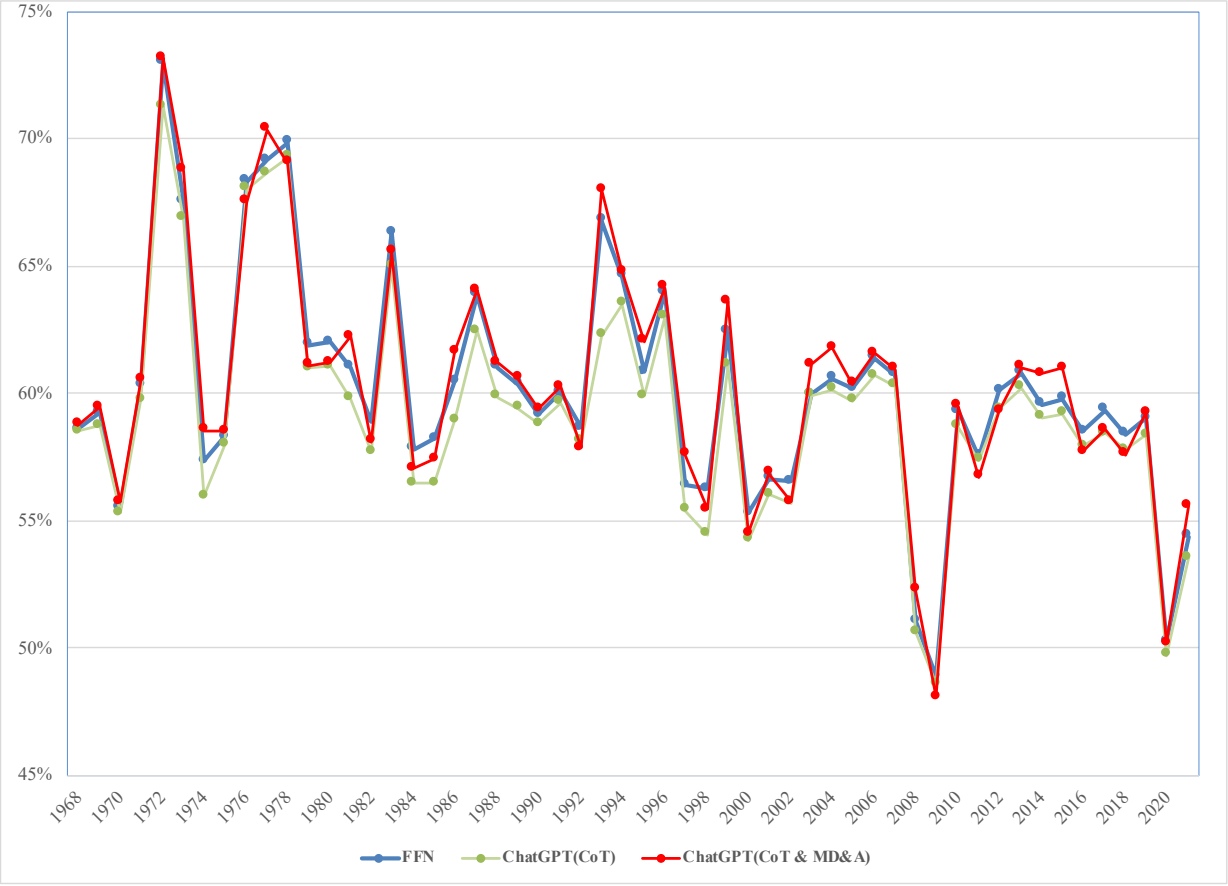

Figure 1: Time Trend for LLMs and FNN

References

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

Bybee, L. (2023). Surveying generative ai's economic expectations. *arXiv preprint arXiv:2305.02823*.

Cao, Y., Chen, Z., Pei, Q., Dimino, F., Ausiello, L., Kumar, P., Subbalakshmi, K., and Ndiaye, P. M. (2024). Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hale, K. and Wetmiller, R. J. (2024). Beyond gaap: A case study analyzing non-gaap financial measures and sec comment letters through the lens of the fasb conceptual framework. *Issues in Accounting Education*, 39(3):147–164.

Hansen, A. L. and Kazinnik, S. (2023). Can chatgpt decipher fedspeak. *Available at SSRN*.

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

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

Kim, A., Muhn, M., and Nikolaev, V. (2024a). Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Kim, A., Muhn, M., Nikolaev, V. V., and Zhang, Y. (2024b). Learning fundamentals from text. *Chicago Booth Accounting Research Center Research Paper, Fama-Miller Working Paper*.

Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., and Zohren, S. (2024). Large language models for financial and investment management: Applications and benchmarks. *Journal of Portfolio Management*, 51(2).

Lopez-Lira, A. and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.