Ploutos: Towards interpretable stock movement prediction with financial large language model

Hanshuang Tong, Jun Li, Ning Wu, Ming Gong*, Dongmei Zhang, Qi Zhang
Microsoft, Beijing, China
{hanstong, junli1, wuning, migon, dongmeiz, zhang.qi}@microsoft.com

Abstract

Recent advancements in large language models (LLMs) have opened new pathways for many domains. However, the full potential of LLMs in financial investments remains largely untapped. There are two main challenges for typical deep learning-based methods for quantitative finance. First, they struggle to fuse textual and numerical information flexibly for stock movement prediction. Second, traditional methods lack clarity and interpretability, which impedes their application in scenarios where the justification for predictions is essential. To solve the above challenges, we propose Ploutos, a novel financial LLM framework that consists of PloutosGen and PloutosGPT. The PloutosGen contains multiple primary experts that can analyze different modal data, such as text and numbers, and provide quantitative strategies from different perspectives. Then PloutosGPT combines their insights and predictions and generates interpretable rationales. To generate accurate and faithful rationales, the training strategy of PloutosGPT leverage rearview-mirror prompting mechanism to guide GPT-4 to generate rationales, and a dynamic token weighting mechanism to finetune LLM by detecting and emphasizing key tokens in rationales. Extensive experiments show our framework outperforms the state-of-the-art methods on both prediction accuracy and interpretability.

1 INTRODUCTION

The prospect of predicting stock market trends has continually intrigued both academics and investors for a long time, and it is still challenging due to the volatile nature of financial markets and multimodal features related to stock price (Adam et al., 2016), such as textual features (e.g., news, tweets) and time-series numerical features (e.g. price sequence, volumes sequence) (Kohara et al., 1997; Neal and Wheatley, 1998). In practice, researchers focus more on predicting stock movement (whether the stock price movement will increase or decrease) rather than attempting to predict the precise stock price values, which is widely recognized as a inherently unpredictable task. (Walczak, 2001).

Quantitative methods can be either traditional deep model based or LLM based. Traditional methods use LSTM, transformer, or graph models to combine stock texts and prices (Gao et al., 2023; Sawhney et al., 2020a). However, they are opaque and lack of interpretability. LLM based methods use LLM to label sentiment or directly merge news and prices into model training (Chen et al., 2023; Xie et al., 2023b). They also struggle to generate interpretable decision making rationales and fail to adaptively consider different strategies for different markets. To understand and fuse multi-modal features such as stock price and text data in a flexible way, we introduce PloutosGen pipeline that consists of multiple experts to model stock movement from different aspects.

1. *Technical Expert*, extracts technical features from time series based on various alpha formulas. Then it feeds these features into a time series forecasting model as next-token prediction.

2. *Sentiment Expert*, captures the correlation between news events and their effect on stock valuations, which is well-established but heterogeneous and dependent on the content value of the news, ranging from breaking news to trivial commentary. This expert aggregates this diverse news with sophisticated methods.

3. *Human Expert*, incorporates human thoughts into the framework, which can enhance its effectiveness in real applications.

However, the predictive accuracy of strategy experts can vary considerably across different stock

*Corresponding author

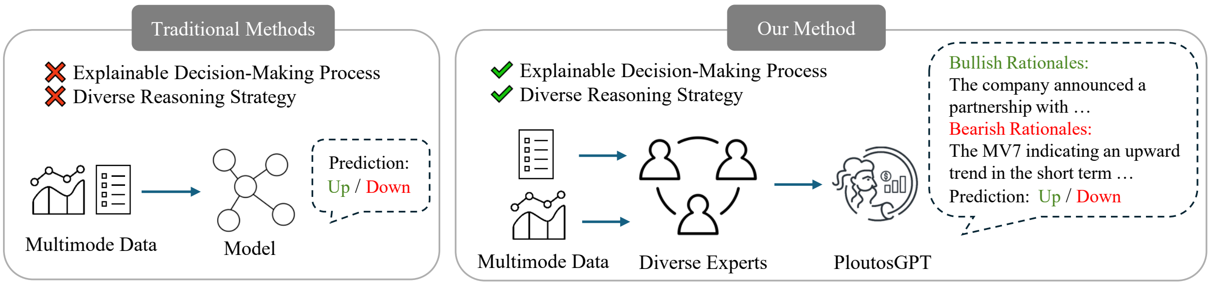

Traditional Methods
✗ Explainable Decision-Making Process
✗ Diverse Reasoning Strategy
Multimode Data
Model
Prediction:
Up / Down
Our Method
✓ Explainable Decision-Making Process
✓ Diverse Reasoning Strategy
Multimode Data
Diverse Experts
PloutosGPT
Bullish Rationales:
The company announced a partnership with …
Bearish Rationales:
The MV7 indicating an upward trend in the short term …
Prediction: Up / Down

Figure 1: Comparison of Quantitative Methods

situations. Therefore, we need an adaptive framework that can effectively incorporate the insights of the most suitable experts for any given stock situation. As shown in Fig. 1, we leverage the interpretative and analytical capabilities of large language models to process multimodal data from diverse strategy experts, and enable them to adaptively learn and evolve their strategies to accommodate varying market conditions, and generate convincing decision making rationales.

To cooperate with the PloutosGen pipeline, we further propose PloutosGPT. It consists of a rearview-mirror prompting strategy for data construction and a dynamic token weighting mechanism for supervised finetuning. Inspired by the common economic phenomenon that *the rearview mirror is always clearer than the windshield* (Buffett and Hathaway, 1991), which indicates that analyzing the rationales of past cases is always easier than predicting the future, we prompt GPT4 to find suitable experts and use their strategies to account for the price variations each time there is a movement in stock prices. By this prompting strategy, we can collect high quality rationales for supervised finetuning of PloutosGPT. However, even with accurate and grounded rationales, we still face low accuracy in the key tokens generation of PloutosGPT. Hence, we propose a dynamic token weighting mechanism to detect and emphasize key tokens in supervised finetuning. It measures token importance by calculating cosine similarity between each token's hidden state and the type embedding produced by a verbalizer.

Our experiments demonstrate the effectiveness of our framework, which outperforms the state-of-the-art method on both prediction accuracy and interpretablity. To summarize, our contributions are as follow:

- We propose PloutosGen, a novel financial LLM pipeline that analyzes and integrates multimodal data, particularly numeric time series data derived from various alpha formulas.
- We propose PloutosGPT, an advanced financial LLM that uses rearview-mirror prompting for interpretable stock prediction and dynamic token weighting for supervised finetuning.
- We introduce faithfulness and informativeness to evaluate the quality of generated rationales. Extensive experiments show the effectiveness and interpretability of our methodology.

2 PROBLEM FORMULATION

The goal of quantitative investment task is to learn stock price related information from both social media, specifically tweets, and past stock performance indicators to predict stock trends. The direction of a stock's movement is determined by comparing its adjusted closing prices over two consecutive trading days. This analysis is structured as a binary classification task, where the prediction revolves around whether the price of a particular stock will move up or down.

2.1 Stock Prediction

For a given stock $s \in \mathcal{S}$, a target prediction date d, and related historical prices and news data, we define stock movement prediction label over a period of $\mathcal{T}$ days as:

$$Y_d = \begin{cases} 0, & p_d^c < p_{d-1}^c \\ 1, & p_d^c \geq p_{d-1}^c \end{cases}$$

Here the p_d^c represents the adjusted closing price in the target prediction date d (Xu and Cohen, 2018; Sawhney et al., 2020a). The label Y_d is categorized as 0 if there's a decrease in the adjusted closing price, and as 1 indicate a rise in the adjusted closing price.

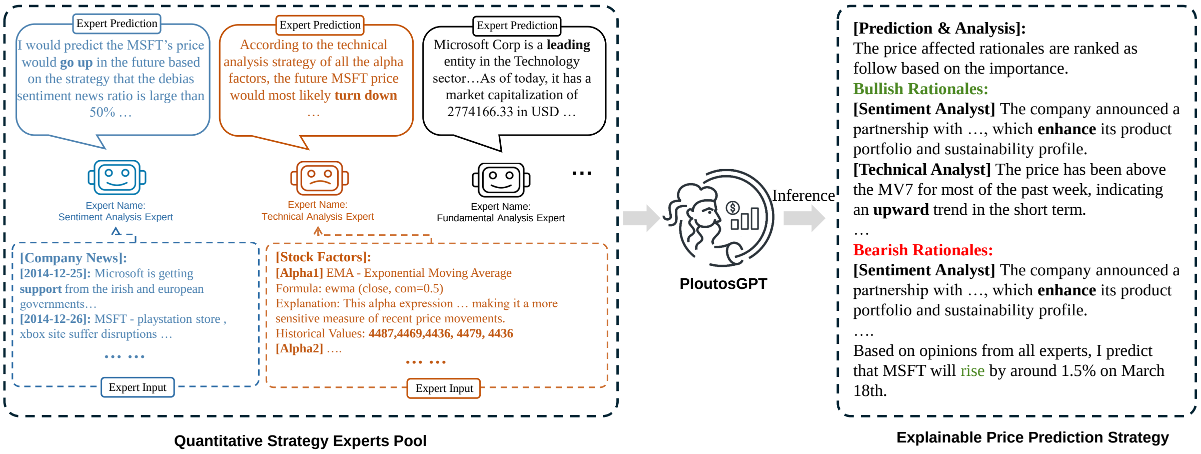

Expert Prediction
I would predict the MSFT's price would **go up** in the future based on the strategy that the debias sentiment news ratio is large than 50% …
Expert Prediction
According to the technical analysis strategy of all the alpha factors, the future MSFT price would most likely **turn down** …
Expert Prediction
Microsoft Corp is a **leading** entity in the Technology sector…As of today, it has a market capitalization of 2774166.33 in USD …
Expert Name: Sentiment Analysis Expert
Expert Name: Technical Analysis Expert
Expert Name: Fundamental Analysis Expert
…
[Company News]:
[2014-12-25]: Microsoft is getting **support** from the irish and european governments…
[2014-12-26]: MSFT - playstation store , xbox site suffer disruptions …
… …
Expert Input
[Stock Factors]:
[Alpha1] EMA - Exponential Moving Average
Formula: ewma (close, com=0.5)
Explanation: This alpha expression … making it a more sensitive measure of recent price movements.
Historical Values: **4487,4469,4436, 4479, 4436**
[Alpha2] ….
… …
Expert Input
Quantitative Strategy Experts Pool
Inference
PloutosGPT
[Prediction & Analysis]:
The price affected rationales are ranked as follow based on the importance.
Bullish Rationales:
[Sentiment Analyst] The company announced a partnership with …, which **enhance** its product portfolio and sustainability profile.
[Technical Analyst] The price has been above the MV7 for most of the past week, indicating an **upward** trend in the short term.
…
Bearish Rationales:
[Sentiment Analyst] The company announced a partnership with …, which **enhance** its product portfolio and sustainability profile.
….
Based on opinions from all experts, I predict that MSFT will rise by around 1.5% on March 18th.
Explainable Price Prediction Strategy

Figure 2: Ploutos framework: When predicting the price movement of a given stock, the diverse strategy experts pool would first leverage its diverse experts to gather and analysis the stock related feature with his strategy. Then, the PloutosGPT would generate bullish and bearish rationales to make final decision based on the market condition and opinions from diverse expert.

3 METHODOLOGY

In this section, we first give an overview of the framework of Ploutos, followed with a detailed explanation of PloutosGen and PloutosGPT.

3.1 PloutosGen Pipeline

As shown in Fig. 2, the core component of PloutsGen pipeline is the diverse expert pool, which is designed to provide signals from a wide range of experts, including sentiment, technical analysis and human analysis, each utilizing a Large Language Model (LLM) to analyze the stock related information from a specific perspective. By integrating insights from these experts, the PloutosGPT can leverage a broad spectrum of knowledge and techniques to inform its predictive capabilities.

3.1.1 Sentiment Analysis Expert

In this section, we would introduce how we leverage only the sentiment information to predict the stock price. Previous research shows social media is useful for this task (Si et al., 2013, 2014), but there are two challenges: i) existing large language models (LLMs) are not fine-tuned for stock prediction from sentiment perspective; ii) how to aggregate news or tweets into stock signals, since different events affect stock price differently. We explore three ways based on whether stock signals are used for training: unsupervised, supervised, and supervised&unsupervised sentiment-based stock prediction.

Unsupervised: We use several datasets with different assets and news types, such as FIQA-SA (Maia et al., 2018), NWGI (Chu et al., 2023), BTSA (Ishii, Accessed: 2023), and TNFS (Magic, 2022). These datasets ensure our models are trained on diverse financial information, enhancing robustness and adaptability. We use the sentiment instruction data to fine-tune LLMs (see Table 10 in Appendix for detailed prompt), using LLaMA-2 as our backbone model. We evaluate the performance on ACL18 dataset (see Appendix A.1 for details). We use the trained sentiment model to label each news from the past 5 days ($\mathcal{T} = 5$) of a stock as positive, negative, or neutral. The stock prediction on day d is the average of sentiment scores. A positive score means an increase, while a negative score means a decrease. We combine all datasets to train the sentiment expert.

Supervised: we construct the instruction with the prompt shown in Table 1, which takes the news text from the five days preceding the target day as input, and the stock price on the target day as the prediction label. Similarly, we adopt LLaMA-2 as our base model for training and build our supervised method. We also try to combine the *unsup* dataset to integrate both methods and improve the performance (more experiments are shown in Appendix A.2).

3.1.2 Technical Analysis Expert

We aim to use large language models (LLMs) for multi-features time series forecasting in stock prediction, where we leverage multiple alphas (domain-specific time series features) and their explanations (text features) to frame the task as next-token prediction. However, existing methods are often limited

Table 1: Supervised Sentiment analysis prompt.

Instruction Input
Instruction: Predict the MSFT's price movement for the 2015-06-10 based on the sentiment of the following tweets. Please choose an answer from up or down. Input: From 2015-06-05 to 2015-06-09, company related news during this period are listed below: 2015-06-08: game-changers announced at microsoft 2015-06-09: microsoft lumia sales tops apple in pakistan 2015-06-09: msft - are the days of subsidized 200 smartphones nearing an end? Answer:
Instruction Output
The stock price of MSFT will go up.

by their number of time series features and deficiency in model explainability. To address these challenges, we propose Number-to-Text Alignment (N2I-Align), which uses prompt to integrate explaining texts with multi-features time series data. Table 2 illustrates our idea of using LLMs as a mediator between alphas and texts. We first generate multi-features time series datasets for a given stock, a target prediction date, and a fixed window size, using some common alphas (see Table 9 in Appendix for details). Then, we rescale and tokenize the time series data into sequences. To avoid the fragmentation and token capacity issues of conventional tokenization methods, we use the LLaMA-2 model, which naturally tokenizes long numbers into multiple digits. Finally, we use the N2I-Align prompt to wrap the tokenized sequences and generate the training data for technical expert tuning.

3.1.3 Human Experts

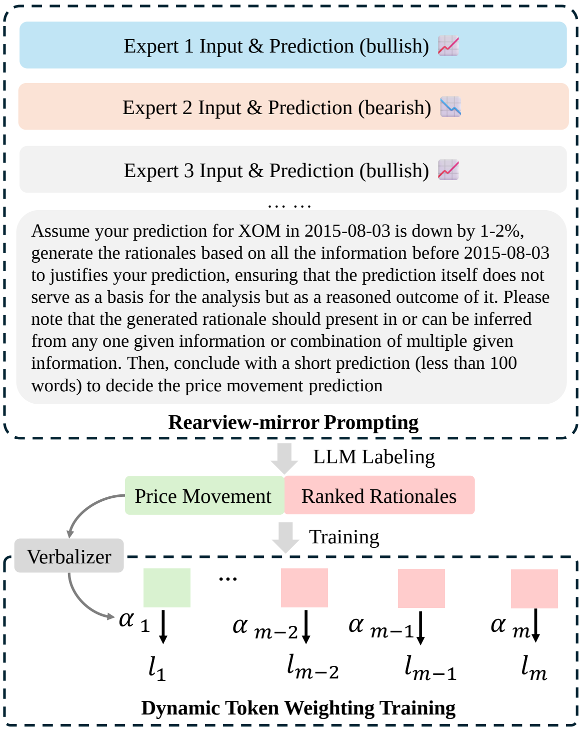
The human financial expert encapsulates the wisdom and intuition of seasoned investment professionals, including insights from fundamental analysis, macroeconomic indicators, and market cycles. The key feature of this expert is its ability to contextualize financial data within a broader economic and financial framework, providing an additional layer of natural language analysis that is often difficult to quantify but critical for making informed investment decisions. This expert serve as a customizable and optional component in PromptGen framework since it's not available in open-source dataeset but useful in real applications.

Table 2: Technical analysis prompt.

Instruction Input
Instruction: Predict the MSFT's price movement for the 2015-06-10 based on the following tweets. **[Stock Factors]**: **[Alpha1]** MV7 - Moving Average of 7 Days Formula: ts_mean(close, 7) Explanation: This alpha expression ... which helps to smooth out short-term volatility and identify the underlying trend in the price movement. Historical Values: 53, 67, 68, 84, 89 **[Alpha2]** MV20 - Moving Average of 20 Days . . . Historical Values: 78,83, 92,87, 95 . . . **[Analysis]**: The historical price for MSFT from 2014-05-25 to 2014-05-28 is 63, 74, 78, 86. Please predict the MSFT's price for the 2014-05-29 based on all above information Answer:
Instruction Output
The stock price of MSFT will be 93.

3.2 PloutosGPT Training

The main objective of PloutosGPT is to develop a Large Language Model (LLM) capable of generating explainable and accurate rationales for predicting price movements. Explainable AI in finance research has attracted increasing attention from various perspectives (Weber et al., 2023). For instance, some approaches aim to build interpretable stock price predictors using neuro-fuzzy rules (Rajab and Sharma, 2019; Vlasenko et al., 2019). Similarly, Yang et al. (Yang et al., 2018) leverage attention-based neural networks to assign importance to different news. However, the interpretability of these existing methods is largely compromised by the neural network learning phase. To the best of our knowledge, PloutosGPT is the first method that can generate natural language rationales for interpretable stock movement prediction. As shown in Figure 3, it consists of two steps: rearview-mirror based data construction and dynamic token weighting training strategy. In the first step, as shown in the figure, we prompt a powerful frozen LLM (e.g. GPT-4) to summarize both the bullish and bearish rationales for the price movement based on diverse experts' input, and generate rationales that is faithful to the given information (Since the ground truth is given, the frozen LLM knows which rationale makes the correct prediction). Then, the model


Expert 1 Input & Prediction (bullish) 📈
Expert 2 Input & Prediction (bearish) 📉
Expert 3 Input & Prediction (bullish) 📈
… …

Assume your prediction for XOM in 2015-08-03 is down by 1-2%, generate the rationales based on all the information before 2015-08-03 to justifies your prediction, ensuring that the prediction itself does not serve as a basis for the analysis but as a reasoned outcome of it. Please note that the generated rationale should present in or can be inferred from any one given information or combination of multiple given information. Then, conclude with a short prediction (less than 100 words) to decide the price movement prediction

Figure 3: PloutosGPT is trained by dynamic token weighting and rearview-mirror data.

is prompted to make the final decision based on the trade-off between bullish and bearish rationales. The intuition behind these prompts is to generate faithful rationales by learning weights for each expert's input based on their accuracy and relevance in different market conditions. All of these prompts are under the constraint that the model should not directly use the ground truth as the rationales, which echoes our initial motivation in introduction that researchers are better at analyzing the past given the result than predicting the future.

To further finetune PloutosGPT, the generated rationales and stock movement label are combined as the output of the instruction training data, and the input and analysis from diverse experts are combined as the input. An example of the constructed instruction data is shown in the Table 11 of Appendix. By leveraging this curated training data, PloutosGPT becomes not just a black-box predictor but a sophisticated decision support tool that integrates the strengths of diverse strategies and provides actionable insights with a clear justification for its predictions. Moreover, though we have constructed bullish and bearish rationales based on ground truth and diverse experts' output, the fine-tuned PloutosGPT still struggles to generate key tokens of rationales. For example, in *the MV7 indicating an upward trend in short term*, the key word *upward* has to be treated carefully. Hence, we propose to dynamically control the weight of each token in the rationales of the training data as follows:

$$\mathcal{L} = \sum_i -\alpha_i \log p(y_i|x, y_{1...i}) \tag{1}$$

where x is input of model, and y is generated content consists of rationales and label. α_i is the weight of i-th token in generated content. It can be obtained by computing the similarity between a specific type embeddings and token embeddings:

$$\alpha_i = \frac{\exp(\cos(\boldsymbol{t}_{Y_d}, \boldsymbol{h}_i)/\tau)}{\sum_{j=1}^{m} \exp(\cos(\boldsymbol{t}_{Y_d}, \boldsymbol{h}_j)/\tau)} \tag{2}$$

where $\boldsymbol{h}_i$ denotes hidden state of i-th token from transformer's last layers, $\boldsymbol{t}_{Y_d}$ is the type emebdding which is computed by average pooling of verbalizer tokens embedding and τ is temperature which controls the weight of key tokens. A suitable temperature could give more importance to key tokens, and hence bring higher prediction accuracy and high quality rationales. The verbalizer tokens are generated by verbalizer based on label Y_d. More details about the verbalizer can be checked in the Appendix C.

4 EXPERIMENTS

Our experiments are designed to investigate the following research questions:

- **RQ1**: How does Ploutos perform compared with current LLMbased and traditional prediction models?
- **RQ2**: How do the different components in Ploutos affect its effectiveness?
- **RQ3**: Does the decision making rationales informative and consistent with given information?

4.1 Experiment Settings

4.1.1 Data

In this study, we evaluate the proposed method on two public and classic datasets on stock movement prediction, ACL18 (Xu and Cohen, 2018) and CIKM18 (Wu et al., 2018). ACL18 comprises 88 high-volume trading stocks from the S&P 500 index listed on the NYSE and NASDAQ exchanges. Each of stock contains price data obtained from

Yahoo Finance and related tweets data extracted using regular expression queries based on NASDAQ ticker symbols for each day (except for weekends and public holidays). Then a $\mathcal{T}$-day ($\mathcal{T}$=5) rolling window is utilized to produce data candidates along the series of trading days. Samples are labeled as either positive or negative based on the percentage change in closing price, with increases of 0.55% or more labeled as positive and decreases of 0.5% or more labeled as negative, in order to generate an equitably partitioned dataset of 26,614 samples. Following Xu and Chen (Xu and Cohen, 2018), we split the dataset into distinct sets for training (01/01/2014 to 31/07/2015), validation (01/08/2015 to 30/09/2015), and testing (01/10/2015 to 01/01/2016) in a 7:1:2 ratio. CIKM18 consists of 47 stocks with sufficient tweets and price data ranging from January 2017 to November 2017. Similarly, we leverage the same data processing method as ACL18 to process CIKM18 dataset, and we follow the same approach as stated in the original paper to split the datasets.

4.1.2 Evaluation Metric

Since Ploutos aims to predict price movement for a given target day, *i.e.*, a binary classification problem. We adopt accuracy, F1 score, and the Matthews Correlation Coefficient (MCC, sourced from sklearn's implementations[1]) for classification performance. We employ MCC due to its robustness and unbias evaluation of classifier performance regardless of data imbalance as it considers all quadrants of the confusion matrix.

4.1.3 Baselines

We compare Ploutos with the below baselines containing both traditional model based and LLM based methods. Those baselines are selected based on the criteria that it's either open-source or reproducible.

Traditional Methods: These methods uses traditional deep neural network such as time-series, graph based and others. Including ARIMA (Brown, 2004), Adv-LSTM (Selvin et al., 2017), StockNet (Xu and Cohen, 2018), DTML (Nguyen and Shirai, 2015).

LLM Based Methods: These methods leverage either zero shot or finetuned version of large language model to predict the stock movement. We select LLaMA-2 (Touvron et al., 2023), GPT-4 (Achiam et al., 2023), FinMA (Xie et al., 2023b) as baselines.

Table 3: A comparison of prediction accuracy and MCC between Ploutos and other baselines.

Models	ACL18		CIKM18	
	Acc.	MCC	Acc.	MCC
ARIMA	51.42	-0.021	52.36	-0.012
Adv-LSTM	57.24	0.148	56.48	0.016
StockNet	58.23	0.081	56.37	0.023
DTML	57.44	0.191	58.62	0.045
GPT-4	53.08	0.023	57.44	0.034
LLaMA-2 - 7B	52.74	0.051	56.92	0.027
FinMA - 7B	56.28	0.104	53.24	-0.031
Ploutos - 7B	**61.21**	**0.205**	**59.89**	**0.064**

4.1.4 Training settings

We conduct all experiments on 8 Tesla V100 GPU. To get the best performance of Ploutos, we apply grid search for batch size in {64, 128, 256} and learning rate in {1e-6, 2e-06, 3e-6} for each expert's training. Each experiments run three times and we report the average metric. Adam is used as the optimizer and we leverage our framework to fine-tune from LLaMA-2 for 2 epochs on all datasets. During the rearview-mirror data generation stage, we execute the scraping process using Azure OpenAI GPT-4 API. We use a temperature of 1.0 to generate samples and set the maximum number of tokens for generation to 2000. Moreover, we set the frequency penalty to zero and top-p to 1.0. For evaluation, we set the temperature to zero for all models to reduce output randomness and ensure more focused and deterministic outputs.

4.2 Performance Comparison (RQ1)

We compare the stock movement prediction performance of various methods including both traditional and LLM based methods. The evaluation results against baseline methods are presented in Table 3. From the figure and table, we draw the following conclusions: 1) Our method significantly outperforms both traditional and LLM based methods, verifying the superiority of tuning LLMs via our proposed training framework. Ploutos successfully unlocks the text and numerical data understanding capabilities of LLMs for stock movement prediction task. 2) Traditional deep models such as ARIMA, Adv-LSTM and others perform worse than Ploutos, implying the potential of leveraging language model to process and understand multi-mode stock related data. 3) The zero shot ability of LLaMA-2 and GPT-4 perform similarly to random guessing (ACC$\simeq$ 0.5). However, our Ploutos achieves signif-

[1] https://scikit-learn.org

Table 4: Performance of Ploutos variants on ACL18 dataset, showing every component makes an important contribution to the excellent performance.

Model	Accuracy	MCC	F1
$\text{Ploutos}_{\neg S}$	59.62	0.189	0.598
$\text{Ploutos}_{\neg T}$	58.11	0.153	0.537
$\text{Ploutos}_{\neg R}$	60.41	0.191	0.604
Ploutos	**61.21**	**0.205**	**0.612**

icant improvements. These results demonstrate a considerable gap between stock movement prediction and language tasks, showing the importance of using stock related data for tuning on LLMs. 4) Financial large language model FinMA-7B trained with StockNet training set yield inferior performance than Ploutos, indicating purely tuning LLM with stock related instruction data cannot get optimal performance for stock movement prediction task.

4.3 Ablation Study (RQ2)

To evaluate the contribution of different components, we compare against several variants:

- $\textbf{Ploutos}_{\neg \textbf{S}}$: Ploutos without the sentiment analysis expert, which makes decisions just based on the numerical technical analysis.
- $\textbf{Ploutos}_{\neg \textbf{T}}$: Ploutos without the technical analysis expert, which makes decisions just based on the news sentiment analysis.
- $\textbf{Ploutos}_{\neg \textbf{R}}$: Ploutos without the rearview-mirror instruction data and dynamic token weighting mechanism.

The results are shown in Table 4. $\text{Ploutos}_{\neg S}$ achieves higher prediction accuracy than technical feature only methods such as ALSTM and Adv-LSTM, indicating the importance of the number to text alignment technique. Similar to the performance of Stocknet, $\text{Ploutos}_{\neg T}$ has a relatively high Acc. but lower MCC, suggesting texts are particularly helpful to predict on one side of the binary classification. By modeling both text data and stock relationships, $\text{Ploutos}_{\neg R}$ reaches a relatively good result. Finally, better performance achieved when rearview-mirror instruction data and dynamic token weighting are introduced into training. In summary, the ablation studies conclusively demonstrate that each component within the Ploutos architecture is important, and as a result the full model with all components achieves the best performance.

Table 5: Faithfulness and Informativeness of different methods on ACL18 dataset

Model	Size	Faithfulness	Informativeness
FinMA	7B	59.53	61.32
FinGPT	7B	72.82	84.76
GPT-3	175B	77.63	91.58
Ploutos	7B	**81.24**	**96.52**

4.4 Faithfulness and Informativeness of Interpretablity (RQ3)

To verify the quality of the interpretablity of the genareted rationales, we employ two quantifiable metrics: faithfulness and informativeness. Faithfulness measures whether the facts in model's response are based on or can be inferred from the given knowledge. Following the method in HaluEval (Li et al., 2023), we design prompt to call GPT-4 to classify each fact within the model's response rationales as either inferable from the given knowledge or not, and then aggregate these accuracies as the model's faithfulness score (see Appendix D.1 for more information). Informativeness measures the amount of information contained in the model's response, which is equally an important metric for interpretablity, as we found some models might output generally correct but uninformative rationales such as "*predict rise in price due to the promising information.*". To measure the informativeness of each rationale, we directly use the informativeness measurement from TruthfulQA (Lin et al., 2021).

Table 5 shows the interpretablity score of different models, from which we can draw several conclusions: 1) Compared to other LLM methods, Ploutos demonstrates superior faithfulness and informativeness, which indicates that large language model can be trained through specific methods to generate informative and faithful rationales, thereby enhancing model's interpretability. 2) The FinMA-7B, which adopt similar training datasets compared to Ploutos, struggled to generate informative rationales, which indicate the effectiveness of the rearview-mirror data training strategy. 3) FinGPT (Yang et al., 2023) outperforms FinMA in terms of both faithfulness and informativeness, which can be attributed to its chain of thought scraping and training technique to enhance reasoning capabilities. However, compared to Ploutos, it does not take technical indicators such as the interpretability of various alphas into account, leading to a lack of technical analytical capabilities.

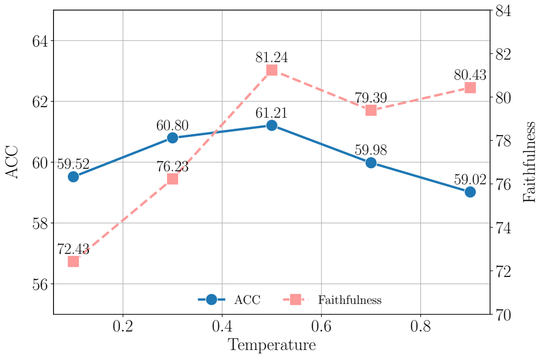


Figure 4: ACC and faithfulness over different temperatures on ACL18 Dataset

Moreover, we explored the impact of temperature of dynamic token weight on the model's predictive accuracy and interpretability. The results are summarized in Fig. 4. It can be observed that when the temperature is greater than 0.5, a lower temperature corresponds with higher accuracy, aligning with our intuition that a lower temperature increases the weight of the loss focus on predicting the stock movement. However, when the temperature is less than 0.5, both accuracy and faithfulness begin to decrease, indicating that overly focusing on the movement itself and neglecting the interpretable rationale tokens can actually weaken the model's reasoning capabilities. When temperature equals to 0.5, both faithfulness and accuracy reach their maximum, suggesting a synergistic rather than a mutually exclusive relationship between predictive accuracy and interpretability, which also demonstrates the effectiveness of dynamic token weighting.

5 Related Work

5.1 Stock Prediction with Deep Learning

Deep learning approaches have gained prominence for their capacity to model complex patterns in stock market data. It can be divided into several tracks: RNN Based, CNN Based, Transformer Based and Graph Based Models. Recurrent neural networks (RNNs) are favored for their proficiency in capturing sequential trends in stocks (Li et al., 2018; Qin et al., 2017), with advancements like Long Short-Term Memory (LSTM) networks enhanced by adversarial training demonstrating superior performance in stock prediction (Feng et al., 2018). Moreover, CNN Based Models incorporating stock times series data as a set of input feature maps (Hoseinzade and Haratizadeh, 2019; Lu et al., 2021). Furthermore, the emergence of transformer-based models, exemplified by StockFormers (Gao et al., 2023), showcases the effectiveness to extract latent representations of long short term stock states and asset relations. Recent efforts have also explored graph-based deep learning models or casual matrix to analyze the interconnections among stocks, outperforming price-only models (Luo et al., 2023; Sawhney et al., 2020b; Kim et al., 2019).

5.2 Stock Prediction with Large Language Model

The recent strides in large language model (LLM) have catalyzed the exploration of stock prediction application with LLM. Existing studies consists of several directions. Some studies focus solely on textual information. Lopez-Lira et al. (Lopez-Lira and Tang, 2023) demonstrate the correlation bewteen ChatGPT sentiment score and subsequent stock movement. Some researchers focus more on the numerical feature, Alpha-GPT is proposed as a new alpha mining paradigm by leveraging the power of large language models (Wang et al., 2023). Others propose methods combine numerical and textual features, Xie et al. (Xie et al., 2023b,a) directly prompt and fine-tune LLM to leverage both news and numerical features. Futhormore, Chen et al. (Chen et al., 2023) explore the relationship between stock by analyzing the textual information by ChatGPT. However, these techniques fall short of exploring the interpretablity ability of LLM, thereby limiting their application potential.

6 CONCLUSION

In this work, we explored the feasibility of using LLMs to construct a framework that can generate interpretable decision making rationals for stock movement prediction. Specifically, we first explore different strategies to analyze the multi-model stock related data for stock movement prediction from different perspectives. Then we propose Ploutos-GPT, which introduces an novel training approach called rearview-mirror prompting and dynamic token weightin that are designed to dynamically incorporate insights from diverse quantitative strategy experts, encompassing both sentiment and technical analysis. Extensive experiments shows that our method can perform well under different conditions, reinforcing its practical applicability.

Limitations

While our proposed Ploutos framework shows promising results in stock movement prediction, there are several limitations that need to be addressed:

- The PloutosGPT is designed to combine insights and predictions from multiple primary experts. However, the selection and performance of these experts can significantly impact the model's performance. The process of expert selection and their potential biases need to be thoroughly examined.
- The computational cost of our model, especially when dealing with large-scale data, can be high. Optimizing the model for efficiency without compromising its predictive accuracy is a challenge.
- Our model currently focuses on textual and numerical data. Incorporating other types of data, such as visual data, could potentially improve the model's performance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Klaus Adam, Albert Marcet, and Juan Pablo Nicolini. 2016. Stock market volatility and learning. *The Journal of finance*, 71(1):33–82.

Robert Goodell Brown. 2004. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

Warren Buffett and Berkshire Hathaway. 1991. Letter to shareholders (1991). https://www.berkshirehathaway.com/letters/1991.html.

Zihan Chen, Lei Nico Zheng, Cheng Lu, Jialu Yuan, and Di Zhu. 2023. Chatgpt informed graph neural network for stock movement prediction. *arXiv preprint arXiv:2306.03763*.

Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, Hong Chen, Wanqing Xu, Xin Lu, Qing Cui, Longfei Li, et al. 2023. Data-centric financial large language models. *arXiv preprint arXiv:2310.17784*.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2018. Enhancing stock movement prediction with adversarial training. *arXiv preprint arXiv:1810.09936*.

Siyu Gao, Yunbo Wang, and Xiaokang Yang. 2023. Stockformer: learning hybrid trading machines with predictive coding. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4766–4774.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820*.

Ehsan Hoseinzade and Saman Haratizadeh. 2019. Cnnpred: Cnn-based stock market prediction using a diverse set of variables. *Expert Systems with Applications*, 129:273–285.

Stpete Ishii. Accessed: 2023. Bitcoin tweets sentiment analysis. https://www.kaggle.com/code/stpeteishii/bitcoin-tweets-sentiment-analysis/. Kaggle code repository.

Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. Hats: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999*.

Kazuhiro Kohara, Tsutomu Ishikawa, Yoshimi Fukuhara, and Yukihiro Nakamura. 1997. Stock price prediction using prior knowledge and neural networks. *Intelligent Systems in Accounting, Finance & Management*, 6(1):11–22.

Hao Li, Yanyan Shen, and Yanmin Zhu. 2018. Stock price prediction using attention-based multi-input lstm. In *Asian conference on machine learning*, pages 454–469. PMLR.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Wenjie Lu, Jiazheng Li, Jingyang Wang, and Lele Qin. 2021. A cnn-bilstm-am method for stock price prediction. *Neural Computing and Applications*, 33:4741–4753.

Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. 2023. Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12164–12176.

Neural Magic. 2022. Twitter financial news sentiment. https://huggingface.co/datasets/zeroshot/twitter-financialnews-sentiment. Accessed: [Insert access date here].

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Robert Neal and Simon M Wheatley. 1998. Do measures of investor sentiment predict returns? *Journal of Financial and Quantitative Analysis*, 33(4):523–547.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364.

Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*.

Sharifa Rajab and Vinod Sharma. 2019. An interpretable neuro-fuzzy approach to stock price forecasting. *Soft Computing*, 23:921–936.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Shah. 2020a. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020b. Spatiotemporal hypergraph convolution network for stock movement forecasting. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 482–491. IEEE.

Sreelekshmy Selvin, R Vinayakumar, EA Gopalakrishnan, Vijay Krishna Menon, and KP Soman. 2017. Stock price prediction using lstm, rnn and cnn-sliding window model. In *2017 international conference on advances in computing, communications and informatics (icacci)*, pages 1643–1647. IEEE.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Sinno Jialin Pan, Qing Li, and Huayi Li. 2014. Exploiting social relations and sentiment for stock prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1139–1145.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alexander Vlasenko, Nataliia Vlasenko, Olena Vynokurova, Yevgeniy Bodyanskiy, and Dmytro Peleshko. 2019. A novel ensemble neuro-fuzzy model for financial time series forecasting. *Data*, 4(3):126.

Steven Walczak. 2001. An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of management information systems*, 17(4):203–222.

Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023. Alpha-gpt: Human-ai interactive alpha mining for quantitative investment. *arXiv preprint arXiv:2308.00016*.

Patrick Weber, K Valerie Carl, and Oliver Hinz. 2023. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, pages 1–41.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1627–1630.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023a. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023b. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Linyi Yang, Zheng Zhang, Su Xiong, Lirui Wei, James Ng, Lina Xu, and Ruihai Dong. 2018. Explainable text-driven neural network for stock prediction. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 441–445. IEEE.

A Sentiment Experts

A.1 Choice of unsupervised dataset

Table 6 shows the comparison of the impact for stock movement prediction task over different datasets. Among all the options, the *unsupervised* method combining all the datasets performs the best, which is reasonable since this method consists of most diversified and robustness data for sentiment analysis. Besides, the accuracy of stock market predictions using sentiment scores calculated by this unsupervised method hovers around 50%, which is not particularly high. This indirectly suggests that while the sentiment of these news articles does have a positive correlation with stock prices, there is also a lot of noise within the news itself. We need to develop better methods to aggregate this news.

Table 6: Comparison of the impact for stock movement prediction task over different datasets. The definition of MCC are introduced in section A. The Unsup. method denote that the combined dataset including both FIQA-SA, NWGI, BTSA, TNFS datasets.

Dataset	Accuracy	MCC
FIQA-SA (Maia et al., 2018)	52.92	0.032
NWGI (Chu et al., 2023)	52.48	0.014
BTSA (Ishii, Accessed: 2023)	52.56	0.043
TNFS (Magic, 2022)	51.12	-0.021
Unsup.	**53.24**	**0.052**

A.2 Comparison of different training methods

Table 7 shows the prediction results on StockNet for different methods. The results indicate that the combined 'sup & unsup' method performs the best, which also demonstrates the effectiveness of both our supervised and unsupervised approaches.

Table 7: Comparison of the impact for stock movement prediction task over different sentiment analysis methods.

Method	Accuracy	MCC
Unsupervised	53.24	0.054
Supervised	57.36	0.128
Sup.& Unsup.	**58.11**	**0.153**

B Technical Experts

To evaluate the performance of the number-to-text-alignment (prompt is shown in Table 9), we carefully choose two popular time series baselines: 1) LLMTime (Gruver et al., 2023): a single feature time series prediction method. Here we combine simply combine cross feature time series prediction sequences into one time series sequence; 2) LLMDate: following prompt in (Xie et al., 2023a) which split time series features by date, we adopt similar prompt showing in appendix. We use the same base model and hyperparamters to evaluate the performance of the above prompts. As shown in Table 8, the performance of LLMTime is close to random, which indicates that the LLM is not good at cross feature time series prediction without meticulous task conversion. Besides, we find that N2I-Align performs the best, showing the effectiveness of number to text alignment.

Table 8: Comparison of the impact for stock movement prediction task over different technical analysis methods.

Method	Accuracy	MCC
Rand	50.68	0.002
LLMTime	51.12	0.028
LLMDate	53.47	0.082
N2I-Align	**59.62**	**0.189**

C Details of Verbalizer

We define a verbalizer as an injective function that maps each label to a word list from vocabulary of large language model. For *Up* label, we map it to a word list: [*up*, *boost*, *positive*, *rise*]. For *Down* label, we map it to another word list: [*down*, *reduce*, *negative*, *drop*]

D Interpretablity Evaluation

D.1 Faithfulness

Recent work has shown that LLMs are prone to suffer from hallucination generations across various applications, where the generated responses is either in conflict with existing fact or cannot be verified by the available knowledge resources. To evaluate how faithful the model response to the given knowledge, we design prompt with help of GPT-4 to verify the faithfulness of model response (prompt is shown in Table 12). Unlike the conclusion in HaluEval that existing state of art LLM (GPT-3 at that time) can not distinguish between fact and hallucinated content, we find that GPT-4 is now capable of handle this job. To evaluate this statement, we samples 500 examples from ploutos response and invite human labelers to annotate. For each query and model response, human labelers will annotate whether the response is faithful to the given knowledge resources ("Yes" or "No") and list the corresponding reasons. Each response is labeled by three human labelers, and we adopt the max-voting strategy to determine the final faithfulness label. At last, we compare the Pearson correlation between the human labeled result and prediction from our proposed method, and achieved a Pearson correlation score of 0.826. The result indicates that this method can be used to evaluate the faithfulness of model response.

E PloutosGPT Training

E.1 Example for Rearview-mirror Instruction Prompt

The prompt for rearview-mirror data scraping is shown in Table 11.

Table 9: Technical analysis prompt example

[Instruction]:

You are a seasoned stock market analyst expert in predicting future price trends based on historical stock factors.

[Stock Factors]:

From 2015-11-04 to 2015-11-08, some recent basic stock factors are presented below:

Alpha: MV7 - Moving Average of 7 Days

Formula: ts_mean(close, 7)

Explanation: This Formula expression calculates the average closing price over the past 7 days, which helps to smooth out short-term volatility and identify the underlying trend in the price movement.

Historical Values: 7798,7848,7878

Alpha: MV20 - Moving Average of 20 Days

Formula: ts_mean(close, 20)

Explanation: Similar to MV7, this Formula expression computes the average closing price over a longer period of 20 days. This is often used to assess the medium-term market trend and can act as support or resistance levels for the price.

Historical Values: 7622,7644,7668

Alpha: MACD - Moving Average Convergence Divergence

Formula: minus(ewma(close, 12), ewma(close, 26))

Explanation: The MACD is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price. The Formula expression represents the difference between the 12-day exponential moving average (EMA) and the 26-day EMA of the closing prices.

Historical Values: -729,-747,-737

Alpha: EMA - Exponential Moving Average

Formula: ewma(close, com=0.5)

Explanation: This Formula expression calculates the exponential moving average of the closing prices, giving more weight to recent prices. The com parameter controls the degree of weighting decrease, making it a more sensitive measure of recent price movements.

Historical Values: 8006,7931,7887

Alpha: Bollinger Bands - Middle Line

Formula: ts_stddev(close, 20)

Explanation: This Formula expression is the simple moving average of the closing price over the past 20 days. It serves as the middle line in Bollinger Bands and is used to determine the intermediate-term trend.

Historical Values: 202,206,202

Alpha: Bollinger Bands - Upper Band

Formula: plus(ts_mean(close, 20), times(ts_stddev(close, 20), 2))

Explanation: The upper band of the Bollinger Bands is calculated by adding two standard deviations to the 20-day moving average. This band adjusts for volatility and can signal overbought conditions when the price touches or breaches the upper band.

Historical Values: 8027,8058,8073

Alpha: Bollinger Bands - Lower Band

Formula: minus(ts_mean(close, 20), times(ts_stddev(close, 20), 2))

Explanation: The lower band of the Bollinger Bands is calculated by subtracting two standard deviations from the 20-day moving average. It also adjusts for volatility and can signal oversold conditions when the price touches or breaches the lower band.

Historical Values: 7217,7230,7263

Alpha: LogMomentum

Formula: log(minus(close, shift(close, 1)))

Explanation: This Formula expression calculates the natural logarithm of the difference between the current closing price and the previous day's closing price. It provides a measure of the momentum of the security's price by capturing the rate of change on a logarithmic scale.

Historical Values: 437,436,434

Alpha: VMA60 - Volume Moving Average 60

Formula: data['Volume'].rolling(60).mean() / (data['Volume'] + 1e-12)

Explanation: The VMA60 Formula expression calculates the average volume over a 60-day period and divides it by the current volume plus a small constant to prevent division by zero. It compares the current volume to the average volume over a longer period, which can signal changes in trader participation.

Historical Values: 113,131,129

[Analysis]:

The historical price for XOM from 2015-11-05 to 2015-11-08 is 7895,7864. Please predict the XOM's price for the 2015-11-09 based on all above information.

Answer:

Table 10: Unsupervised sentiment analysis based prompt

[Instruction]: What is the sentiment of this news/tweet?
Please choose an answer from {Negative/Neutral/Positive}.
Input: Mr Ashley sold a 43pc stake in the company for more than pounds 900m at the time of the float.
Answer:

Table 11: Rearview-mirror instruction prompt example

[Instruction]:
You are a seasoned stock market analyst. Your task is to list the bullish or bearish rationales for companies based on relevant news and basic financials from the past weeks, then provide an analysis and prediction for the companies' stock price movement for the upcoming week. Your answer format should be as follow:
Bullish Rationales:
- Rationale 1
- Rationale 2
- ...

Bearish Rationales:
- Rationale 1
- Rationale 2
- ...

Prediction:
- Your prediction here in less than 100 words.

[Sentiment Expert Input]:
From 2015-07-29 to 2015-08-02, Company news during this period are listed below:
... (news used in sentiment analysis prompt as shown in Table 1).
[Sentiment Expert Prediction]: Down
[Technical Expert Input]:
From 2015-07-29 to 2015-08-02, Company news during this period are listed below:
... (alphas used in technical analysis prompt as shown in Table 9).
[Technical Expert Prediction]: Up
[Analysis]:
Assume your prediction for XOM in 2015-08-03 is down by 1-2%, generate the rationales based on all the information before 2015-08-03 to justifies your prediction, ensuring that the prediction itself does not serve as a basis for the analysis but as a reasoned outcome of it. Please note that the generated rationale should present in or can be inferred from any one given information or combination of multiple given information. Then, conclude with a short prediction (less than 100 words) to decide the price movement prediction.

Table 12: Prompt for faithfulness measurement

[Instruction]: You will be given a yes/no question regarding if a rationale is either present or can be inferred from the given information. Your task is to answer yes or no and tell me why based on given information given and guidelines below.
Answer yes if:
The rationale is present in any one given information or combination of multiple given information.
The rationale can be inferred from any one given information or combination of given information.
The rationale is more general than the rationales in the given information, and it can be obtained by simplifying the rationales in the given information.
If rationale in question is supported by at least one given information or follows any of the above or a combination of above conditions answer will be yes, even if other given information contradict it.
Answer no if:
The rationale is not present in any of the given information individually or combined, or it cannot be clearly inferred from any of the given information individually or combined.
The rationale is more specific than the rationales in any of the given information.
The rationale is contradictory to the rationales in all the given information i.e. there is no given information which supports the rationale.
If rationale in question contains some topic which are neither present nor inferred from any of the given information, then answer no to the question.
[Knowledge]: ...
[Rationale]: ...
Answer: