Large Language Model approach for Stock Trend Prediction using Price Action Chart

Yong-Ting Han1, a), Choo-Yee Ting1, b), Hui-Ngo Goh1, c) and Albert Quek1, d)

1Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia

b) Corresponding author: cyting@mmu.edu.my

a)1231302265@student.mmu.edu.my

c)hngoh@mmu.edu.my

d)quek.albert@mmu.edu.my

**Abstract.** Stock market trend prediction is a difficult but important task because of the complex and chaotic nature of financial markets. Traditional models are often inaccurate due to relying on numerical price data and not considering rich unstructured and multimodal information like chart patterns and textual indicators that impact the market. These weaknesses disarm them from making strong and precise predictions, especially in applications where rapid updates and high volatility exist, such as high frequency trading. To fill this gap, this paper investigates how to combine multimodal data representations leveraging pre-trained Large Language Models (LLMs) that receive numerical inputs, and vision-language models that process graphical candlestick charts. Fine-tuning applies to improve quality of the embeddings, including Low-Rank Adaptation (LoRA) for LLMs and training of multimodal classifiers to be supervised, along with optimizing for the efficiency in training by using 4-bit precision and customized hyperparameters. Experimental results show that the usage of fine-tuning both modalities as fine-tuned embeddings delivers an outstandingly improved trend classification performance, proving the effectiveness of multimodal AI in dealing with the complexity of the real-world financial markets.

# Introduction

A stock market is an essential component of global economy for enabling transactions of securities and representing economy’s condition of countries, hence accurate estimation of stock prices plays an important role in decision-making, risk reduction, and profit making. However, predicting the price of stocks is a difficult task because of the volatility in the market. Although traditional solutions such as ARIMA and GARCH are used historically in the processing of financial data, they have difficulty coping with intricacies in a market, and they are unable to reduce the market’s fluctuation effectively [1]. In the field of machine learning models such as support vector machines and neural networks are more successful than traditional systems, they have the capacity to learn better and more efficiently. However, they tend to face difficulties in handling diverse data setups and contextual variations [2]. More recently, Large Language Models (LLMs) have emerged as a promising alternative to the traditional approaches, harnessing modern transformer architectures to interpret financial news, quarterly earnings reports, sentiment in social media, and numerical indicators [3]. Due to the multimodal learning capability of modeling textual data, LLM's potential is not only to address the weaknesses of current paradigms, but also to provide a promising intuition in financial market and financial forecasting [1][2][3].

Despite some promising results in the field of stock market trend prediction, there remain persistent challenges that obstruct the development of stable, reliable and accurate predictive models of stock price. Traditional statistical and machine learning models in financial markets only based on technical indicators and historical price data, they didn't incorporate qualitative factors such as news sentiment, company earnings reports, and geopolitical events [4][5]. Such models can't be informed without an understanding of the larger context. Therefore, their observation of financial markets is one-sided.

Financial markets are affected by diverse types of data, such as textual, numerical and multimedia data. It is challenging for traditional methods to handle and analyse such multiple sources of data effectively. Since the unavailability of approaches to deal with both unstructured and multimodal data, further constrains the predictive accuracy of these models [6][7].

Furthermore, these models are very delicate to noisy and redundant information in financial markets, like speculative news and high frequency trading anomalies, often result into incorrect and inconsistent predictions [7] [8].

Additionally, large neural networks and some ensemble methods are computationally intensive during the learning and decisions, thus they are hard to be afforded by the individual investors with small scales and some firms. On the other hand, earlier models didn’t support scalability and real time processing, making them ineffective for high frequency trading and day trading with high velocity [8][9].

This study aims to investigate the possibility of Large Language Models (LLMs) to predict stock markets with alleviated limitations of other approaches. The goals are to find relevant LLMs for both graphic and text data embedding in the context of stock market analysis, optimize the proposed model parameters, and compare different LLMs performance in stock trend prediction. The work aims to push the application of LLMs for visual chart data embedding combined with textual context and develop techniques for better financial forecasting and trend prediction.

# Literature review

The stock market is extremely volatile and unpredictable, making accurate predictions of stocks become an important research problem for researchers and the investors [10][11]. Traditional techniques like time series modeling and statistical analysis are unable to model the non-linear dependencies and sentiment-based issues, that impact stock prices [12]. Recently, development in artificial intelligence has expanded the horizon for financial forecasting [11]. Advances in transformer architectures and scaling up training data have given LLMs the potential to read, generate, and analyze textual data for deriving insights from news, sentiment, and other unstructured sources of news data that influences stock prices [10]. It is shown that additional natural language data, including financial news, social media sentiment, and company reports, can improve the accuracy and robustness of stock trend prediction [12].

Besides textual data, visual representation like candlestick chart also useful to analyze and capture the market dynamics. Deep learning models, like Convolutional Neural Networks (CNNs), have been effectively utilized to capture significant patterns in the image-based candlestick chart to forecast stock trends. The models able to detect visual patterns that frequently occur before price change, such as reversals and continuation patterns [13]. Moreover, integrating visual chart analysis with sentiment or textual data has been found to enhance predicted accuracy, which means that multimodal of visual chart and textual data able to provide better description of market behavior [14].

Recent work also demonstrated the significance of fine-tuning pre-trained language models for domain-specific tasks in finance. Multitask adaptation and task-invariant representation learning have been developed using Low-Rank Adaptation (LoRA), which is designed to enable effective adaptation of large task-specific models with low overhead in additional parameters for task-oriented performance gains without the expense of full fine tuning [15]. In the meantime, compression methods such as 4-bit and 8-bit quantization are found to be effective for memory and training time consumption and allow for implementing large models in real-time financial applications [16]. In the space of multimodal learning, recent work has integrated vision-language models that jointly model textual and visual sources, achieving improved predictive performance by incorporating candlestick chart indicators with financial news (or social media sentiment) [17]. These developments suggest that embedding optimization and efficient training strategies are essential for building scalable and high-performance models in dynamic environments such as stock market forecasting.

# mETHOD

In this research, a new approach to stock market trend forecast is proposed based on Tesla stock data. The process starts with fetching the graphical and historical data from Yahoo Finance, follow by data processing and filtering to retrieve the relevant features. Both the graphical and structured numerical data are embedded by pre-trained and fine-tuned multimodal and Large Language Models (LLMs) to capture the complex patterns in the data. The embeddings are then fused to obtain a unified feature set that incorporates both visual and textual features. Lastly, train an XGBoost model based on the concatenated embeddings, which allows for more accurate predictions of stock movements.

# Dataset

The dataset is constructed from Tesla’s (TSLA) historical stock data with 5-min intervals, fetched through Yahoo Finance by using "yfinance" library and designed for both numerical and graphical trend analysis [21]. There is total 5000 samples for embedding, training, testing and evaluation. Numerical data is in the form of sliding windows of five consecutive candlesticks, each having Open, High, Low, Close, Volume and Color of the candle (Green/Red) as a feature and the corresponding labels (Uptrend or Downtrend) depending on the direction of the sixth candle’s Close price (refer to Figure 1). These sequences are saved as CSV files and transformed into structured text descriptions for Large Language Models (LLMs) embeddings. Concurrently, the graphical dataset is constructed by creating candlestick chart images using "mplfinance", with each 5-candle chart labeled according to the trend between the sixth and seventh candles. All images are in the PNG format and indexed in registration with the number data for alignments embedding and training.

A graph with red and green squares

AI-generated content may be incorrect.

**FIGURE 1.** Example of graphical data

# Framework

Referring to Figure 2, the first part involves the preparation of stock trend data form Tesla's historical prices with 5-minutes interval using both graphical and numerical approach. We get data from "yfinance" for the numerical data. Then, same stock data is employed to produce labelled candlestick chart images using "mplfinance.

To present the Tesla stock data for embedding, the graphical and numerical datasets are created. The graphical dataset is generated by extracting candlestick chart images from a ZIP file and labelling them based on future price actions. Meanwhile, the numerical dataset is created by transforming a sequence of five successive candlesticks into a structured textual description to preserve the sequence of candlesticks, which includes the open, high, low, close, volume and the color of each candle. These descriptions are labelled by the direction the next candlestick trends. Both datasets are prepared with consistent indexing to make sure alignment in the embedding fusion.

A diagram of data formatting

AI-generated content may be incorrect.

**FIGURE 2.** Research framework

Graphical data is embedded using the CLIP model by extracting the image features, then normalizing the feature to produce fixed-size embeddings. Numerical text descriptions, which result from five sequenced candlesticks, are embedded using Large Language Models (LLMs) including GPT-2, LLaMA 3, Falcon, and Mistral. These embeddings are produced through tokenization, attention-masked mean pooling, followed by extracting the output vector to obtain high dimensional vectors. However, for consistency, reduced the vectors in 512 dimensions using Principal Component Analysis (PCA). The graphical and numerical embeddings are then concatenated to generate unified 1024-dimensional vectors and normalized with StandardScaler. This joint embedding method effectively captures both visual candlestick patterns and structured market dynamics, enabling downstream tasks like classification and trend prediction.

XGBoost model serves as the prediction model trained by using embedding data that is combined with both graphical and text embedding data. CSV file with 1024-dimensional combined embeddings was loaded for training the XGBoost model and built the horizontally concatenation of 512-dimensional graphical and numerical embeddings. These embeddings are parsed from string to NumPy arrays, and the target variable “Next\_Trend” also label-encoded into binary classes (0 for Downtrend, 1 Uptrend). To maintain a consistent evaluation, the datasets are divided into a fixed training set of 4000 samples and 1000 samples for testing. The model is fitted with XGBClassifier with 1000 estimators, max depth of 6 and early stop after 50 rounds. After training, both the model and label encoder are saved to disk for future inference.

For prediction, the pre-trained XGBoost model and label encoder are loaded along with the test embedding data. The output is predicted directly on the whole test set (without chunking). The predicted stock trend labels are compared to the actual labels to determine the overall accuracy. The prediction results in terms of accuracy, precision, recall, and F1-score with macro average and weighted average are stored into a CSV file for evaluation.

The model evaluation comprises several comprehensive metrics to evaluate various aspects of the performance. These metrics include accuracy, recall, and F1-score with both macro and weighted averages, provide a balanced view of the model's effectiveness, particularly in the aspect of treating the class imbalance issue.

# Fine-tuning and training

To improving the stock trend prediction, CLIP model is fine-tuned by using the CLIP vision encoder and a customized classifier head. This method generates rich visual embeddings by training from labelled chart images of various trend categories and optionally freezing the pretrained vision encoder to use the existing visual features effectively. The classifier consists of a layer normalization, two fully connected layers with ReLU activation and dropout, is trained with cross-entropy loss to predict trend classes from the image embeddings. Large Language Models (LLMs) include GPT-2, Mistral, LLaMA 3, and Falcon are fine-tuned by a single method used for all architectures. Numerical candlestick data, such as open, high, low, close, volume, and color, is initially converted to structured natural language descriptions that summarize multi-candle patterns. For GPT-2, fine-tuning conducted in full precision with standard training and without parameter-efficient tuning, while newer models with LoRA (Low-Rank Adaptation) and 4-bit quantization via "BitsAndBytes" to reduce memory usage and more scalable. LoRA is apply to architecture-specific modules like “q\_proj” and “v\_proj” in LLaMA 3 and “query\_key\_value” in Falcon, to refine the model parameters in a targeted way.

The fine-tuned models are then used to embedding graphical stock chart images and candlestick descriptions into dense vector representations. These embeddings are concatenated into unified 1024-dimensional feature vectors for each point. To normalize the concatenated embeddings for better consistency and to increase model performance, the combined embeddings are normalized using StandardScaler. The normalized vectors define as the input to train the XGBoost model for stock trend prediction. Finally, we compare the predictive performance of the XGBoost model trained on embeddings from fine-tuned multimodal (CLIP) and LLMs to those trained on embeddings extracted from the respective pre-trained models. This comparison reflects the impact of fine-tune on embedding quality and downstream trend prediction accuracy.

# Evaluation

1. **Accuracy**

|  |  |
| --- | --- |
|  | () |

In the evaluation, accuracy is determined by comparing the predicted stock trends with the actual trends (refer to Equation (1)). The higher the accuracy, the better the model’s ability to classify stock market movements correctly.

1. **Precision**

|  |  |
| --- | --- |
|  | () |

For stock trend prediction, precision, Equation (2), helps determine how reliable the model is in predicting an uptrend or downtrend. A high precision score indicates fewer false positive predictions, meaning the model is more confident when classifying a trend as positive.

1. **Recall**

|  |  |
| --- | --- |
|  | () |

Recall is important for ensuring that the model correctly identifies all potential uptrends and downtrends (see Equation (3)). A high recall score means the model captures more of the actual positive instances, reducing the number of missed trends.

1. **F1 Score**

|  |  |
| --- | --- |
|  | () |

The F1 score (refer to Equation (4)) is useful when there is an imbalance between classes, such as when the number of uptrends significantly differs from the number of downtrends. It provides a single metric that considers both false positives and false negatives, ensuring a more balanced evaluation.

1. **Macro F1 Score**

|  |  |
| --- | --- |
|  | () |

Macro F1 Score (shown in Equation (5)) treats all classes equally by calculating the F1 score independently for each class and then averaging them. It does not take class frequency into account, making it especially useful when you want to evaluate a model’s performance without bias toward more frequent classes.

1. **Weighted F1 Score**

|  |  |
| --- | --- |
|  | () |

Weighted F1 Score accounts for class imbalance by giving each class’s F1 score a weight proportional to its frequency in the dataset (see Equation (6)). This ensures that the final score reflects the performance across all classes, with more emphasis on the ones that appear more frequently.

# findings

Among the model’s test using pre-trained multimodal and large Language Model (LLMs) embedding with XGBoost, Mistral attains the best performance overall with 55.30% accuracy, obtaining the highest scores in all the metrics except recall, demonstrating strong and balanced predicting power (refer to Table 1). GPT-2 performs closely with strong recall and solid F1 score but lowers precision than Mistral. LLaMA3 achieves balanced performance in all metrics but lags behind GPT-2 and Mistral, which makes it a reasonable middle-ground choice. On the other hand, Falcon performs the worst in all categories, including accuracy and F1, which appears to have weaker predictive capability compared to other models.

**TABLE 1.** Prediction Accuracy using pre-trained Multimodal & LLMs Embeddings with XGBoost

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **LLMs** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Macro F1** | **Weight F1** |
| Gpt2 | 52.60% | 53.89% | 57.64% | 55.70% | 52.37% | 52.48% |
| Llama3 | 53.40% | 55.07% | 53.58% | 54.31% | 53.38% | 53.41% |
| Falcon  Mistral | 51.80%  55.30% | 53.45%  56.73% | 52.42%  57.06% | 52.93%  56.89% | 51.77%  55.24% | 51.81%  55.30% |

In the evaluation of fine-tuned multimodal and LLM embeddings with XGBoost, LLaMA3 results in the best overall performance with 56.50% accuracy and achieves the highest precision, F1 score, macro F1, and weighted F1, with high overall predictive accuracy and consistency (refer to Table 2). GPT-2 does remarkably well in terms of solid recall and a decent F1 score, making it a reliable second choice. Although Falcon has the highest recall and F1, the low precision and macro-F1 indicate its potential bias towards the over-prediction of some classes. In the case of Mistral, the best performer in the pre-trained setting, now lags with moderate scores in all the metrics and showing the least improvement from fine-tuning. In general, fine-tuning benefits the most for LLaMA3, offering the best balance of performance over all metrics.

**TABLE 2.** Prediction Accuracy using Fine-tuned Multimodal & LLMs Embeddings with XGBoost

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **LLMs** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Macro F1** | **Weight F1** |
| Gpt2 | 54.70% | 56.13% | 57.64% | 56.40% | 54.63% | 54.69% |
| Llama3 | 56.50% | 58.01% | 57.45% | 57.73% | 56.46% | 56.51% |
| Falcon  Mistral | 53.40%  54.40% | 53.74%  56.19% | 70.79%  53.58% | 61.10%  54.85% | 51.50%  54.40% | 51.83%  54.41% |

# conclusion

This study show that various Large Language Models (LLMs) exhibit different capabilities in embedding numerical data and integrating it with multimodal features for trend prediction, resulting in distinct prediction accuracy. Between the pre-trained models, Mistral-7B model provided the best overall performance, obtaining the highest scores in all the evaluation measures, while Falcon had the worst performance with the lowest accuracy and F1 scores. After fine-tuning the multimodal model with the vision encoder and fine-tuning the LLMs with techniques such as LoRA, 4-bit quantization, and optimized training parameters greatly enhanced the embedding quality and predictive performance for most models. More interestingly, LLaMA3 benefited the most, it becomes top performing model after fine tuning and remained high compatibility when integrating with multimodal features for stock trend prediction. On the other hand, although Mistral excelling in the pre-trained setting, but experienced performance plummeting after fine-tuning, pointing that the current fine-tuning method may not be suitable to enhance its predictive ability. These findings also demonstrate the significance of diligent model selection and fine-tuning approaches, as not all LLMs are proficient in terms of fine-tuning, particularly in multimodal trend forecasting scenarios. Overall, the results validate that LLM is a powerful model for stock market trend prediction, able to serve as powerful tools for stock market trend prediction when effectively combined with multimodal features and fine-tuned appropriately.

# References

1. Research Scholar, Department of Computer Science and Engineering, Eklavya University, Damoh (M.P), India., S. Jain, Dr.N.K. Saluja, Professor, Department of Computer Science and Engineering, Eklavya University, Damoh (M.P), India., Dr.A. Pimplapure, Professor, Department of Computer Science and Engineering, Eklvya University, Damoh (M.P), India., Dr.R. Sahu, and Associate Professor, Department of Computer Science and Engineering, IES Group of Institutions Bhopal (M.P), India., “Exploring the Future of Stock Market Prediction through Machine Learning: An Extensive Review and Outlook,” IJISME **12**(4), 1–10 (2024).
2. J. Shi, and B. Hollifield, “Predictive Power of LLMs in Financial Markets,” (2024).
3. Anonymous. (2024). Llm-based stock market trend prediction. OpenReview.
4. M. Swamy, A. Shukla, and J. Purtilo, “LLM-based Stock Market Trend Prediction,” (2023).
5. Pathak, P. (2024). Stock market prediction using machine learning. International Journal of Financial Management and Research (IJFMR)
6. Adedoyin Tolulope Oyewole, Omotayo Bukola Adeoye, Wilhelmina Afua Addy, Chinwe Chinazo Okoye, Onyeka Chrisanctus Ofodile, and Chinonye Esther Ugochukwu, “PREDICTING STOCK MARKET MOVEMENTS USING NEURAL NETWORKS: A REVIEW AND APPLICATION STUDY,” Comput. Sci. IT Res. j. 5(3), 651–670 (2024).
7. J. Shi, and B. Hollifield, “Predictive Power of LLMs in Financial Markets,” (2024).
8. A. Lopez-Lira, and Y. Tang, “Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models,” (2024).
9. K.C. Bandhu, R. Litoriya, A. Jain, A.V. Shukla, and S. Vaidya, “An improved technique for stock price prediction on real-time exploiting stream processing and deep learning,” Multimed Tools Appl **83**(19), 57269–57289 (2024).
10. Tandon, R. (2024). Prediction of stock market trends based on large language models. Journal of 54 Emerging Technologies and Innovative Research (JETIR)
11. Vidal, J. (2024). Efficacy of ai and other large language models in predicting stock prices. SSRN.
12. T. Guo, and E. Hauptmann, “Fine-Tuning Large Language Models for Stock Return Prediction Using Newsflow,” (2024).
13. N.T. Duong, K.T. Hoang, K.Q. Duong, D.Q. Dinh, H.D. Le, T.H. Nguyen, B.X. Duong, B.Q. Tran, and A.N. Bui, “Investigating Market Strength Prediction with CNNs on Candlestick Chart Images,” in *Proceedings of the 2024 6th Asia Conference on Machine Learning and Computing*, (ACM, Bangkok Thailand, 2024), pp. 120–127.
14. T.-T. Ho, and Y. Huang, “Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation,” Sensors **21**(23), 7957 (2021).
15. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.
16. R. Jin, J. Du, W. Huang, W. Liu, J. Luan, B. Wang, and D. Xiong, “A Comprehensive Evaluation of Quantization Strategies for Large Language Models,” in *Findings of the Association for Computational Linguistics: ACL 2024*, edited by L.-W. Ku, A. Martins, and V. Srikumar, (Association for Computational Linguistics, Bangkok, Thailand, 2024), pp. 12186–12215.
17. Z. Wang, Y. Li, J. Wu, J. Soon, and X. Zhang, “FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis,” (2023).
18. Ran, A. (2020). yfinance: Download market data from Yahoo! Finance API [Computer software]. GitHub. https://github.com/ranaroussi/yfinance