Temporal Data Meets LLM - Explainable Financial Time Series Forecasting

Xinli Yu
xinliy@amazon.com
Amazon AWS
Seattle, WA, USA

Zheng Chen
zgchen@amazon.com
Amazon Alexa AI
Seattle, WA, USA

Yuan Ling
yualing@amazon.com
Amazon Benchmarking
Seattle, WA, USA

Shujing Dong
shujdong@amazon.com
Amazon Benchmarking
Seattle, WA, USA

Zongyi Liu
joeliu@amazon.com
Amazon Benchmarking
Seattle, WA, USA

Yanbin Lu
luyanbin@amazon.com
Amazon Alexa AI
Seattle, WA, USA

ABSTRACT

This paper presents a novel study on harnessing Large Language Models' (LLMs) outstanding knowledge and reasoning abilities for explainable financial time series forecasting. The application of machine learning models to financial time series comes with several challenges, including the difficulty in cross-sequence reasoning and inference, the hurdle of incorporating multi-modal signals from historical news, financial knowledge graphs, etc., and the issue of interpreting and explaining the model results. In this paper, we focus on NASDAQ-100 stocks, making use of publicly accessible historical stock price data, company metadata, and historical economic/financial news. We conduct experiments to illustrate the potential of LLMs in offering a unified solution to the aforementioned challenges. Our experiments include trying zero-shot/few-shot inference with GPT-4 and instruction-based fine-tuning with a public LLM model Open LLaMA. We demonstrate our approach outperforms a few baselines, including the widely applied classic ARMA-GARCH model and a gradient-boosting tree model. Through the performance comparison results and a few examples, we find LLMs can make a well-thought decision by reasoning over information from both textual news and price time series and extracting insights, leveraging cross-sequence information, and utilizing the inherent knowledge embedded within the LLM. Additionally, we show that a publicly available LLM such as Open-LLaMA, after fine-tuning, can comprehend the instruction to generate explainable forecasts and achieve reasonable performance, albeit relatively inferior in comparison to GPT-4.

CCS CONCEPTS

• **Computing methodologies → Time series analysis**; *Natural language processing*; *Explainable AI*.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

Time Series, Temporal Data, Large Language Model, Explainable AI, Multi-Modal Learning

ACM Reference Format:
Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2018. Temporal Data Meets LLM - Explainable Financial Time Series Forecasting. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The rapid advancements in Machine Learning (ML) and Artificial Intelligence (AI) technologies over the past few years have opened up numerous opportunities and challenges across various domains, including the realm of financial markets [4, 32, 49]. In particular, the task of financial time series forecasting, a key element in strategic decision-making and policy formation, has witnessed significant technological innovations, from statistical/econometric time series techniques [2, 7, 21, 46], to machine learning techniques [31, 33, 68], to deep learning [13, 28, 34, 35, 52]. Despite these advancements, there are several inherent challenges associated with the deployment of ML/AI models in finance.

One challenge lies in the realm of **cross-sequence reasoning and inference**, a vital aspect for understanding temporal patterns and making accurate predictions. The current approaches include time-series correlation analysis [8, 11, 20, 48] and clustering [1, 3, 50]. Deep learning has recently been leveraged to learn from the complex latent dependencies among time series [25, 40, 42, 55]. Despite these advancements, existing methods have yet to effectively capture the intricate dependencies characteristic of time series data. The varying design, implementation, and data requirements of these methods further creates a barrier for their widespread application in the field.

Another notable hurdle involves handling **complex multi-modal financial temporal data** that extends beyond numeric sequences. The data may encapsulate diverse sources such as historical news, financial knowledge graphs, social media activities, and various other market indicators. There has been recent effort leveraging *statistical inference* [29], RNN/CNN with text embedding [62], *graph neural networks* [9], etc. to integrate the complex information.

Lastly, but of utmost importance, the issue of **interpretability and explainability** poses significant challenges to the trustworthiness of machine learning and deep learning models. The majority of existing deep learning models operate as black boxes, offering little insight into their decision-making processes. This lack of transparency sometimes raises concerns about the reliability of their results and impedes user trust. This is particularly relevant in sensitive fields like finance, where substantial investments and assets are at stake. There is recent study trying to understand deep-learning based predictions mainly through the attention scores [24], but such insight is still not readily human readable and still requires substantial interpretation effort.

The recent advancement of *Large Language Models* (LLMs) [5, 6, 43, 59] potentially lend us a powerful tool to address all above challenges in a unified, flexible way.

First, **LLMs can learn the complex relations among sequences**. LLMs are so far the most powerful Transformer-based models, and there has been abundant previous researches showing Transformer-based models are capable of learning the underlying complex relations among textual sequences [15, 51, 67, 69, 70] and solving quantitative problems [26, 36, 64]. It is reasonable to expect the potential of LLMs understanding complex dependencies among numeric time series augmented by temporal textual sequences.

Secondly, **LLMs have demonstrated outstanding reasoning and inference capability over multi-modal data**. By design, LLMs are proficient at learning from a broad spectrum of data sources and types. They are trained on a vast amount of texts from the internet, encompassing a wide range of topics, styles, and formats. This equips them to handle diverse input data, such as numerical, textual, structured data [53, 65]. This multi-modal data handling capability could be particularly useful for financial forecasting, where crucial information often comes from disparate sources, such as numerical market data, textual news articles, and social media posts.

Lastly, **LLMs are natural explainers that generate human readable explanations providing insight into a decision**. One of the key advantages of LLMs is their ability to generate natural language text that is coherent, contextual, and comprehensive. This allows them to provide human-readable explanations for their decisions [72]. Furthermore, through Chain-of-Thoughts (COT) or step-by-step thinking [38, 64, 71], beyond a few sentences of explanation, LLMs can even generate detailed step-by-step reasoning to reveal the decision-making process.

The following summarizes the main contributions of this paper,

- This paper takes a novel exploration to study LLMs' potential to the valuable task of explainable financial time series forecasting. For this paper, we focus on the NASDAQ-100 stock price time series. To the best of our knowledge, there is not yet public studies on this topic to date.
- We experiment with a combination of zero-shot/few-shot inference techniques with the state-of-the-art AI model GPT-4 [43], and instruction-based fine-tuning using Open LLaMA [18]. Our experiment results also show that the technique of chain-of-thoughts helps boost the performance in most of the experiments.
- We compare our proposed LLM approaches with existing methods, which include an ARMA-GARCH model and a gradient-boosting tree model. We demonstrate even zero-shot inference using GPT-4 can outperform a boosting-tree model with about ~300 features.

2 RELATED WORKS

The field of financial time series forecasting has been a subject of extensive research, with various methodologies being proposed over the years. While traditional statistical methods and machine learning techniques have made significant contributions to this field, the advent of LLMs presents new and significant potentials.

2.1 Traditional Statistical/Econometric Methods

Traditional statistical/econometric methods have long been the cornerstone of financial time series forecasting. Techniques such as ARMA-GARCH models have been widely used due to their ability to capture dependencies and volatility clustering in financial time series [2, 14, 17, 22]. These models have been extended and modified in various ways to better capture the complexities of financial markets [19, 23, 39, 56]. Other popular statistical/econometric methods for financial time series include Vector Autoregressive Models (VAM) [73], State-Space Models and the Kalman Filter [12], Diffusion Models [16], Vector Error Correction Model (VECM) [27], Dynamic Stochastic General Equilibrium (DSGE) [54], etc.

2.2 Machine Learning Techniques

With the advent of machine learning, a variety of models have been applied to financial forecasting. Decision trees, support vector machines, etc., have been actively studied for financial time series prediction [37, 45, 60, 61, 63, 66]. More recently, deep learning techniques, such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformer models, have been applied to this task, demonstrating their ability to capture complex, non-linear relationships in the data [13, 28, 34, 35, 52].

2.3 Large Language Models

The recent development of Large Language Models (LLMs) has opened up new possibilities for financial time series forecasting. LLMs, such as GPT-3 [5] and GPT-4 [43], LLaMA[58] (including Alpaca[57], Vincuna[10]), have demonstrated remarkable capabilities in reasoning and understanding complex dependencies in the heterogeneous data, and the ability to generate human-readable explanations for their decisions [38, 64, 71, 72]. However, the application of LLMs in financial time series forecasting with explanation is still a relatively unexplored area, and this paper aims to contribute to this emerging field.

3 METHODOLOGY

In this study, we focus on the NASDAQ-100 stock price time series, supplemented by metadata about the stock company and relevant financial news data concerning both the specific stock and the broader financial/economic landscape. Our primary focus is on forecasting weekly/monthly *stock returns* (defined as the percentage change in stock price from the beginning to the end of the week/month) with accompanying explanations. This focus aligns well with the expertise of Large Language Models (LLMs).

We demonstrate our structured design of prompts for LLMs and apply the state-of-the-art GPT-4 model [44] for zero-shot and few-shot inference tasks. For fine-tuning, we utilize the publicly available Open LLaMA [18]. We also incorporate the technique of Chain of Thoughts (COT) [38, 64], which has been found to enhance the effectiveness of LLMs in other research studies.

3.1 Data

3.1.1 Stock Price Data. We download daily NASDAQ-100 stock[1] price data from Yahoo Finance using yfinance package (pypi.org/project/yfinance/).

In this paper, we first normalize the numeric price time series as a percentage-change time series, and then categorize the percentage change into bins. For example, for weekly forecasting, we categorize price change between this week and last week into 12 bins "D5+", "D5", "D4", "D3", "D2", "D1", "U1", "U2", "U3", "U4", "U5", "U5+", where "D5+" means price dropping more than 5%, "Di" (i=5,4,3,2,1) means price dropping between $(i-1)\%$ and $i\%$, "U5+" means price rising more than 5%, "Ui" (i=1,2,3,4,5) means price rising between $(i-1)\%$ and $i\%$. The number of bins might vary for inference at different granularity. For example, for monthly inference, we allow i be up to 10, and we have corresponding "D10+" and "U10+" categories.

3.1.2 Company Profile Data. We use GTP-4 to generate company description, general positive/negative factors that might impact the company's stock price. Figure 1 is an example of the prompt to ask GPT-4 to generate the company profile, and the GPT-4 response.

3.1.3 Finance/Economy News Data. We use *Google Custom Search API* to obtain stock top-5 news stories on a weekly basis for each of the NASDAQ-100 stocks. After that, we use GPT-4 to generate a summary and extract keywords from each obtained news article. An example of prompt and GPT-4 response is shown in Figure 2.

In addition, a similar method is applied to obtain top-5 news stories about macro economy and finance status of each week.

To reduce input size, We further generate *meta summary and keywords* for each week using GPT-4, given all the top story summaries and keywords of the week. We only use the meta summary and the keywords in this paper's experiments. An example of meta summary and keywords is shown in Figure 3; they look similar to the example in Figure 2, but much condensed.

3.2 Instruction-Based Zero-shot/Few-shot Inference with LLMs

In zero-shot and few-shot inference, LLMs demonstrate their ability to generate responses either without any additional examples (zero-shot) or based on a minimal number of examples beyond the original training set (few-shot).

利用

In our zero-shot/few-shot inference experiment, we utilize an instruction-based prompt. The structure of our prompt, illustrated in Figure 4, includes instructions, the company profile, a historical temporal news summary/keywords sequence intermixed with the categorized stock price time series, and *cross-sequence few-shot learning examples*.

To avoid unnecessary repetition in the prompt text, we purposely provide few-shot learning examples from stocks similar to the subject of interest. This design also assists us in demonstrating that the LLM can consider cross-sequence information from various stocks. To identify similar stocks, we query GPT-4 with a question such as "List the top 3 NASDAQ stocks most similar to AAPL". A typical response, such as "MSFT, GOOGL, AMZN", showcases the LLM's comprehension of the relationships between various financial entities and concepts. In employing an LLM, we are implicitly leveraging its extensive knowledge of financial entities and concepts.

The prompt structure and instructions have been empirically tweaked. For instance, we divided the instruction into two parts, positioning them at the beginning and end of the prompt, which aids the model in better recognizing its task: to predict next week's summary and keywords, rather than summarizing historical data. The predicted summary and keywords serve as the explanation for the corresponding stock return prediction.

We also experimented with the Chain-of-Thoughts approach [38, 64, 71], i.e., the idea of "step-by-step thinking", by appending the instruction "Can you reason step by step before finalizing the output?" to the end of the prompt. To our surprise, this notably improved the performance by a few points (see Section 4.2). The result of the step-by-step thinking process in response to Figure 4 is illustrated in Figure 5, where it is evident that GPT-4 identifies a previously overlooked crucial point about "earnings reports" when explicit reasoning steps are generated.

3.3 Instruction-based Fine-tuning with Open LLaMA

We perform instruction-based fine-tuning using Open LLaMA 13B model to see how well a publicly available model could perform in comparison to GPT-4, especially after fine-tuning. The Open LLaMA 13B model, in its zero-shot inference, typically tends to replicate portions of the prompt rather than executing the instructions in the prompt effectively. Therefore, it is incapable of properly handling instruction-based prompts as shown in Figure 4 without undergoing a process of fine-tuning. Therefore we focus on fine-tuning with the Open LLaMA model in this paper.

Instruction-based fine-tuning has been recently shown to be effective in guiding the model's training process with specific directives [47, 57]. We created a dataset of 30K weekly forecasting plus 7K monthly forecasting, derived from 5-year historical data spanning from Jun 2017 to June 2022. Unlike GPT-4 that supports up to 8K token size, we need to compress the prompt into 1K tokens for fine-tuning Open LLaMA, due to model and hardware constraints. For each fine-tuning example, we employ GPT-4 to condense the full historical meta news summary/keywords (e.g. from week 8 to the last week as shown in Figure 4) into a single, even more concise summary/keywords pair. Simultaneously, the "Company Profile" and "Forecasting Examples" sections of the prompt are also respectively condensed into more succinct summary paragraphs.

[1]Using the stock list of Jun 01, 2023

Prompt: Generate a short description for stock AAPL' s company. Also list general positive and negative factors that might impact the stock price; be brief and use keywords. Consider diverse general factors, such as macro economic situation (e.g. inflation, CPI growth), business factors (e.g. sales, investment, products), technology factors (e.g. innovation), and others. Use format Description: ..., Positive Factors: ..., Negative factors: ...

GPT-4 Response:
Description:
AAPL (Apple Inc.) is a world-renowned technology company based in Cupertino, California. Known for its sleek designs and innovative technologies, the company produces a wide range of products including the iPhone, iPad, Mac, Apple Watch, and services like iCloud, App Store, and more.

Positive Factors:
Macroeconomic: Economic growth, increased consumer spending
Business: Strong sales, expansion into new markets, positive earnings reports, strategic acquisitions
Technology: Innovation, new product launches, advancements in artificial intelligence and augmented reality
Brand: Strong global brand, loyal customer base
Legal/Regulatory: Favorable policy changes, successful patent applications

Negative Factors:
Macroeconomic: Inflation, economic downturn, trade restrictions
Business: Decreased sales, poor earnings reports, lack of new investments
Technology: Slow pace of innovation, delays in product launches
Brand: Damage to brand reputation, loss of customer loyalty
Legal/Regulatory: Antitrust investigations, unfavorable policy changes, patent disputes
Environmental: Failure to meet sustainability goals, negative environmental impact

Figure 1: An example of a stock's company profile consisting of the company description, the general positive/negative factors impacting the company's stock prices.

Prompt [template, those inside {} are placeholders]: Please summarize the following noisy but possible news data extracted from web page HTML, and extract keywords of the news. The news text can be very noisy due to it is HTML extraction. Give formatted answer such as Summary: ..., Keywords: ... The news is supposed to be for {symbol} stock. You may put 'N/A' if the noisy text does not have relevant information to extract.

News: {news}

GPT-4 Response:
Summary: Apple Inc is postponing perks for some corporate divisions and expanding a cost-cutting effort, joining its peers in streamlining operations. The company proposes to forego bonuses or promotions for April, with the payments occurring only in October. Apple is also restricting hiring and leaving additional positions open after employee departure. This is triggering fear of downsizing among employees.
Keywords: Apple, bonus, hiring, cost-cutting, promotions, downsizing, employees

Figure 2: An example of news summary and keywords extracted from one news article for a stock's company (AAPL in this case), generated by GPT-4. The original news is at https://sports.yahoo.com/apple-joins-cost-cut-bandwagon-145845685.html. The prompt in this example is a template. Texts inside [ ] are comments and are not really in the prompt when we submit it to the LLM, and we will fill stock symbol and the news into the placeholders enclosed by {}.

While it would be ideal for Open LLaMA to manage its own end-to-end experiment, including the task of prompt compression for fine-tuning, we still resort to using GPT-4 right now. This is due to Open LLaMA 13B model's zero-shot summarization capability is considerably inferior to those of GPT-4 in practice. The summaries and keywords extracted by Open LLaMA 13B model often fall short of usability.

Summary: Apple Inc.'s stock is displaying rising relative strength, although shy of a key benchmark, and the company has been highlighted as a top-performing stock due to its sales of 1.3 billion iPhones and a significant $19.4 billion profit last quarter. As the NASDAQ surges with Apple as a big cap, Apple expands cost-cutting measures, postpones perks, restricts hiring, and leaves positions open, instigating downsizing concerns. The overall stock market performance is boosted by a $30 billion deposit infusion for struggling firms.
Keywords: Apple Inc., stock, relative strength, cost-cutting, downsizing, NASDAQ, $30 billion deposit infusion, iPhones, profit.

Figure 3: An example of one week's meta summary and keywords condensed from all the company's summaries and keywords from the week.

Once fine-tuned, the Open LLaMA 13B model demonstrates a much more satisfactory comprehension of the instruction, resulting in the generation of a forecast and an accompanying explanation that appears coherent. This is illustrated in Figure 6. As per the result in section 4.2, when it comes to binary classification, the Open LLaMA model's performance is competitive compared to GPT-4. However, we've noticed that the Open LLaMA model has a tendency to produce more extreme predictions, such as U5+ or D5+, which result in a relatively higher squared error.

4 EXPERIMENTS

4.1 Experiment Setup

4.1.1 Data Time Window. The details of the data used in the experiments is as described in Section 3.1. We focus on NASDAQ-100 stock return forecasting for this paper.

- The *training/fine-tuning time window* contains 5-year data from Jun 12 2017 to Jun 05 2022. The data in this time window is used for training of the baseline models, and the fine-tuning of the Open LLaMA 13B model.
- The *evaluation time window* has 52 weeks spanning from Jun 06 2022 to Jun 04 2023. The evaluation of baseline models, the zero/few-shot inference experiments with GPT-4, and the evaluation of fine-tuned Open LLaMA 13B model, are based on data in this time window.

4.1.2 Baseline Models. To evaluate the performance of our approach, we include a heuristic baseline using the most-frequent historical bin (i.e. the most frequent bin from historical weeks before the target week to forecast) as the prediction, an ARMA-GARCH model ($p = q = 1$) [39, 56], and a gradient-boosting tree model [41] implemented by LightGBM package [30]. These baseline models are trained on the training/fine-tuning data time window, and evaluated on the evaluation time window.

For the gradient-boosting tree model, we include the following features. There are total about 300 features for the tree.

(1) Historical price time series available in the daily stock price data, including open, close, min, max prices, and the daily trading volume.
(2) The average, medium, min, max, and stddev of a rolling window of size 2, 5, 10, 30, 60, 90 for the above time series.
(3) The stock sector information, and the stock historical earning obtained from Alpha Vantage (https://www.alphavantage.co/documentation/)

4.1.3 Evaluation Metrics. We perform weekly and monthly stock return forecasting with the baselines and LLM-based methods. We treat 4 weeks as one month for convenience, and therefore there are 13 "month"s in the 52-week evaluation time window.

To evaluate the performance of our forecasting models, we employ three metrics.

- Binary precision assesses the model's ability to correctly predict the general direction of stock price movement, i.e., "Up" (U) or "Down" (D).
- Bin precision, on the other hand, evaluates the model's accuracy in predicting the exact bin from a full list of bins such as "D5+", "D5", "D4", ..., "D1", "U1", ..., "U5", "U5+".
- The MSE of consecutive bin ordinals (e.g., -6 for "D5+", -5 for "D5", ..., 0 for "U1", ..., 4 for "U5", 5 for "U5+") is used to measure the average squared differences between the model's predictions and the actual values. This metric helps to understand the model's tendency to make extreme forecasts when its predictions are incorrect.

To evaluate the quality of the forecasting explanation (the predicted next-week/month summary/keywords), we employ ROGUE-1 and ROGUE-2 scores to compare with the actual summary/keywords by GPT-4 extracted from the actual top news of the next week/month.

4.2 Performance Evaluation

The results of our experiments are summarized in Table 1 and 2. Table 1 provides a comparative analysis of our LLM-based methods and the baseline models in terms of their performance in forecasting stock returns. Table 2, on the other hand, evaluates the quality of the explanations generated by the LLMs.

In summary, our results show the effectiveness of LLMs in financial time series forecasting, with "GPT-4 few-shot with COT" consistently showing the best performance in both prediction accuracy and explanation quality. The results also highlight the technique of Chain-of-Thoughts (COT) consistently boosts performance, and the potential of instruction-based fine-tuning with publicly available LLMs like Open LLaMA to achieve reasonable performance in comparison to GPT-4 through fine-tuning with COT.

4.2.1 Stock Price Forecasting. From the results of Table 1, we observe that both GPT-4 and Open LLaMA 13B model outperform the ARMA-GARCH model and the gradient-boosting tree model in terms of both binary and bin precision. GPT-4, in particular, shows superior performance in both zero-shot and few-shot settings, with

Prompt [template, those inside {} are placeholders]:
Instruction: Forecast next week stock return (price change) for symbol, given the company profile, historical weekly news summary, keywords, and stock returns, and optionally the examples from other stocks of a similar company.

The trend is represented by bins "D5+", "D5", "D4", "D3", "D2", "D1", "U1", "U2", "U3", "U4", "U5", "U5+", where "D5+" means price dropping more than 5%, D5 means price dropping between 4% and 5%, "D4" means price dropping between 3% and 4%, "U5+" means price rising more than 5%, "U5" means price rising between 4% and 5%, "D4" means price rising between 3% and 4%, etc.
Company Profile: {company_profile} [company description, general positive/negative factors as shown in Figure 1.]
Recent News: News are ordered from oldest news to latest news.
8 Weeks ago.
{meta_news_summary_company} [meta news summary and keywords of the week for the company]
{meta_news_summary_macro} [meta news summary and keywords for the macro economy and finance status of the week]
...
7 weeks ago.
...
Last week.
...
Forecasting Examples: {few_shot_learning_examples_from_similar_stocks} [We use forecasting examples from top-3 similar stocks as the few-short learning examples. The example is structured similar to this prompt, with company profile and recent news, but without the instruction texts.]

Now predict what could be the next week's Summary, Keywords, and forecast the Stock Return. The predicted Summary/Keywords should explain the stock return forecasting. You should predict what could happen next week. Do not just summarize the history. The next week stock return need not be the same as the previous week. Use format Summary: ..., Keywords: ..., Stock Return: ...

GPT-4 Response [Forecasting week 05/01/2023 to 05/07/2023]:
Summary: Apple's stock is expected to see fluctuations in the week to come against a backdrop of mixed news. The company's announced expansion into India, coupled with the introduction of high-yield savings accounts, initially provided a boost to the stock. However, growing concerns surrounding reduced PC shipments and apprehensions about the company's upcoming earnings injected a degree of uncertainty. Despite these challenges, Apple's unwavering focus on cost control and the anticipation surrounding innovative offerings, such as the forthcoming mixed-reality headset, shored up investor confidence. Consequently, the stock maintained a general upward trend throughout the week, albeit punctuated by a few dips. The broader economic climate on Wall Street also had a role to play, with a recovering sentiment in Q1 towards high-tech companies like Microsoft and Alphabet. These companies released pivotal financial reports and ground-breaking products during the period, influencing the tech sector's performance, potentially impacting Apple's stock trajectory.

Keywords: AAPL, mixed-reality headset, India expansion, high-yield savings account, cost control strategies, Bernstein, Wall Street, economic recovery, tech sector performance

Stock Return: U1

Ground Truth [Summary from actual news from 05/01/2023 to 05/07/2023 and the actual return]:
Summary: Apple's stock saw significant gains after outperforming Wall Street's fiscal second quarter targets, driven by robust iPhone sales and strong services, despite posting a decrease in revenue for the second consecutive quarter and facing potential competition in the VR headset market from Reliance Jio's JioDive. The company also noted an improvement in supply and strength in emerging markets, further boosting investor confidence.

Keywords: Apple, stock, Wall Street, fiscal second quarter, iPhone, services sales, quarterly earnings, revenue, declining revenue, emerging markets, improved supply, Virtual Reality, competition, JioDive, Reliance Jio.

Stock Return: U3

Figure 4: The prompt structure for the experiments in this paper with LLMs. We also give an example of GPT-4 response to a concrete prompt constructed from information on and before 04/30/2023. We notice the cross-sequence information and macro-economy information are obviously considered in the LLM's reasoning. The stock return forecast U1 is for the next week from 05/01/2023 to 05/07/2023. The actual market performance for AAPL is U3. We note Apple 2023 Q2 earning call happened on May 04 beat expectation, which maybe the major contributor to the higher-than-forecasting gain in the week.

Prompt: {The same prompt as in Figure 4.}
Can you reason step by step before the finalized output?

GPT-4 Response [Forecasting week 05/01/2023 to 05/07/2023]:
Given the information provided, we can see that Apple (AAPL) has had a number of positive developments and a few potential concerns in the recent past.
Positive Developments:

- Recent sales and earnings have been strong, and the company appears to be exploring new markets such as India.
- The upcoming mixed-reality headset and the "Apple Pay Later" service show that Apple is still innovating and pushing into new product areas.
- The cost-cutting measures and the plan to use 100% recycled cobalt show that Apple is taking steps to improve its efficiency and sustainability, which could be attractive to certain investors.
- Winning the lawsuit against Britain's antitrust regulator and avoiding potential penalties.
- AAPL stock has been trending upwards recently, with multiple weeks of upward movement (U5, U4, U3, U2), and only a single day of slight downward movement (D1).

Potential Concerns:

- Apple is cutting costs, delaying bonuses, and slowing down hiring in certain areas, which could potentially affect employee morale and productivity.
- The negative sentiment from analyst Toni Sacconaghi about the overestimation of Q3 and Q4 revenues by Wall Street could potentially impact the stock price.

In addition to above factors, the general direction of the stock market appears to be on an upward trajectory, as evidenced by the recent increases in the Dow Jones, S&P 500, and Nasdaq. In recent weeks, the stock performances of high-tech giants such as Microsoft and Alphabet have also shown considerable strength. Furthermore, recent positive developments in the macroeconomic environment, including a surge in consumer spending, seem encouraging.

Now, we can generate a prediction for next week's Summary, Keywords, and Stock Return. However, it's important to note that while the information provided is used as a basis, this is a prediction and the actual outcome may be different.

Summary: Apple's mixed-reality headset previews have received positive feedback from industry experts, increasing investor confidence. However, apprehension about cost-cutting measures and fears of slower hiring might add some uncertainty. Wall Street anticipates a strong earnings report, boosting stock morale. Furthermore, Apple's expansion into India continues to show promise.

Keywords: Apple, mixed-reality headset, investor confidence, cost-cutting, hiring, Wall Street, earnings report, expansion, India.

Stock Return: U2. Given the positive trends and promising developments, the stock is likely to experience upward movement. However, cost-cutting concerns may moderate gains, leading to a prediction of "U2", suggesting a rise of 1% 2% next week.

Figure 5: GPT-4 outputs its detailed reasoning steps if we simply add an instruction "Can you reason step by step before the finalized output?" to the end of the prompt in Figure 4. With detailed reasoning steps, GPT-4 captures a previously missed point "Wall Street anticipates a strong earnings report, boosting stock morale", and amends the stock return forecast as U2.

Summary: The stock market is likely to react favorably to Apple's continued focus on cost optimization and revenue generation through new product launches and market expansions. Despite facing regulatory challenges, the company's efforts towards environmental responsibility and diversification of financial offerings are likely to have a positive impact on its long-term prospects. As a result, we expect Apple's stock to see another strong week of growth, with a projected return of U5+.
Keywords: Cost-Cutting, New Product Launches, Expansion, Regulation, Supply Chain
Stock Return: U5+

Figure 6: Explainable forecasting by fine-tuned Open LLaMA 13B model.

the few-shot setting with COT achieving the best performance. In terms of MSE, "GPT-4 few-shot with COT" also achieves the lowest error, indicating that it not only best predicts the direction of the price change but also provides a more accurate estimate of the magnitude of the change.

Open LLaMA 13B model, after fine-tuning, shows competitive performance compared to GPT-4 in terms of binary precision. However, its bin precision is obviously worse, indicating it lacks competitive fine-grained reasoning capability to pick the right bin. It also tends to produce more extreme predictions, resulting in a relatively higher MSE.

4.2.2 Explanation Quality. Table 2 shows the quality of the explanations generated by the LLMs (GPT-4 and fine-tuned Open LLaMA), evaluated using ROUGE-1 and ROUGE-2 scores for both the summary (S) and keywords (K) of the news.

Again, the results show that "GPT-4 few-shot with COT" achieves the highest ROUGE scores, indicating that it generates the most relevant and accurate explanations for the predictions. Open LLaMA, after fine-tuning with COT, also shows reasonable explanation quality in parallel with GPT-4 results without COT.

5 CONCLUSION

In this paper, we explored the application of Large Language Models (LLMs) in the field of financial time series forecasting, with a particular focus on NASDAQ-100 stock return prediction. We demonstrated how LLMs, specifically GPT-4 and Open LLaMA, can be utilized to generate forecasts and provide human-readable explanations for their predictions. Our approach involved the use of structured prompts, which included company profile data, historical stock price data, and financial news data, to guide the LLMs in their forecasting tasks.

Our experimental results revealed that LLMs could surpass traditional statistical models and machine learning techniques, such as the ARMA-GARCH model and a gradient-boosting tree model, in performance. Notably, the integration of a step-by-step reasoning process, inspired by the Chain of Thought (COT) approach, significantly enhanced the performance of our LLM-based models. Moreover, our fine-tuning experiments with Open LLaMA demonstrated the feasibility of effectively tuning publicly available LLMs for this task, thereby addressing the inherent challenges of cross-sequence reasoning, multi-modal signals integration, and result interpretability in the financial sector.

In conclusion, our preliminary exploration into the application of LLMs in financial forecasting yielded promising initial results. Despite being in the early stages, the encouraging outcomes provide a strong motivation for further exploration in this direction. As we continue to expand the application of LLMs in the financial domain, we envision a future where financial forecasting is not only more precise but also more comprehensible and transparent. This development could significantly contribute to the transformation of financial decision-making across the entire sector.

Future research will dive deeper into these methodologies, including but not limited to: 1) extending the studies to include more stock indexes such as SP500 and Russell 2000, 2) integrating the research with more data types, such as the macro economy time series, stock trading volumes, and social network data, and 3) exploring the fine-tuning of larger publicly available models, such as a 30B model, to enhance reasoning capabilities.

REFERENCES

[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. Time-series clustering–a decade review. *Information systems* 53 (2015), 16–38.
[2] Torben Gustav Andersen, Richard A Davis, Jens-Peter Kreiß, and Thomas V Mikosch. 2009. *Handbook of financial time series.* Springer Science & Business Media.
[3] M Suresh Babu, N Geethanjali, and B Satyanarayana. 2012. Clustering approach to stock market prediction. *International Journal of Advanced Networking and Applications* 3, 4 (2012), 1281.
[4] Arash Bahrammirzaee. 2010. A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications* 19, 8 (2010), 1165–1195.
[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *ArXiv* abs/2005.14165 (2020).
[7] Cathy WS Chen, Feng-Chi Liu, and Mike KP So. 2011. A review of threshold time series models in finance. *Statistics and its Interface* 4, 2 (2011), 167–181.
[8] Zheng Chen, Xinli Yu, Yuan Ling, Bo Song, Wei Quan, Xiaohua Hu, and Erjia Yan. 2018. Correlated anomaly detection from large streaming data. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 982–992.
[9] Dawei Cheng, Fangzhou Yang, Sheng Xiang, and Jin Liu. 2022. Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition* 121 (2022), 108218.
[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://vicuna.lmsys.org
[11] Thomas Conlon, Heather J Ruskin, and Martin Crane. 2009. Cross-correlation dynamics in financial time series. *Physica A: Statistical Mechanics and its Applications* 388, 5 (2009), 705–714.
[12] Piet De Jong and Ben Zehnwirth. 1983. Claims reserving, state-space models and the Kalman filter. *Journal of the Institute of Actuaries* 110, 1 (1983), 157–181.
[13] Alexiei Dingli and Karl Sant Fournier. 2017. Financial time series forecasting-a deep learning approach. *International Journal of Machine Learning and Computing* 7, 5 (2017), 118–122.
[14] Feike C Drost and Theo E Nijman. 1993. Temporal aggregation of GARCH processes. *Econometrica: Journal of the Econometric Society* (1993), 909–927.
[15] Vijay Prakash Dwivedi and Xavier Bresson. 2020. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699* (2020).
[16] Jianqing Fan. 2005. A selective overview of nonparametric methods in financial econometrics. *Statist. Sci.* (2005), 317–337.
[17] Christian Francq and Jean-Michel Zakoian. 2004. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* 10, 4 (2004), 605–637.
[18] Xinyang Geng and Hao Liu. 2023. *OpenLLaMA: An Open Reproduction of LLaMA.* https://github.com/openlm-research/open_llama
[19] M Ghahramani and A Thavaneswaran. 2006. Financial applications of ARMA models with GARCH errors. *The Journal of Risk Finance* 7, 5 (2006), 525–543.
[20] Parameswaran Gopikrishnan, Vasiliki Plerou, Yan Liu, LA Nunes Amaral, Xavier Gabaix, and H Eugene Stanley. 2000. Scaling and correlation in financial time series. *Physica A: Statistical Mechanics and its Applications* 287, 3-4 (2000), 362–373.
[21] Wolfgang Härdle, Helmut Lütkepohl, and Rong Chen. 1997. A review of nonparametric time series analysis. *International statistical review* 65, 1 (1997), 49–72.
[22] Jan S Henneke, Svetlozar T Rachev, Frank J Fabozzi, and Metodi Nikolov. 2011. MCMC-based estimation of Markov Switching ARMA–GARCH models. *Applied Economics* 43, 3 (2011), 259–271.
[23] Altaf Hossain and Mohammed Nasser. 2011. Comparison of the finite mixture of ARMA-GARCH, back propagation neural networks and support-vector machines in forecasting financial returns. *Journal of Applied Statistics* 38, 3 (2011), 533–551.
[24] Tsung-Yu Hsieh, Suhang Wang, Yiwei Sun, and Vasant Honavar. 2021. Explainable multivariate time series classification: a deep neural network which learns

	Weekly			Monthly (Every 4 Weeks)		
	Binary Precision	Bin Precision	MSE	Binary Precision	Bin Precision	MSE
Most-Frequent Historical Bin	50.7%	16.4%	43.5	51.4%	17.2%	155.1
ARMA-GARCH	52.4%	11.1%	22.1	50.5%	6.2%	90.1
Gradient Boosting Tree Model	60.8%	26.4%	24.3	56.4%	17.7%	85.6
GPT-4 Zero-Shot	64.5%	31.2%	20.5	64.8%	26.0%	60.1
GPT-4 Few-Shot	65.8%	32.7%	20.6	65.3%	26.5%	58.2
GPT-4 Few-Shot w/ COT	**66.5%**	**35.2%**	**18.7**	**69.5%**	**28.6%**	**50.4**
Open LLaMA (13B) Fine-Tuned	62.2%	26.5%	23.3	60.1%	22.6%	63.3
Open LLaMA (13B) Fine-Tuned w/ COT	64.7%	30.7%	21.0	62.2%	24.4%	63.5

Table 1: Performance comparison between the baseline models and LLMs for stock price weekly/monthly forecasting.

	Weekly				Monthly (Every 4 Weeks)			
	ROUGE-1 (S)	ROUGE-2 (S)	ROUGE-1 (K)	ROUGE-2 (K)	ROUGE-1 (S)	ROUGE-2 (S)	ROUGE-1 (K)	ROUGE-2 (K)
GPT-4 Zero-Shot	0.2212	**0.0675**	0.1295	0.0447	0.2528	0.0665	0.1335	0.0657
GPT-4 Few-Shot	0.2242	0.0526	0.1304	0.0454	0.2450	0.0634	0.1348	0.0644
GPT-4 Few-Shot w/ COT	**0.2414**	0.0543	**0.2083**	**0.0869**	**0.2645**	**0.0758**	**0.2450**	**0.1025**
Open LLaMA (13B) Fine-Tuned	0.2053	0.0395	0.0927	0.0324	0.2242	0.0474	0.1167	0.0520
Open LLaMA (13B) Fine-Tuned w/ COT	0.2371	0.0434	0.1123	0.0425	0.2436	0.0536	0.1356	0.0834

Table 2: Explanation quality evaluation using ROGUE scores, using the GPT-4 summary/keyword extraction of each week's true top news from google search as the ground truth.

to attend to important variables as well as time intervals. In *Proceedings of the 14th ACM international conference on web search and data mining*. 607–615.
[25] Yuxiu Hua, Zhifeng Zhao, Rongpeng Li, Xianfu Chen, Zhiming Liu, and Honggang Zhang. 2019. Deep learning with long short-term memory for time series prediction. *IEEE Communications Magazine* 57, 6 (2019), 114–119.
[26] Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398* (2023).
[27] Søren Johansen. 1995. *Likelihood-based inference in cointegrated vector autoregressive models*. OUP Oxford.
[28] Norberto Ritzmann Júnior and Julio Cesar Nievola. 2018. A generalized financial time series forecasting model based on automatic feature engineering using genetic algorithms and support vector machine. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[29] Nont Kanungsukkasem and Teerapong Leelanupab. 2019. Financial latent Dirichlet allocation (FinLDA): Feature extraction in text and data mining for financial time series prediction. *IEEE Access* 7 (2019), 71645–71664.
[30] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).
[31] Kyoung-jae Kim. 2003. Financial time series forecasting using support vector machines. *Neurocomputing* 55, 1-2 (2003), 307–319.
[32] Boris Kovalerchuk and Evgenii Vityaev. 2005. Data mining for financial applications. *Data Mining and Knowledge Discovery Handbook* (2005), 1203–1224.
[33] Bjoern Krollner, Bruce J Vanstone, Gavin R Finnie, et al. 2010. Financial time series forecasting with machine learning techniques: a survey.. In *ESANN*.
[34] Pedro Lara-Benítez, Manuel Carranza-García, and José C Riquelme. 2021. An experimental review on deep learning architectures for time series forecasting. *International Journal of Neural Systems* 31, 03 (2021), 2130001.
[35] Tim Leung and Theodore Zhao. 2021. Financial time series analysis and forecasting with Hilbert–Huang transform feature generation and machine learning. *Applied Stochastic Models in Business and Industry* 37, 6 (2021), 993–1016.
[36] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858* (2022).
[37] Wei Li and Jian Liao. 2017. A comparative study on trend forecasting approach for stock price time series. In *2017 11th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID)*. IEEE, 74–78.
[38] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050* (2023).
[39] Junmei Ma and Xinli Yu. 2013. Research on the Pricing of the Basket Credit Default Swap. *Mathematical Computation* 2, 4 (2013).
[40] Romit Maulik, Arvind Mohan, Bethany Lusch, Sandeep Madireddy, Prasanna Balaprakash, and Daniel Livescu. 2020. Time-series learning of latent-space dynamics for reduced-order model closure. *Physica D: Nonlinear Phenomena* 405 (2020), 132368.
[41] Alexey Natekin and Alois Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* 7 (2013), 21.
[42] Nam Nguyen and Brian Quanz. 2021. Temporal latent auto-encoder: A method for probabilistic multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9117–9125.
[43] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).
[44] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
[45] Ping-Feng Pai and Chih-Sheng Lin. 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* 33, 6 (2005), 497–505.
[46] Andrew J Patton. 2012. A review of copula models for economic time series. *Journal of Multivariate Analysis* 110 (2012), 4–18.
[47] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).
[48] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A Nunes Amaral, and H Eugene Stanley. 1999. Universal and nonuniversal properties of cross correlations in financial time series. *Physical review letters* 83, 7 (1999), 1471.
[49] Yuan Qi and Jing Xiao. 2018. Fintech: AI powers financial services to improve people's lives. *Commun. ACM* 61, 11 (2018), 65–69.
[50] Sangeeta Rani and Geeta Sikka. 2012. Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications* 52, 15 (2012).
[51] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems* 33 (2020), 12559–12571.
[52] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing* 90 (2020), 106181.
[53] Li Shen, Yan Sun, Zhiyuan Yu, Liang Ding, Xinmei Tian, and Dacheng Tao. 2023. On Efficient Training of Large-Scale Deep Learning Models: A Literature Review. *arXiv preprint arXiv:2304.03589* (2023).
[54] Frank Smets and Raf Wouters. 2003. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European economic association* 1, 5 (2003), 1123–1175.
[55] Wen Song and Shigeru Fujimura. 2021. Capturing combination patterns of long- and short-term dependencies in multivariate time series forecasting. *Neurocomputing* 464 (2021), 72–82.
[56] Him Tang, Kai-Chun Chiu, and Lei Xu. 2003. Finite mixture of ARMA-GARCH model for stock price prediction. In *Proceedings of the Third International Workshop on Computational Intelligence in Economics and Finance (CIEF'2003), North Carolina, USA*. 1112–1119.

[57] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv* abs/2302.13971 (2023).

[60] Theodore B Trafalis and Huseyin Ince. 2000. Support vector machine for regression and applications to financial forecasting. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Vol. 6. IEEE, 348–353.

[61] Chih F Tsai and Sammy P Wang. 2009. Stock price forecasting by hybrid machine learning techniques. In *Proceedings of the international multiconference of engineers and computer scientists*, Vol. 1. 60.

[62] Manuel R Vargas, Beatriz SLP De Lima, and Alexandre G Evsukoff. 2017. Deep learning for stock market prediction from financial news articles. In *2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA)*. IEEE, 60–65.

[63] Jar-Long Wang and Shu-Hui Chan. 2006. Stock market trading rule discovery using two-layer bias decision tree. *Expert Systems with Applications* 30, 4 (2006), 605–611.

[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[65] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[66] Haiqin Yang, Laiwan Chan, and Irwin King. 2002. Support vector machine regression for volatile stock market prediction. In *Intelligent Data Engineering and Automated Learning—IDEAL 2002: Third International Conference Manchester, UK, August 12–14, 2002 Proceedings 3*. Springer, 391–396.

[67] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems* 34 (2021), 28877–28888.

[68] Paul D Yoo, Maria H Kim, and Tony Jan. 2005. Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, Vol. 2. IEEE, 835–841.

[69] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. *Advances in neural information processing systems* 32 (2019).

[70] Jiawei Zhang, Haopeng Zhang, Congying Xia, and Li Sun. 2020. Graph-bert: Only attention is needed for learning graph representations. *arXiv preprint arXiv:2001.05140* (2020).

[71] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

[72] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[73] Eric Zivot and Jiahui Wang. 2006. Vector autoregressive models for multivariate time series. *Modeling financial time series with S-PLUS®* (2006), 385–429.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009