M.S. Final Project

Financial Statement Analysis with Large Language Models: Are They Analyzing or Just Memorizing?

By

Dong Shu

Advisor: Prof. David Demeter

Committee: Prof. Han Liu

Department of Computer Science

Northwestern University

May 2025

TABLE OF CONTENTS

Chapter 1: Introduction 1

Chapter 2: Related Work 2

2.1 AI in Finance 2

2.2 LLMs in Finance 3

2.2.1 Fine-Tuned Financial LLMs 3

2.2.2 Pretrained-from-Scratch Financial LLMs 4

Chapter 3: Methodology 6

3.1 Problem Statement 6

3.2 Prompt Used 7

Chapter 4: Experiments 9

4.1 General Settings 9

4.1.1 Dataset 9

4.1.2 Metrics & Baseline 9

4.1.3 Model 10

4.1.4 Experiment Details 10

4.2 Proposed Pipeline 10

Chapter 5: Result 12

5.1 Original Input Result 12

5.2 Scaled Input Result 12

Chapter 6: Conclusion and Future Works 14

References 17

CHAPTER 1

INTRODUCTION

Large Language Models (LLMs) have rapidly transformed the landscape of artificial intelligence (AI), demonstrating remarkable capabilities across a wide range of domains, from natural language processing and code generation to creative writing and scientific discovery [1, 2, 3]. In recent years, their potential has increasingly drawn attention in the financial domain, where complex, language-rich data such as earnings reports, financial statements, and market news demand sophisticated modeling [4, 5]. Applications of LLMs in finance now span sentiment analysis, risk assessment, algorithmic trading, and financial forecasting, promising to reshape decision-making processes and enhance predictive accuracy.

Traditionally, the use of LLMs in finance has relied heavily on large-scale models, often with billions of parameters, extended context windows, or domain-specific fine-tuning on financial texts [6, 7, 8, 9]. While these approaches have led to impressive performance gains, they also come with substantial computational costs, making them difficult to deploy in resource-constrained settings or real-time applications. This raises an important question: can smaller, instruction-tuned LLMs deliver meaningful predictive power in financial tasks without the need for massive scale or extensive fine-tuning?

In this paper, we revisit this question by focusing on a relatively lightweight LLM — specifically, 1.5 billion parameter instruction-tuned models — and examine its ability to predict annual trends in company financial statements. Our study not only investigates the predictive performance of these models but also probes a critical underlying issue: are these models genuinely learning to predict financial trends, or are they merely memorizing patterns seen during training? By addressing this question, we aim to provide a more nuanced understanding of the role that compact LLMs can play in financial forecasting tasks and explore their potential as cost-effective, scalable tools for financial analysis.

CHAPTER 2
RELATED WORK

2.1 AI in Finance

The use of AI in the finance domain has a long and rich history, predating the current wave of LLMs [10, 11]. Traditional AI approaches have been widely adopted for a range of financial tasks, leveraging both structured numerical data and unstructured textual data to drive decision-making and improve forecasting accuracy [12].

In finance domain, machine learning models such as linear regression has been widely used to analyze relationships between financial indicators and firm performance, such as modeling the economic rate of return as a function of firm-level variables using data from the processing industry [13]. Support vector machines (SVMs) have been applied to financial time series forecasting, demonstrating superior performance over backpropagation neural networks in predicting futures contracts by leveraging the structural risk minimization principle [14]. Random forests have been employed to predict the direction of stock prices, particularly for clean energy ETFs, by aggregating predictions from multiple decorrelated decision trees to reduce variance and improve accuracy [15]. Gradient boosting methods have been used for fast and accurate pricing of derivatives, providing substantial speed-ups over traditional Monte Carlo simulations while maintaining practical accuracy, and offering interpretable models for practitioners [16]. Furthermore, ARIMA models, rooted in classical statistical time series analysis, have shown strong performance in short-term stock price forecasting and continue to serve as benchmarks against which more complex models are evaluated [17]. GARCH models, particularly GARCH(1,1), are extensively used in financial econometrics to model time-varying volatility, a key feature of asset returns, with implementations supported by software such as GAUSS, RATS, and TSP [18]. More recent deep learning models, such as convolutional neural networks (CNNs) [19] and recurrent neural networks (RNNs)

[20], have been applied to financial stock prediction, sentiment detection, and event extraction, enabling richer representations of linguistic context. Together, these traditional methods form the foundation upon which modern AI and machine learning techniques in finance have evolved.

Despite their success, these traditional AI approaches typically require extensive domain expertise for feature engineering, careful preprocessing of heterogeneous data sources, and specialized model design for each task. Moreover, they often struggle to generalize across companies, industries, or market conditions, limiting their adaptability and scalability in real-world financial applications. This has motivated the exploration of more general-purpose models — a role that LLMs have increasingly begun to play in the finance domain.

2.2 LLMs in Finance

With the rise of LLMs, the financial domain has experienced a surge of interest in applying these models to a wide range of language-intensive tasks. One of the most significant breakthroughs in recent LLM research has been the emergence of chain-of-thought (CoT) reasoning, which is the ability of models to generate intermediate reasoning steps that mimic human-like thought processes [21]. Rather than producing a single, direct answer, CoT reasoning enables LLMs to articulate step-by-step rationales, which improves performance on tasks requiring multi-step reasoning. This capability is particularly valuable in financial applications, where analysis often requires integrating information across time periods, synthesizing qualitative and quantitative data, and making probabilistic or causal inferences about trends and outcomes. In addition to CoT capabilities, recent research has also focused on fine-tuning LLMs or pretraining them from scratch on financial-domain data to further enhance their performance in specialized tasks.

2.2.1 Fine-Tuned Financial LLMs

Recent years have seen an explosion of fine-tuned large language models (LLMs) tailored specifically for financial applications. One major example is FinGPT, which uses a data-centric approach to fine-tune general-purpose LLMs like LLaMA and Falcon on real-time financial data from over

34 sources, including news, social media, and company filings [22]. FinGPT also introduces methods like LoRA and QLoRA to reduce fine-tuning costs and proposes reinforcement learning with stock prices (RLSP) to incorporate market feedback. Applications demonstrated with FinGPT include robo-advisors, sentiment analysis for trading, and low-code development. Similarly, PIXIU introduces FinMA, a financial LLM fine-tuned on LLaMA using a newly constructed multi-task and multi-modal instruction dataset covering sentiment analysis, headline classification, named entity recognition, question answering, and stock movement prediction [23]. It also provides the FLARE benchmark to comprehensively evaluate financial LLM performance across both NLP and prediction tasks, achieving state-of-the-art results on several financial benchmarks. Another strong example is Instruct-FinGPT, which focuses specifically on financial sentiment analysis by instruction-tuning general-purpose LLMs like LLaMA-7B [24]. By reframing sentiment classification as a generation task and applying instruction-tuned fine-tuning, Instruct-FinGPT significantly improves performance, especially in understanding numeric values and financial context, surpassing both FinBERT and general LLMs like ChatGPT on benchmark tasks.

2.2.2 Pretrained-from-Scratch Financial LLMs

Recently, several large-scale financial language models have been developed by training entirely from scratch on domain-specific corpora, aiming to capture the unique linguistic, numerical, and contextual features of financial text. One prominent example is BloombergGPT [7], a 50-billion-parameter model trained on a mixed dataset of 363 billion financial tokens and 345 billion general-domain tokens. BloombergGPT demonstrates superior performance across a wide range of financial tasks, including sentiment analysis, named entity recognition, and question answering, while maintaining competitive performance on general NLP benchmarks. The model is notable not only for its size and capabilities but also for the scale of its proprietary "FinPile" dataset, which combines Bloomberg's extensive archives of financial documents with public datasets, making it one of the most ambitious domain-specific LLM efforts to date. In the Chinese financial NLP landscape, XuanYuan 2.0 represents a milestone as the largest open-sourced Chinese financial chat model,

based on the BLOOM-176B architecture [25]. This model introduces a hybrid-tuning method that integrates general and financial-domain data during both pretraining and instruction-tuning stages, mitigating catastrophic forgetting and allowing the model to retain broad conversational abilities alongside specialized financial reasoning. XuanYuan 2.0 is designed to handle tasks like financial Q&A, report analysis, and social media monitoring, and is publicly available on Hugging Face, contributing a major open resource to the Chinese financial AI community. Another important effort is BBT-FinT5 [26], a Chinese financial language model developed under the Big Bang Transformer project. BBT-FinT5 is built on the T5 architecture, with versions up to 1 billion parameters, and is pretrained on BBT-FinCorpus, a 300GB financial corpus sourced from corporate reports, analyst notes, social media, and financial news. To benchmark the model, the team introduced BBT-CFLEB, a comprehensive Chinese financial evaluation suite covering six datasets for tasks like information extraction, sentiment analysis, and text generation. This work addresses the critical gap in high-quality Chinese financial corpora and benchmarks and highlights the need for knowledge-enhanced pretraining tailored to the financial domain.

CHAPTER 3

METHODOLOGY

3.1 Problem Statement

In this paper, we investigate whether small-scale LLMs can perform financial statement analysis (FSA) — the task of predicting the direction of a company's future earnings. Specifically, we ask whether an LLM, working solely with numerical financial data (without access to narrative context, company identity, or external market information), can extract meaningful economic insights and achieve prediction performance comparable to, or better than, human analysts and specialized machine learning models.

We follow the problem setting proposed by Kim et al. [6]. Formally, let $\mathcal{V}$ denote the vocabulary set. Given an input sequence:

$$\mathcal{X} = \{\text{instruction}, \text{data}, \text{question}\} \in \mathcal{V}^N \tag{3.1}$$

of length N, an LLM f generates a response:

$$\mathcal{Y} = \{\text{CoT}, \text{answer}\} \in \mathcal{V}^M \tag{3.2}$$

of length M. Here, the instruction prompts the LLM to follow a CoT reasoning process; the data consists of a company's historical financial records — specifically, two years of balance sheet data and three years of income statement data; and the question asks the LLM to predict the earnings trend for the next year. In the generated response, the CoT component reflects the LLM's intermediate reasoning steps, as specified by the instruction, and the answer component provides the predicted earnings direction. More formally, the model outputs answer $\in \{\text{increase}, \text{decrease}\}$, predicting the annual earnings trend for year $t+1$ based on input data consisting of balance sheets

for years $\{t, t-1\}$ and income statements for years $\{t, t-1, t-2\}$.

3.2 Prompt Used

As described in the problem statement, our approach relies heavily on prompt engineering to guide the LLM's reasoning process. In this section, we detail the exact prompts used in our experiments.

System prompt:

> *You are a highly skilled financial analyst capable of performing detailed financial statement analysis. Your objective is to analyze the provided balance sheet and income statement to predict the direction of future earnings for a company. The possible directions are: increase, decrease. Follow a systematic and logical approach to provide accurate, insightful, and comprehensive analysis.*

Instruction prompt provided to the LLM:

> *Based on the following financial statements, assess whether EPS will increase or decrease in the next year:*
>
> {*data*}
>
> *Solve this problem step by step:*
>
> *1. Analyze the Balance Sheet and Income Statement:*
>
> - *Extract balance sheet and income statement data, including year-over-year percentage changes for each item.*
> - *Analyze the balance sheet and income statement data to identify key insights and trends.*
>
> *2. Financial Ratio Analysis:*
>
> - *Calculate key financial ratios for the two most recent periods using the balance sheet and income statement data.*
> - *Analyze the calculated financial ratios to extract meaningful insights and trends.*

3. EPS Prediction for Next Year:

- *Assess whether EPS is likely to increase, or decrease based on the findings.*
- *Return the results with the following details:*
  - *Direction of EPS change — provide a one-word response:* `Increase`*, or* `Decrease`*.*
  - *Magnitude of change.*
  - *Certainty of assessment.*
  - *Reasoning behind the assessment.*

Output Format:

```
"reason": "..."
"final prediction": {"prediction": <prediction>,
                     "confidence": <confidence>}
```

CHAPTER 4

EXPERIMENTS

4.1 General Settings

4.1.1 Dataset

We conduct our experiments using financial statement data from the WRDS Compustat database, following the same setup as Kim et al. [6]. Our dataset spans annual financial data from 1971 to 1979 and includes firms listed in the United States across a wide range of industries. We select this time span because, according to Kim et al., this period yielded the highest predictive performance, with an average accuracy of 65.06% and an F1 score of 73.28% for GPT-based predictions, and manual accuracy of 64.28% and 72.71% for human analysts.

For each firm-year observation at year t within the selected period, we extract two years of balance sheet data (t and $t-1$) and three years of income statement data (t, $t-1$, and $t-2$). The full dataset contains over 15,000 unique firms, from which we randomly sample 5,000 firms for our experiments to ensure computational feasibility while maintaining diversity in firm characteristics.

4.1.2 Metrics & Baseline

Following Kim et al. [6], we evaluate model performance primarily using the F1 score, which balances precision and recall to provide a robust measure of classification performance. In their work, both accuracy and F1 score were reported as key metrics. However, Kim et al. did not explicitly address the inherent class imbalance present in real-world financial datasets. Given the imbalanced nature of our dataset, we choose not to report accuracy as a primary metric, as it can be misleading and may overstate performance by favoring the majority class. For the baseline, we directly adopt the metric scores reported in Kim et al. [6], which includes both the GPT-4–based prediction performance (1.75 trillion parameters) and the human analyst (manual) prediction performance.

4.1.3 Model

This study aims to evaluate whether small-scale LLMs, specifically models with 1.5 billion parameters, can successfully predict annual financial trends, achieving performance comparable to that of large-scale LLMs (such as GPT-4) and even human analysts. While prior research has focused on models with tens or hundreds of billions of parameters, we seek to understand whether lightweight, instruction-tuned models can also extract meaningful insights from purely numerical financial data. We selected the following 1.5B-parameter LLMs for our experiments: Qwen2.5-1.5B-Instruct, Qwen2.5-Coder-1.5B-Instruct [27], DeepSeek-R1-Distill-Qwen-1.5B [28]. All models were used in their instruction-tuned versions to align with the prompt-following setup of our task. We used the default parameter settings provided by each model, including the temperature, top-p sampling, and other decoding parameters, without further fine-tuning or modification. This allows us to assess the out-of-the-box capability of these small models in performing financial statement analysis through chain-of-thought reasoning.

4.1.4 Experiment Details

All experiments were conducted on a single NVIDIA Quadro RTX 8000 GPU with 48GB of VRAM. We directly used the models loaded from Hugging Face. To ensure the reproducibility of our results, we set the random seed to 42 across all relevant modules. All models were run in FP16 precision where supported to optimize memory usage and inference speed.

4.2 Proposed Pipeline

Our proposed pipeline consists of two main steps:

1. **Original Input:** We directly input the prompted data, including two years of balance sheet data and three years of income statement data, into the LLM and record its prediction.

2. **Scaled Input:** We systematically divide all numerical values in the balance sheet and income statement by 2 and feed this modified input into the LLM for a second prediction.

The purpose of this two-step process is to assess whether the model is genuinely analyzing the financial data or merely memorizing patterns. If the model is relying on memorization, altering the scale of the input should lead to inconsistent or degraded performance, as the underlying numeric patterns no longer match the memorized templates. Conversely, if the model is truly analyzing the relative relationships and trends within the data, its prediction should remain stable even when the absolute values change proportionally.

CHAPTER 5

RESULT

5.1 Original Input Result

Table 5.1 presents the F1 scores for all models testing on the original input data. The Baseline-GPT and Baseline-ANN scores are taken directly from Kim et al. [6], where Baseline-GPT refers to GPT-based predictions and Baseline-ANN corresponds to manual predictions by human analysts.

Our results show that while GPT achieves the highest F1 score 73.28%, the small-scale 1.5B models we tested deliver competitive performance. Notably, Qwen2.5-Coder-1.5B achieves an F1 score of 68.44%, which is particularly impressive given its significantly smaller size compared to GPT-4. This demonstrates that even small-scale LLMs can extract meaningful patterns from financial data when guided by strong instruction prompt.

We also observe that Qwen2.5-Coder outperforms the regular Qwen2.5 (68.44% vs. 63.03%). This suggests that models instruction-tuned on code data may be better suited for working with numerical, structured inputs like financial statements, and are more effective for tasks requiring precise reasoning and prediction.

Additionally, the DeepSeek-R1-Distill-Qwen-1.5B model (reported as DeepSeek-R1 in the table) achieves an F1 score of 64.27%, outperforming the regular Qwen2.5 model. This indicates that distillation can help preserve or even enhance performance, further validating the potential of small, efficient LLMs in financial forecasting tasks.

5.2 Scaled Input Result

As discussed in the previous section, Qwen2.5-Coder outperformed all other 1.5B-parameter LLMs in our experiments. Therefore, in this section, we further investigate whether this model is truly analyzing the financial data or merely memorizing patterns.

Table 5.1: LLMs performance on the original financial statement input. The highest score is shown in **bold**.

Metric	Baseline-GPT	Baseline-ANN	Qwen2.5	Qwen2.5-Coder	DeepSeek-R1
F1	**73.28**	72.71	63.03	68.44	64.27

Table 5.2: LLMs performance on the scaled financial statement input. The highest score is shown in **bold**.

Metric	Qwen2.5-Coder(Original)	Qwen2.5-Coder(Scaled)
F1	**68.44**	67.21

As shown in Table 5.2, the performance of Qwen2.5-Coder on the scaled input data remains very close to its performance on the original data, with F1 score 67.21%. This result provides strong evidence that the model is not simply memorizing absolute values but is genuinely analyzing the financial structure and relationships within the data. Remarkably, this suggests that even a relatively small 1.5B-parameter LLM is capable of performing meaningful financial analysis, reinforcing the power and flexibility of modern LLM architectures.

CHAPTER 6
CONCLUSION AND FUTURE WORKS

In this paper, we investigated whether small-scale LLMs, specifically 1.5B-parameter instruction-tuned models, can effectively perform financial statement analysis and predict annual earnings trends. By evaluating models such as Qwen2.5, Qwen2.5-Coder, and DeepSeek-R1-Distill-Qwen, we demonstrated that even compact models are capable of extracting meaningful economic insights from purely numerical financial data. Notably, Qwen2.5-Coder achieved an F1 score of 68.44%, approaching the performance of large-scale GPT models and human analysts reported in prior work. Our robustness tests, involving scaled input data, provided further evidence that these models are not simply memorizing patterns but are genuinely analyzing the relationships within financial statements. This highlights the growing potential of small-scale LLMs as efficient, practical tools for financial forecasting tasks, especially in resource-constrained environments.

For future work, we plan to extend this research in several directions. We are planning to expand the range of small-scale LLMs evaluated, including newer architectures and multi-modal models that can process both numerical and textual data. We are also planning on investigating whether lightweight fine-tuning (e.g., LoRA, QLoRA) on domain-specific financial data can further enhance the performance of small-scale LLMs. Overall, our study provides promising evidence that small, accessible LLMs can play a valuable role in the evolving landscape of AI-driven financial analysis.

REFERENCES

[1] B. Desai, K. Patil, A. Patil, and I. Mehta, "Large language models: A comprehensive exploration of modern ai's potential and pitfalls," *Journal of Innovative Technologies*, vol. 6, no. 1, 2023.

[2] M. U. Hadi *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, vol. 3, 2023.

[3] P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From text to transformation: A comprehensive review of large language models' versatility," *arXiv preprint arXiv:2402.16142*, 2024.

[4] Y. Li, S. Wang, H. Ding, and H. Chen, "Large language models in finance: A survey," in *Proceedings of the fourth ACM international conference on AI in finance*, 2023, pp. 374–382.

[5] Y. Nie *et al.*, "A survey of large language models for financial applications: Progress, prospects and challenges," *arXiv preprint arXiv:2406.11903*, 2024.

[6] A. Kim, M. Muhn, and V. Nikolaev, "Financial statement analysis with large language models," *arXiv preprint arXiv:2407.17866*, 2024.

[7] S. Wu *et al.*, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[8] A. H. Huang, H. Wang, and Y. Yang, "Finbert: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, 2023.

[9] C. Zhang *et al.*, "When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments," *arXiv preprint arXiv:2407.18957*, 2024.

[10] A. Benko and C. S. Lányi, "History of artificial intelligence," in *Encyclopedia of Information Science and Technology, Second Edition*, IGI global, 2009, pp. 1759–1762.

[11] I. Aldridge, "The ai revolution: From linear regression to chatgpt and beyond and how it all connects to finance.," *Journal of Portfolio Management*, vol. 49, no. 9, 2023.

[12] L. Cao, "Ai in finance: Challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–38, 2022.

[13] B. Lucian, A. M. Ganea, L. D. CÎRCIUMARU, *et al.*, "Using linear regression in the analysis of financial-economic performances," *Annals of University of Craiova-Economic Sciences Series*, vol. 2, no. 38, pp. 32–43, 2010.

[14] F. E. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," *omega*, vol. 29, no. 4, pp. 309–317, 2001.

[15] P. Sadorsky, "A random forests approach to predicting clean energy stock prices," *Journal of Risk and Financial Management*, vol. 14, no. 2, p. 48, 2021.

[16] J. Davis, L. Devos, S. Reyners, and W. Schoutens, "Gradient boosting for quantitative finance," *Journal of Computational Finance*, vol. 24, no. 4, pp. 1–40, 2021.

[17] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the arima model," in *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, IEEE, 2014, pp. 106–112.

[18] C. Brooks, "Garch modelling in finance: A review of the software options," *The Economic Journal*, vol. 107, no. 443, pp. 1271–1276, 1997.

[19] E. Hoseinzade and S. Haratizadeh, "Cnnpred: Cnn-based stock market prediction using a diverse set of variables," *Expert Systems with Applications*, vol. 129, pp. 273–285, 2019.

[20] S. Hansun and J. C. Young, "Predicting lq45 financial sector indices using rnn-lstm," *Journal of Big Data*, vol. 8, no. 1, p. 104, 2021.

[21] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[22] X.-Y. Liu, G. Wang, H. Yang, and D. Zha, "Fingpt: Democratizing internet-scale data for financial large language models," *arXiv preprint arXiv:2307.10485*, 2023.

[23] Q. Xie *et al.*, "Pixiu: A large language model, instruction data and evaluation benchmark for finance," *arXiv preprint arXiv:2306.05443*, 2023.

[24] B. Zhang, H. Yang, and X.-Y. Liu, "Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models," *arXiv preprint arXiv:2306.12659*, 2023.

[25] X. Zhang and Q. Yang, "Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters," in *Proceedings of the 32nd ACM international conference on information and knowledge management*, 2023, pp. 4435–4439.

[26] D. Lu *et al.*, "Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark," *arXiv preprint arXiv:2302.09432*, 2023.

[27] A. Yang *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[28] A. Liu *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.