Investigating the Predictive Capabilities of Large Language Models in Day Trading by Leveraging Multimodal Data

Adam Horn
IU International University of Applied Sciences
Germany
adam.horn@iu-study.org

Tim Schlippe
IU International University of Applied Sciences
Germany
tim.schlippe@iu.org

***Abstract*—This paper evaluates the predictive capabilities of six LLMs—GPT-4, GPT-4o, Llama 3, Claude 3.5, Mistral 0.3, and Gemma 2—in day trading using multimodal data. The LLMs process diverse inputs, including text-based price histories, news titles, and images. The lowest Mean Absolute Percentage Errors (MAPEs) (1.4%) were achieved by Claude 3.5 and Gemma 2 using only price history text and by Claude 3.5 and Mistral 0.3 with combined price and news history inputs, demonstrating LLMs' potential for accurate financial predictions through prompting without advanced technical expertise. Remarkably, GPT-4 and Claude 3.5 achieve MAPEs of just 1.7% and 1.5%, respectively, by processing only price history images. Furthermore, Gemma 2 achieves a MAPE of 1.5% using only news history inputs, without any information from the price history.**

***Index Terms*—large language models, LLMs, multimodal, natural language processing, NLP**

I. Introduction

Forecasting financial market movements is crucial for risk management, trading strategies, and investment decisions. Yet it remains a challenging task due to the growing complexity and volatility of global markets [1], [2]. Traditional approaches, such as technical and fundamental analysis, often struggle to account for dynamic shifts in market behavior and cannot effectively integrate the vast amounts of data now available, including news, social media, and corporate reports [3]. As a result, more advanced, data-driven techniques are needed.

Recent developments in large language models (LLMs) like GPT-4 and specialized models such as FinBERT have enabled the extraction of meaningful insights from unstructured data, improving the accuracy of financial forecasts [4]. Studies like [5] demonstrate the potential of combining machine learning with Natural Language Processing (NLP) to outperform conventional methods. However, while prior research has primarily focused on isolated data types, the challenge lies in the integration of diverse data sources to adapt to real-world conditions. Furthermore, no study has analyzed and compared multiple state-of-the-art LLMs for stock price forecasting.

To address these gaps, we investigate the capabilities of state-of-the-art LLMs to predict stock prices in day trading by leveraging multimodal data, including the price history as text, the news history as text and the price history as image. Our research builds on existing work by integrating these diverse data streams in a single model, offering a more comprehensive understanding of market trends [1], [4]. Our contributions are as follows:

- We investigate LLMs' performances to dynamically process multimodal data in daily financial forecasting.
- We analyze and compare six state-of-the-art LLMs regarding their predictive capabilities.
- Our approach requires no advanced technical knowledge, allowing users to instruct LLMs with prompts and upload files containing historical price data and relevant news.

II. Related Work

Initial financial forecasting methods use mathematical models like technical and fundamental analysis, which perform well in stable markets but poorly in dynamic conditions [1]. Recent research shows these approaches lack adaptability to complex patterns and multimodal data [1], [6].

Traditional transformer models like BERT, RoBERTa, and FinBERT are powerful for financial forecasting, effectively processing unstructured data such as news through self-attention mechanisms [1], [2]. However, these models often require extensive domain-specific fine-tuning and significant computational resources [1], [4], and they face challenges like overfitting on small datasets [1].

LLMs based on Transformer architectures, offer flexibility and ease of use through techniques like prompting, making them highly versatile for financial analysis tasks [1], [7]. Unlike traditional models, LLMs can process unstructured data efficiently, supporting tasks like sentiment analysis and news-based stock forecasting [2], [4]. [3] developed a multi-agent system combined with LLMs for alpha factor generation, outperforming existing methods on the Chinese stock market.

This research was supported by the IU International University of Applied Sciences (*IU Incubator*) through internal initial funding for the period from October 2023 to September 2025.

[8] introduced a "Denoising-then-Voting" approach, achieving high prediction accuracy for global stock indices. [7] employed GPT-4 and Gemini for predicting Federal Reserve rate decisions, showing strong predictive capabilities. [9] enhanced LLMs with QLoRA for post-earnings stock predictions, while [10] improved stock trend predictions by integrating news and price data. However, to the best of our knowledge, no analysis exists yet which investigates and compares difference state-of-the art LLMs for stock price prediction in day trading when the input are multimodal data.

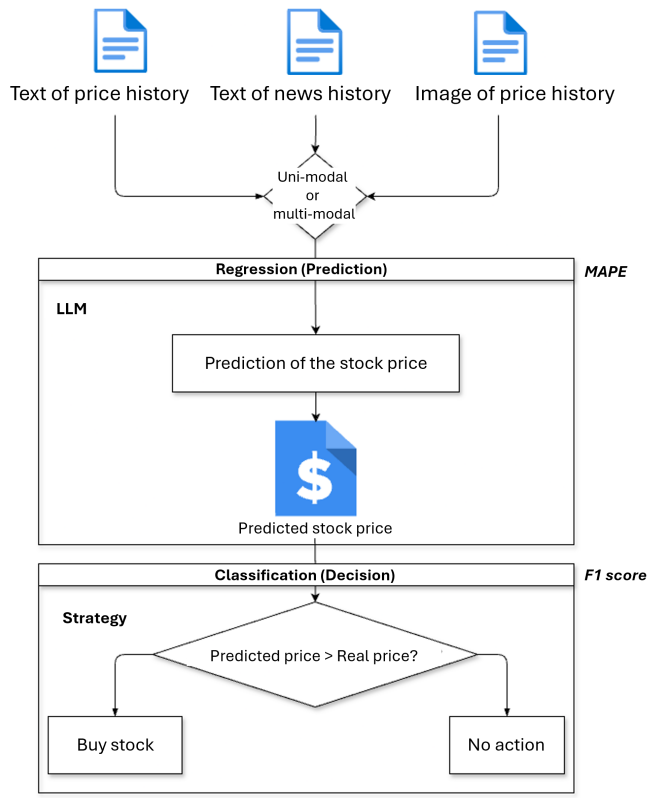

Fig. 1: Overview of our Setup.

III. Experimental Setup

In this section, we will describe our setup for the evaluation of the predictive capabilities of state-of-the-art LLMs, the data which we used for evaluation, as well as the analyzed LLMs. Finally, we will present the prompts which we elaborated to instruct the LLMs to predict the prices based on multi-modal data inputs.

A. Overview of our Setup

Figure 4 demonstrates our setup for the evaluation of the predictive capabilities of state-of-the-art LLMs. First, we instructed the LLMs to predict the price varying different input modalities:

- a text file which contains the price history of the stock in CSV format
- a text file containing the historical news titles related to the stock, listed line-by-line
- an image file of the price history of the stock
- their multimodal combinations

We evaluated prediction quality using the Mean Absolute Percentage Error (MAPE), which measures the average percentage deviation between predicted and actual prices, allowing comparison with other stock price predictions.

Since investors apply a buy strategy based on predicted prices, we also evaluated how well the LLMs implicitly predict price increases and decreases through price prediction. We evaluated this classification of price increases and decreases using the F1 score, the commonly used evaluation metric in classification tasks.

B. Data

Our dataset to evaluate the performance of LLMs in predicting financial market movements and includes historical stock prices along with news articles for Apple (AAPL). It spans from July 2 to August 9, 2024, and includes only weekdays, totaling 28. Stock price data was collected from Yahoo Finance, a well-established source offering extensive financial data, while news articles were obtained using the Reuters API[1]. The stock price data includes daily prices (e.g., opening, closing, highs, lows) and trading volumes. News data consists of the titles of the latest 20 articles per asset in JSON format, converted to text files for efficient processing and model input. TradingView.com was used for generating the images of the price history due to its advanced charting tools and real-time data feeds, providing detailed price trends for analysis.

C. Large Language Models

1) GPT-4: GPT-4, released by OpenAI in 2023, processes up to 32k tokens [11]. Its multimodal version supports image analysis and object recognition [12], [13]. It has been applied to stock prediction by integrating news and market data with stock features to improve accuracy [14].

2) GPT-4o: GPT-4o, an optimized version of GPT-4, processes up to 32k tokens more efficiently [11]. It handles multimodal input, including text and images, for tasks like image interpretation. Although no specific studies on its use in stock prediction exist, early reports indicate potential for financial analysis, similar to GPT-4 [15].

3) Llama 3: Llama 3, developed by Meta AI in 2024, processes up to 8k tokens per input [16]. The LLAMA-3-8B version with 8 billion parameters was used in this study. Though few studies focus on stock prediction, Llama 3 has shown strong performance in financial analysis, particularly when combined with models like Mistral [4]. It handles text but not images.

4) Claude 3.5: Claude 3.5 Sonnet, released by Anthropic in 2024, handles up to 200k tokens per input [17]. Though not specifically studied for stock prediction, it can process images. Its large context size makes it promising for tasks involving both text and visual data.

[1]https://api.reuters.com/news

Prompt to Predict the Closing Price Leveraging a Text of the Price History:

Attached you will find a CSV file of the daily closing prices of the <SHARE NAME> share from the period of 1 year ago until today. The last closing price of the share was <PRICE>. Based on the information in the attached file, make a prediction of what the closing price will be today which is <DATE>. Please output only the predicted price in $. Please refrain from any other outputs.

Prompt to Predict the Closing Price Leveraging a Text of the News History:

Attached you will find a file containing the last 20 news about the <SHARE NAME> share from the period of 1 year ago until today. Every news text contains the corresponding date, news title and the news content. The last closing price of the share was <PRICE>. Based on the information in the attached file, make a prediction of what the closing price will be today which is <DATE>. Please output only the predicted price in $. Please refrain from any other outputs.

Prompt to Predict the Closing Price Leveraging an Image of the Price History:

Attached you will find an image file of the daily closing prices of the <SHARE NAME> share from the period of 1 year ago until today. The last closing price of the share was <PRICE>. Based on the information in the attached file, make a prediction of what the closing price will be today which is <DATE>. Please output only the predicted price in $. Please refrain from any other outputs.

Fig. 2: Prompts to instruct LLMs to predict the closing price leveraging text of the price history, a text of the news history and an image of the price history.

5) Mistral 0.3: Mistral 0.3, released by Mistral AI in 2024, processes up to 4k tokens [18]. The 7-billion-parameter version, used in this study, is available on Hugging Face. Though not studied for stock prediction, it has been tested in other financial areas [4]. It handles text only and does not support images.

6) Gemma 2: Gemma 2, from Google DeepMind's Gemini project, processes up to 8k tokens using sliding-window attention [19]. Gemma 2 is focused on text processing and cannot handle images. The 9-billion-parameter version, used in this study, is available on Hugging Face. To the best of our knowledge, Gemma 2 has not been directly applied to stock prediction.

D. Prompts

Figure 2 demonstrates the prompts which were most successful to instruct our tested LLMs to predict the closing price of each day given text of the price history, a text of the news history and an image of the price history individually. To reach a predictions leveraging the multimodal combinations, we combined the prompts accordingly.

IV. Experiments and Results

Using the prompts outlined in Section III-D, we instructed our six LLMs to predict the daily closing price at the start of each trading day between July 2 and August 9, 2024, based on the opening price and variations of the unimodal and multimodal data described in Section III-B.

A. Stock Price Prediction

Figure 3 depicts the deviations in MAPE for the 29 predictions of the Apple stock's closing price. A lower MAPE indicates higher average prediction accuracy. As Llama 3 and Mistral 0.3 could not process images at the time of the experiments, no MAPE values are reported for these models in image contexts. The lowest MAPE of 1.4% was achieved by Claude 3.5 and Gemma 2 using only the *text of the price history*, as well as by Claude 3.5 and Mistral 0.3 with *text of the price history + text of the news history*, and Claude 3.5 with *text of the price history + text of the news history + image of the price history*. The 2nd lowest MAPE of 1.5% was obtained by Gemma 2 using only *text of the news history*, Claude 3.5 using *image of the price history*, and Gemma 2 using *text of the price history + text of the news history*. Notably, the image processing and analysis capabilities of GPT-4 and Claude 3.5 are so advanced that they achieve a MAPE of just 1.7% and 1.5%, respectively, using only the *image of the price history*. In contrast, GPT-4 and GPT-4o demonstrate the highest MAPEs when relying solely on the *text of the price history*, with values reaching up to 10.5%.

B. Stock Price Increase/Decrease Prediction

Figure 3 shows the prediction quality of price increases and decreases in F1 score based on LLMs' predicted Apple stock prices. A higher F1 score indicates better prediction for the classes *increase* and *decrease*. Comparing F-scores to MAPEs reveals that most LLMs with low MAPEs provide reliable estimates of stock movement. But the highest F1 score of 83% was achieved with the classification based on the price prediction of Llama 3 and *text of price history + text of news history*.

V. Conclusion and Future Work

Our study highlights the effectiveness of LLMs in forecasting financial market movements by leveraging multimodal data. The results show that Claude 3.5 and Gemma 2 achieved the lowest MAPE of 1.4% using only price history text, underscoring the potential of simple input formats for accurate predictions. Furthermore, combining price history with news history improved performance for Claude 3.5 and Mistral 0.3, indicating that integrating diverse data sources enhances forecasting capabilities.

Future work should explore the LLMs' performance on other stocks, the inclusion of additional modalities, such as social media sentiment, and investigate LLM-based multi-agent scenarios for stock prediction.

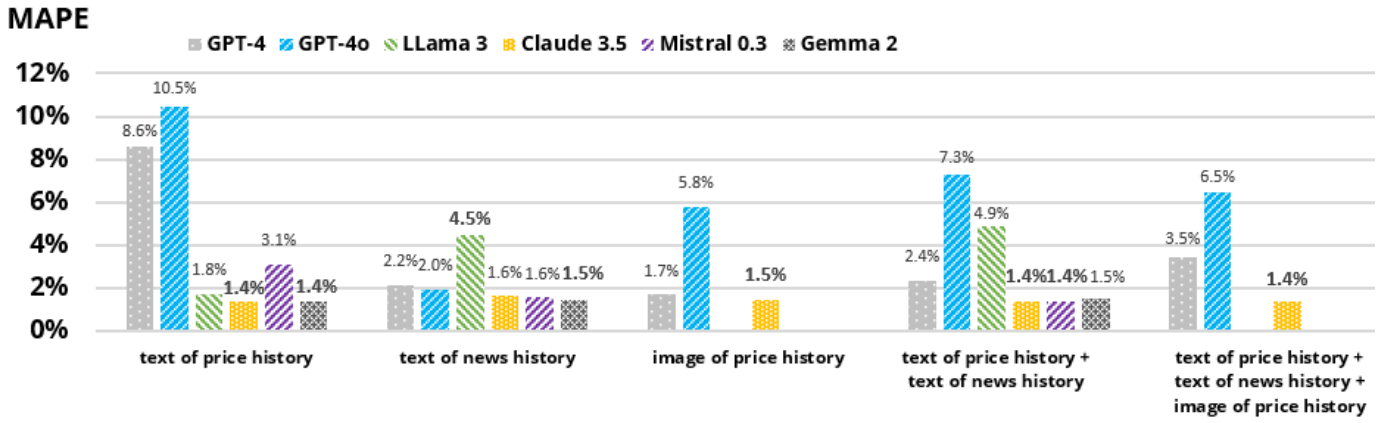

Fig. 3: LLMs' Apple stock price prediction quality in MAPE.

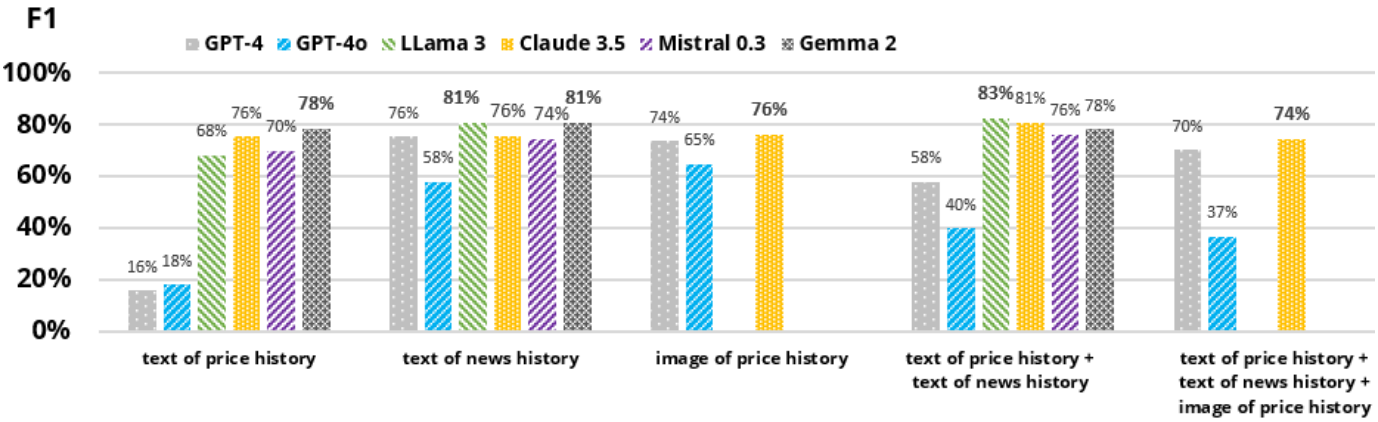

Fig. 4: Prediction of increases and decreases (F1) based on the LLMs' predicted Apple stock prices.

REFERENCES

[1] Y. Cao, Z. Chen, Q. Pei, N. J. Lee, K. Subbalakshmi, and P. M. Ndiaye, "ECC Analyzer: Extract Trading Signal from Earnings Conference Calls using Large Language Model for Stock Volatility Prediction," in *Proceedings of the ACM Conference*. ACM, 2024, p. 9. [Online]. Available: https://arxiv.org/abs/2404.18470

[2] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," 2019. [Online]. Available: https://arxiv.org/abs/1908.10063

[3] Z. Kou, H. Yu, J. Peng, and L. Chen, "Automate Strategy Finding with LLM in Quant Investment," 2024. [Online]. Available: https://arxiv.org/abs/2409.06289

[4] T. Guo and E. Hauptmann, "Fine-Tuning Large Language Models for Stock Return Prediction Using Newsflow," *arXiv preprint*, vol. abs/2407.18103, 2024. [Online]. Available: https://arxiv.org/abs/2407.18103v2

[5] M. Wang, K. Izumi, and H. Sakaji, "LLMFactor: Extracting Profitable Factors through Prompts for Explainable Stock Movement Prediction," *arXiv preprint*, vol. abs/2406.10811, 2024. [Online]. Available: https://arxiv.org/abs/2406.10811

[6] J. Kaur and K. Dharni, "Data Mining-based Stock Price Prediction using Hybridization of Technical and Fundamental Analysis," *Data Technologies and Applications*, 2023. [Online]. Available: https://doi.org/10.1108/DTA-04-2022-0142

[7] C. H. Daube and V. Krivenkov, "Generative AI Tools zur Prognose von Leitzins-Entscheidungen: eine Fallstudie am Beispiel der Leitzinsentscheidungen der Federal Reserve," ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg, IUCF Working Paper 6/2024, 2024. [Online]. Available: https://hdl.handle.net/10419/293992

[8] Y. Deng, X. He, J. Hu, and S.-M. Yiu, "Enhancing Few-Shot Stock Trend Prediction with Large Language Models," *arXiv preprint*, vol. abs/2407.09003, 2024. [Online]. Available: https://arxiv.org/abs/2407.09003

[9] H. Ni, S. Meng, X. Chen, Z. Zhao, A. Chen, P. Li, S. Zhang, Q. Yin, Y. Wang, and Y. Chan, "Harnessing Earnings Reports for Stock Predictions: A QLoRA-Enhanced LLM Approach," *arXiv preprint*, vol. abs/2408.06634, 2024. [Online]. Available: https://arxiv.org/abs/2408.06634v1

[10] Q. Zheng, "How can we Use ChatGPT Better: A Research of API-Enhanced ChatGPT in Stock Prediction," M.A. Thesis, University of Chicago, June 2024. [Online]. Available: https://doi.org/10.6082/uchicago.11888

[11] OpenAI, "GPT-4 Technical Report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[12] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep Learning-Enabled Medical Computer Vision," *npj Digital Medicine*, vol. 4, p. 5, 2021. [Online]. Available: https://doi.org/10.1038/s41746-020-00376-2

[13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, September 2022. [Online]. Available: https://doi.org/10.1145/3505244

[14] Y. Ding, S. Jia, T. Ma, B. Mao, X. Zhou, L. Li, and D. Han, "Integrating Stock Features and Global Information via Large Language Models for Enhanced Stock Return Prediction," 2023. [Online]. Available: https://arxiv.org/abs/2310.05627

[15] E. Zhang, R. Yao, H. Liu, J. Yu, and J. Wang, "First Multi-Dimensional Evaluation of Flowchart Comprehension for Multimodal Large Language Models," 2024. [Online]. Available: https://arxiv.org/abs/2406.10057

[16] M. AI, “The Llama 3 Herd of Models,” 2024. [Online]. Available: https://arxiv.org/abs/2407.21783
[17] Anthropic, “Claude 3.5,” 2024, accessed: 2024-10-10. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet
[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, “Mistral 7B,” 2023. [Online]. Available: https://arxiv.org/abs/2310.06825
[19] Google, “Gemma 2-9B Model,” 2024, accessed: 2024-10-10. [Online]. Available: https://huggingface.co/google/gemma-2-9b-it