LLM-Driven Stock Prediction: Capturing Market Trends with LLaMA

Rahul TAK*
Bucharest University of Economic Studies, Bucharest, Romania
[*]Corresponding author, takrahul24@stud.ase.ro

Daniel Traian PELE
Bucharest University of Economic Studies, Bucharest, Romania
Institute for Economic Forecasting, Romanian Academy
danpele@ase.ro

Abstract. *Stock price forecasting remains a challenging task due to the dynamic nature of financial markets and the influence of external factors such as investor sentiment and macroeconomic events. Traditional time series statistical models often struggle to capture complex nonlinear dependencies and market signals embedded in unstructured data. With advancements in Large Language Models (LLMs), it is now possible to integrate textual information from financial news, social media, and earnings reports to enhance predictive accuracy. In this study, we leverage the LLaMA family of LLMs to improve stock price forecasting by combining historical price with news. We evaluate the performance of LLaMA 3.3 against LLaMA 3.1 and the benchmark ARIMA model to assess its effectiveness in capturing time series patterns and textual signals. Our results indicate that LLaMA 3.3 outperforms both LLaMA 3.1 and ARIMA, demonstrating its superior capability in modeling complex financial relationships. Additionally, our analysis confirms that market sentiment has an impact on stock returns, with sentiments influencing short-term price fluctuations. By incorporating news sentiment in LLMs prompt, we achieve improved forecasts compared to models without news. This highlights the importance of integrating both structured (numerical time series) and unstructured (news sentiment) data for enhanced financial modeling. Our findings suggest that LLM-driven forecasting methods hold substantial promise for traders, analysts, and financial institutions seeking more accurate market predictions. Future work will explore fine-tuning LLaMA models for domain-specific financial tasks and improving interpretability in decision making processes.*

Keywords: LLMs, LLaMA, ARIMA, News Sentiment, Time series forecasting, Stock Price.

Introduction

Financial markets are complex and dynamic systems influenced by a multitude of factors including macroeconomic indicators, investor sentiment, and historical price movements. Traditional algorithmic trading strategies have long relied on statistical models and machine learning techniques to capture these relationships. However, these approaches often struggle to incorporate unstructured data sources, such as financial news and social media sentiment, and fail to adapt quickly to evolving market conditions. Despite the rise of algorithmic trading, 90% of hedge fund trades are still executed through hardcoded procedures (M. Chang, 2018).

Earlier technical analysis-based algorithms were built by simply replicating human trading behavior using traditional technical indicators, such as the Moving Average (MA), Relative Strength Index (RSI), and Bollinger bands (BOLL). These indicators can provide incremental information (Lo, Mamaysky & Wang, 2000) for a trading system and are still effective and widely used today. However, they suffer from instability, large drawdowns, poor generalization, and vulnerability to changing market conditions. Research by Yu et al. (2013) suggests that simple

DOI: 10.2478/picbe-2025-0043

technical trading rules become less profitable as markets become more efficient. Consequently, more advanced methods such as Pairs Trading (PST) and machine learning strategies have gained significant attention. Additionally, techniques like regression analysis have been widely applied to explore the relationships between multiple financial factors and asset prices (Chen & Fang, 2013).

Before the advent of Large Language Models (LLMs), traditional financial forecasting models primarily relied on fundamental and quantitative approaches (Zeng et al., 2023). Fundamental analysis involves three main components: (i) macroeconomic analysis (e.g., GDP, CPI) to assess economic conditions, (ii) industry analysis to estimate a company's valuation based on sector trends, and (iii) company analysis to evaluate the intrinsic value based on financial and operational performance. Statistical models, which assume linearity, stationarity, and normality, have been used for market analysis, often employing Perceptually Important Points (PIP) for dimensionality reduction and template matching to identify stock price patterns (Chen & Chen, 2016). Research by Velay and Daniel (2018) highlights correlations between market patterns and future price trends. However, these models struggled to incorporate unstructured data, such as financial reports and news articles, without extensive manual preprocessing.

Recent advancements in LLMs have revolutionized financial market analysis by enabling automated processing and interpretation of vast textual datasets. Initial research repurposed general-purpose LLMs, such as GPT models and LLaMA models , for financial applications. More specialized models, including FinBERT, BloombergGPT, and FinGPT , have since emerged, demonstrating improved performance in understanding and predicting market movements. These models, pre-trained on financial-specific corpora, exhibit greater proficiency in analyzing market sentiment and financial reports. However, effectively integrating insights from diverse data sources including historical stock prices, sentiment, and fundamental indicators remains a challenge.

This paper explores how general purpose LLM can leverage the varying amount of market information consisting of both structured and unstructured data sources to make stock price forecasts without the need to alter model architecture or training custom models for each dataset. We conduct a comparative evaluation of LLaMA 3.1, LLaMA 3.3, and the ARIMA model in predicting stock prices. As the research is currently ongoing, results for the performance evaluation of LLaMA 3.2 are under process and will be shared in the future work. Additionally, we examine the impact of news sentiment and historical price data on LLaMA's predictive ability. Our approach aims to improve trading decision-making by integrating structured (closing price time series) and unstructured financial data (news), enabling models to adapt dynamically to changing market conditions.. Furthermore, we perform news sentiment scoring using LLMs and analyze its correlation with historical stock returns, demonstrating that market sentiment significantly influences stock price movements. The results suggest that fine-tuning LLMs can further improve their predictive accuracy, allowing them to capture both time series patterns and sentiment-driven market trends more effectively. Our findings highlight the potential of LLMs in financial market forecasting and provide insights into its practical applications for traders and quantitative analysts. For ease of access and transparency all the codes used to conduct the analysis are published on Quantlet and can be accessed here .

Literature review

The field of financial time series forecasting is a critical area of research in financial markets, leveraging both traditional statistical models and advanced machine learning techniques to predict future movements. Traditional statistical methods have made significant contributions to this field.

However, the non-linear and chaotic nature of stock markets has driven the adoption of machine learning methods, including neural networks, support vector machines, and ensemble models, which offer improved predictive accuracy by capturing complex patterns. The advent of LLMs presents new and significant potential to incorporate unstructured data, multimodal learning, temporal pattern recognition, etc for making more accurate forecasts.

Statistical Methods

Traditional statistical methods have long served as the foundation for financial time series forecasting. Models such as Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) have been extensively utilized due to their effectiveness in capturing dependencies and volatility clustering within financial data (Gustav et al., 2009). Over time, these models have undergone various modifications and extensions to better account for the complexities and evolving dynamics of financial markets (Hossain & Nasser, 2011).

Machine Learning

Certain statistical models are constrained by their inherent assumption of linearity, limiting their ability to capture complex financial market dynamics. emerging machine learning algorithms offer a powerful framework for addressing non-linearity, optimizing trading system design, and adapting to dynamic environments. Among these, Artificial Neural Networks (ANNs) and Support Vector Regression (SVR) have been widely studied for financial forecasting. More recently, advanced deep learning architectures, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformer models, have been applied to financial time series prediction, demonstrating superior capability in modeling complex, non-linear dependencies within the data (Pedro et al., 2021)

Large Language Models

The advancement of Large Language Models (LLMs) has introduced new opportunities for financial time series forecasting. Several specialized LLM architectures, including Time-LLM, LLM4TS, CHRONOS, and TimesFM, have been developed to analyze time series data based on human-generated prompts. FinGPT (Yang, Liu & Wang, 2023) established an open-source framework tailored for financial LLM applications, while InvestLM (Yang, Tang & Tam, 2023) demonstrated the effectiveness of instruction tuning for investment-related tasks. Additionally, FinReport introduced an automated system for financial report generation, and AlphaFin incorporated retrieval-augmented generation (RAG) techniques to enhance financial analysis (Li et al., 2024a,b). These advancements highlight the evolution of LLMs for financial NLP modeling, expanding the capabilities of LLMs in financial applications.

Data and Methodology

Data

For the purpose of this analysis, we selected 10 assets from Yahoo finance as shown in Table 1. These assets were selected to ensure diversity in financial instruments, including financial firms, technology companies, energy indices, cryptocurrencies, and stock market indices across multiple countries. The daily historical closing price data was sourced from Yahoo Finance for each ticker the period spanning January 1, 2024, to November 31, 2024, still a difference in the number of observations is observed due to difference in the number of market open dates for each asset. For every trading day within this timeframe, the top five stock-related news headlines for each asset were retrieved from the Google News US database using the GNews API in Python. The selection of data post-December 2023 was based on the most recent training period of the LLaMA 3 series of models (Dubey et al., 2024). The closing price prediction for a given day was generated by providing the LLM with historical closing prices within a defined lag window, alongside the cumulative news headlines for that day as part of the model prompt. Additionally, a daily sentiment score data was derived by prompting LLaMA 3.3 70B Instruct to evaluate the aggregated daily news headlines on a [-1,1] scale, reflecting their estimated impact on the financial market.

Table 1. Assets and number of observations used for analysis

Ticker	Name	# Observations
XLE	Energy Select Sector SPDR Fund	231
CBU	Community Financial System Inc.	231
GOOG	Alphabet Inc.	231
^STOXX	STOXX Europe 600	232
^S&P 500	Standard and Poor 500	231
^FTSE	FTSE 100 Index	234
^GDAXI	DAX performance-index	231
^N225	Nikkei 225	224
^BSESN	S&P BSE Sensex	224
BTC-USD	Bitcoin-US Dollar	334

Source: Yahoo finance ("^" represents an index fund).

Methodology

In this study, we use instruction finetuned LLaMA 3 models with direct prompting to forecast future time series values. A *rolling window* forecasting was used to predict the next day closing price of the asset based on the past 15 days closing price and daily aggregated news headlines to generate multiple forecasts for the daily closing price of the asset. The 15-day lookback window was selected as a starting point, though further research is required to determine the optimal window size. To evaluate forecasting performance, we compare the quality of forecasts made by instruction based Zero-shot and Few-shot inference, with and without the daily news for LLaMA 3.1 8B Instruct. These results were further compared with zero-shot news-aware LLaMA 3.3 70B Instruct forecast and ARIMA model forecasts with Root Mean Squared Error (RMSE) serving as the evaluation metric. The Ollama Python framework was used to generate LLM outputs and simulated 1000 forecasts for each input prompt, the median of these thousand values is selected as the LLM prediction. We used ARIMA model as the benchmark to assess the performance of the predictions made by LlaMA models. In our experiments, we tested three different prompting styles for LLM-based forecasting, defining the system's role as a financial forecasting assistant

```
response = ollama.chat(model="llama3.*",
messages=[{"role": "system", "content": "You are a financial forecasting assistant."},
            {"role": "user", "content": prompt}])
```

In the prompts given below, variable *asset_symbol* corresponds to stock ticker symbol on Yahoo finance, *window_size* is the lookback window range which in our case is 15 days, *historical_window* is a list of closing price in the lockback window as string, *news_window* is a list of daily aggregated news headlines in the lookback window and *news_current* is the current day aggregated news. Codes used to conduct the following analysis is available on Quantlet at the following link: .

Prompt 1 (*P1*) – Historical Price-Based Forecasting: The model was prompted with the asset symbol, lookback window range, and daily historical prices. Strict formatting instructions were provided to ensure a numerical output.

```
prompt =  "Given this data:"
          "Historical prices for {asset_symbol} over the past {window_size} days:"
          "{historical_window}"
          "Based on the historical prices, predict the price for {asset_symbol} for the next
           day."
          "Provide your prediction as a single number (e.g., 38500). Do not include any text
            other than the number."
```

Prompt 2 (*P2*) – Zero-Shot Forecasting with News: The LLM received the asset symbol, lookback window range, historical prices, and the top five aggregated news headlines for the forecasted day, using a zero-shot instruction-based approach.

```
prompt =  "Given this data:"
          "Historical prices for {asset_symbol} over the past {window_size} days:"
          "{historical_window}"
          "Recent news headlines related to {asset_symbol}:"
          "{news_current}"
          "Based on the historical prices and recent news, predict the price for {asset_symbol}
           for the next day."
          "Provide your prediction as a single number (e.g., 38500). Do not include any text
           other than the number."
```

Prompt 3 (P3) – Few-Shot Forecasting with News: The model was provided with five example instances from the past 15-day lookback window, where each instance included 10 days of historical prices, aggregated daily news, and the actual closing price as the expected output. Using these examples, the model was then prompted to forecast the current day's closing price based on the asset symbol, lookback window range (10 days), historical prices, and current day's news headlines.

```
prompt =  "Following examples shows past {window_size-5} days historic prices list, recent
           news related to {asset_symbol} as input and next price as their corresponding
           output:"
          "Example 1: </input> [Prices:[{historical_window[:window_size-5]}],
                            News:{news_window[-5]}],
                 </output> {historical_window[-5]}"
                                 . . .
                                 . . .
          "Example 5: </input> [Prices:[{historical_window[4:window_size-1]}],
                            News:{news_window[-1]}],
                 </output> {historical_window[-1]}"
          "Based on the insights learned from the above examples, return the next price for
```

```
{asset_symbol} using the following input:"
"</input> [Prices:[{historical_window[:window_size-5]}], News:{news_current}]"
"Make sure to capture the effect of Prices and News for predicting the output"
"Provide your output as a single number (e.g., 38500.00). Do not include any text
 other than the number."
```


Furthermore, we also analysed the direct classification ability of the positive, neutral and negative sentiment class in predicting the direction of returns assuming that positive sentiment will give positive returns and negative sentiment will have negative returns. This classification based on only the class information was not sufficient, thus we analysed the correlation of the daily news sentiments derived from LLaMA with actual market returns. For deriving the news sentiment we specified the role of LLM as expert financial analyst and prompted LLaMA 3.3 with each day's news and the instruction to rate the sentiment from [-1, 1] based on the perceived effect of news on the market. Here, a negative score corresponds to the negative market returns and vise-versa. We used the following prompt to derive sentiment scores for the news.

```
prompt = "Given the following news about stock {asset_symbol}:"
         "{news}"
         "Provide a single sentiment score between -1 to 1 for this financial news based on its
          probable impact on the stock price of {asset_symbol} ticker"
         "Give a negative score for negative news and a positive score for positive news."
         "Do not include any text other than the number."
```

Here, *asset_symbol* is the ticker symbol for a stock on Yahoo finance and *news* is the concatenation of daily news headlines. We analysed the returns in positive, neutral and negative sentiment class, and computed the correlation of sentiment scores and percentage returns for actual stock closing price at different lags.

Metrics

We employ both regression and classification metrics to comprehensively evaluate predictive accuracy of the LLM given prediction and classification effectiveness of the returns direction based on generated news sentiment. Specifically, we utilize Root Mean Squared Error (RMSE) for regression tasks and Accuracy, Precision, Recall, and F1-score for classification tasks.

1. Root Mean Squared Error (RMSE)

RMSE is a widely used metric for evaluating the performance of regression models. It measures the root of average magnitude of squared errors between predicted closing price for the day $\hat{y}_i$ and actual closing price values y_i, penalizing larger errors more than smaller ones. The RMSE is computed as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

where n is the total number of observations, y_i represents the actual values, and $\hat{y}_i$ represents the predicted values. A lower RMSE indicates better predictive accuracy.

2. Accuracy

Accuracy is a fundamental metric for classification tasks, measuring the proportion of correctly classified instances relative to the total number of predictions. It is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

where: TP (True Positives): Correctly predicted positive instances, TN (True Negatives): Correctly predicted negative instances, FP (False Positives): Incorrectly predicted positive instances, FN (False Negatives): Incorrectly predicted negative instances. Higher accuracy indicates a better-performing classification model, but it may not be reliable for imbalanced datasets.

3. Precision

Precision measures the proportion of correctly identified positive instances out of all instances classified as positive. It is given by:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

A high precision score indicates that the model produces fewer false positives, making it suitable for applications where minimizing false alarms is crucial.

4. Recall

Recall (or Sensitivity) quantifies the ability of the model to correctly identify all actual positive instances. It is computed as:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

A high recall score is desirable in scenarios where missing positive cases is costly, such as medical diagnosis.

5. F1-score

The F1-score provides a balanced measure between precision and recall, making it useful when the dataset is imbalanced. It is defined as the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

A high F1-score indicates a well-balanced model with good classification performance.

Results and Discussion

This section evaluates the performance of the forecasting methods discussed earlier, applied to the asset tickers obtained from Yahoo Finance given in Table 1. For each asset, 1,000 forecasts of the next-day closing price were generated based on historical prices and news data. The median forecast was taken as the final prediction, while the 5% and 95% quantiles were also computed to estimate the uncertainty range. Figure 1 illustrates the predicted closing prices for the N225 index, using different LLaMA architectures and prompting styles. The red dashed line represents the median forecast, the shaded grey region indicates the 90% quantile range, and the blue line corresponds to the actual asset price on a given day.

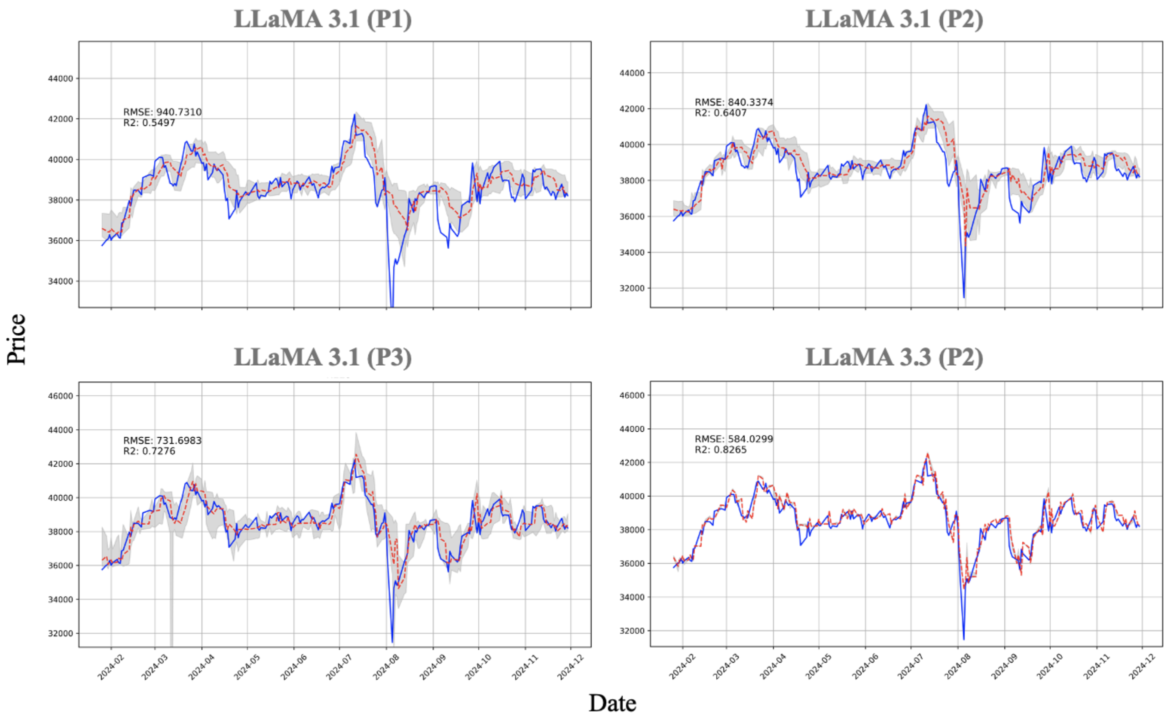

Figure 1. Quantile forecast for Nikkei 225 index using LLaMA for different prompts

Source: Authors' own research.

In Table 2, we computed the RMSE error for each ticker using different methods discussed before, cells corresponding to the best performance model for each ticker are shaded dark grey and the second best performance is highlighted with the lighter shade. The performance metrics indicate that the LLaMA 3.3 70B Instruct model achieved the lowest RMSE score, making it the most accurate forecasting model among those evaluated. Additionally, this model exhibited the smallest spread in the predicted distribution, suggesting higher confidence in its outputs generated by the LLM.

- The LLaMA 3.1 8B Instruct model with multi-shot learning mostly outperformed its zero-shot counterpart, highlighting the effectiveness of incorporating few relevant examples to improve predictive accuracy.
- The LLaMA 3.1 8B Instruct model without news data yielded the worst results, underscoring the importance of incorporating news sentiment into stock price forecasting using LLMs.
- The ARIMA model, used as a benchmark, was outperformed by the LLaMA 3.3 70B Instruct model with zero-shot news-aware forecasting for all assets. Additionally, the LLaMA 3.1 8B multi-shot model delivered superior performance to ARIMA in most cases.

Table 2. Price forecast performance: RMSE for prediction using different methods

Asset	LLaMA 3.3 (P2)	ARIMA	LLaMA 3.1 (P3)	LLaMA 3.1 (P2)	LLaMA 3.1 (P1)
XLE	1.0689	1.4229	1.3045	1.5807	1.9145
CBU	1.2564	1.6682	1.4513	1.8372	2.0669
GOOG	3.1390	3.8061	4.2904	4.3582	5.1576
^STOXX	4.1183	4.9162	4.3697	4.1187	5.4029
^S&P 500	49.1617	53.6039	61.7586	71.0451	66.3374

Asset	LLaMA 3.3 (P2)	ARIMA	LLaMA 3.1 (P3)	LLaMA 3.1 (P2)	LLaMA 3.1 (P1)
^FTSE	53.7127	68.4965	59.5877	58.2894	73.5906
^GDAXI	171.4903	211.7257	188.8334	261.9454	288.9932
^N225	584.0299	715.7150	731.6983	840.3374	940.7310
^BSESN	776.8133	790.3234	822.0617	878.1615	1024.9329
BTC-USD	2008.2004	2870.0984	2196.6372	2370.9456	2805.4461

Source: Authors' own research.

To assess the impact of news on market performance, we generated sentiment scores by prompting LLaMA 3.3 70B Instruct with daily aggregated news headlines. A preliminary analysis was conducted on the distribution of returns across different sentiment classes, direct return direction prediction ability of the sentiment score and cross-correlation between daily sentiment scores and percentage returns at various time lags

Table 3. Average percentage returns for different news sentiment class

Sentiment	CBU	^GDAXI	GOOG	XLE	^STOXX	^N225	^BSESN	^FTSE	BTC-USD
Negative	-0.5284	-0.4570	-0.6022	-0.2125	-0.2748	-0.8837	-0.4741	-0.3144	-1.8155
Neutral	0.1321	0.4215	-	0.0130	-0.0427	-	0.2882	-	-
Positive	0.3427	0.2780	0.2772	0.0831	0.2465	0.4631	0.3374	0.2473	0.5823

Source: Authors' own research.

Table 4. Returns classification using news sentiment: Classification metrics

Metric	CBU	^GDAXI	GOOG	XLE	^STOXX	^N225	^BSESN	^FTSE	BTC-USD
Accuracy	0.2217	0.6502	0.6321	0.4651	0.6620	0.6618	0.7039	0.7083	0.5942
Precision	0.1468	0.4237	0.5829	0.2897	0.4359	0.6549	0.4635	0.7028	0.5784
Recall	0.3822	0.4577	0.6012	0.3626	0.4559	0.6739	0.4877	0.7169	0.6412
F1 Score	0.1897	0.4302	0.5810	0.2945	0.4449	0.6494	0.4715	0.7014	0.5347

Source: Authors' own research.

As shown in Table 3 and Table 4, negative sentiment was associated with the lowest average returns, while neutral sentiment corresponded to higher returns, and positive sentiment yielded the highest average returns. This pattern was consistent across all assets and aligns with findings from previous research (Xu et al, 2022; Alamsyah et al., 2019). Although accuracy scores exceeded 0.50 for most assets, the F1-score was not sufficiently high to establish a strong directional relationship between sentiment classification and market movement. To incorporate both magnitude and directionality, cross-correlation was computed between sentiment scores and percentage returns for lags ranging from -5 to +5. The cells with highest correlation for each ticker are highlighted with darker shades and the second largest with lighter shade and as shown in Table 5. These results suggests that:

- Same-day sentiment scores exhibited the highest correlation with same-day returns, suggesting that news sentiment is immediately reflected in asset prices.
- A positive correlation was observed at lag -1, implying that previous-day returns may influence the sentiment of current-day news.
- A similar correlation at lag +1 suggests that current-day news sentiment may impact next-day closing prices.

These findings suggest that news sentiment has a rapid and observable effect on market prices, often materializing within hours.

Table 5. Correlation in news sentiment and percentage returns at different lags

Ticker	-5	-4	-3	-2	-1	0	1	2	3	4	5
XLE	0.1551	0.0699	-0.0609	-0.0825	0.0193	0.1537	0.1487	0.0824	0.0806	0.0622	0.1783
CBU	0.0071	-0.0038	0.0812	0.0168	0.0080	0.2050	0.1694	0.0135	0.0499	0.0900	0.0160
GOOG	0.0313	-0.1115	-0.0953	0.0687	0.1836	0.3187	0.1688	-0.0252	-0.0173	-0.0034	-0.0093
^STOXX	-0.0665	-0.0557	-0.0434	-0.0650	0.3797	0.5356	0.0537	0.0197	0.0014	0.0413	-0.0430
^FTSE	-0.0624	-0.0376	-0.0105	-0.0664	0.2902	0.6302	0.2069	0.0417	-0.0140	0.0596	-0.0506
^GDAXI	-0.1681	-0.0633	0.0659	-0.0228	0.3179	0.4647	0.1765	0.0861	0.0335	0.0340	-0.0609
^N225	-0.0346	-0.0270	-0.0286	0.1492	0.3369	0.4468	0.0142	-0.0043	0.1632	0.0663	0.0006
^BSESN	-0.0742	-0.0643	-0.0355	0.0614	0.3353	0.4825	0.0353	0.0778	-0.0092	0.0037	-0.0980
BTC-USD	0.0084	-0.0468	-0.0896	-0.0177	0.2263	0.4126	0.3398	0.1361	0.1193	0.1489	0.1886

Source: Authors' own research.

The results demonstrate that The LLaMA 3.3 70B Instruct model consistently outperforms the ARIMA benchmark and the LLaMA 3.1 8B Instruct model in stock price forecasting, The ability to incorporate additional context such as financial news into the LLM prompts enhances prediction accuracy and News sentiment plays a critical role in stock price movement, making it an essential component of effective financial forecasting models. These findings highlight the potential of large language models in financial time series forecasting and emphasize the need for sentiment-aware predictive frameworks to enhance forecast reliability.

Conclusion

In this paper, we demonstrate that general-purpose large language models (LLMs) can effectively capture both structured time series data and unstructured market sentiment from news sources yielding superior forecasts compared to traditional statistical models like ARIMA. The findings serve as a proof of concept that LLM architectures can be used to capture multi dimensional structured and unstructured information which affects the returns in the market. Market factors like public sentiment, government policies, company fundamentals, history, etc. have a significant impact on asset price movement. LLM provides us a tool to capture this diverse information without the need to train custom models for each component. The extensive pretraining of LLMs on large corpus of knowledge allows them to identify patterns in financial data, though task-specific fine-tuning could further enhance performance for particular financial applications. The increasing availability of high-performance open-source LLMs continues to drive advancements in model capabilities and real-world applications across various domains, including finance.

Current findings suggest that to effectively capture the information in time series data instruction fine tuned LLMs specifically trained on time series data should be used, as these may better capture the dependencies and patterns in historic time series data. Such specific purpose LLMs that are trained on time series data include Chronos (Ansari et al., 2024) and TimeGPT (Garza et al., 2024; Liao et al., 2024) are better suited for capturing temporal dependencies and patterns in historical price data. Additionally, integrating embeddings from specialized time series models with broader market information is essential for improving predictive accuracy. While generating multiple forecast samples to estimate the median enhances robustness, it is also computationally expensive. Therefore, employing larger models with superior predictive capabilities and output confidence may provide a more optimal balance between forecast accuracy and computational efficiency.

Future research should focus on enhancing model explainability and efficiency, particularly for high-reliability sectors like finance, where transparency and interpretability are critical. While LLMs hold significant potential for financial applications, it is crucial to validate their predictions against established theoretical frameworks to ensure reliability and practical usability. Continued advancements in LLM-based financial forecasting will depend on refining methodologies that incorporate domain-specific knowledge, improved interpretability, and computational scalability

Acknowledgement

This paper is supported through the project "IDA Institute of Digital Assets", CF166/15.11.2022, contract number CN760046/ 23.05.2023; the project "AI for Energy Finance (AI4EFin)", CF162/15.11.2022, contract number CN760048/23.05.2023, financed under the Romania's National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8; and the Marie Skłodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym DIGITAL, Project No. 101119635.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Alamsyah, A., Ayu, S. P., & Rikumahu, B. (2019, July). Exploring relationship between headline news sentiment and stock return. In 2019 7th International Conference on Information and Communication Technology (ICoICT) (pp. 1-6). IEEE.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., ... & Wang, Y. (2024). Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815.

Chen, T. L., & Chen, F. Y. (2016). An intelligent pattern recognition model for supporting investment decisions in stock market. Information Sciences, 346, 261-274.

Chen, X., & Fang, Y. (2013). Enterprise systems in financial sector–an application in precious metal trading forecasting. Enterprise Information Systems, 7(4), 558-568.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Garza, A., & Mergenthaler-Canseco, M. (2023). TimeGPT-1. arXiv preprint arXiv:2310.03589.

Hossain, A., & Nasser, M. (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), 533-551.

Lara-Benítez, P., Carranza-García, M., & Riquelme, J. C. (2021). An experimental review on deep learning architectures for time series forecasting. *International journal of neural systems*, *31*(03), 2130001.

Li, X., Shen, X., Zeng, Y., Xing, X., & Xu, J. (2024, May). FinReport: Explainable Stock Earnings Forecasting via News Factor Analyzing Model. In Companion Proceedings of the ACM on Web Conference 2024 (pp. 319-327).

Liao, W., Wang, S., Yang, D., Yang, Z., Fang, J., Rehtanz, C., & Porté-Agel, F. (2025). TimeGPT in load forecasting: A large time series model perspective. Applied Energy, 379, 124973.

Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. The journal of finance, 55(4), 1705-1765.

M. Chang. (2018). How A.I. Traders Will Dominate Hedge Fund Industry. [Online]. Available: https://www.youtube.com/watch?v=lzaBbQKUtAA

Torben Gustav Andersen, Richard A Davis, Jens-Peter Kreiu, and Thomas V Mikosch. 2009. Handbook of financial time series. Springer Science & Business Media.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
Velay, M., & Daniel, F. (2018). Stock chart pattern recognition with deep learning. arXiv preprint arXiv:1808.00418.
Xu, Y., Liang, C., Li, Y., & Huynh, T. L. (2022). News sentiment and stock return: Evidence from managers' news coverages. Finance Research Letters, 48, 102959.
Yang, H., Liu, X. Y., & Wang, C. D. (2023). Fingpt: Open-source financial large language models. arXiv preprint arXiv:2306.06031.
Yang, Y., Tang, Y., & Tam, K. Y. (2023). Investlm: A large language model for investment using financial domain instruction tuning. arXiv preprint arXiv:2309.13064.
Yu, H., Nartea, G. V., Gan, C., & Yao, L. J. (2013). Predictive ability and profitability of simple technical trading rules: Recent evidence from Southeast Asian stock markets. International Review of Economics & Finance, 25, 356-371.
Zeng, S., Yu, Y., & Li, W. (2023). Prediction and evaluation of the energy structure under the green finance development in Chongqing municipality, China. Heliyon, 9(12).