

PREDICTION ALGORITHM OF NORTHBOUND CAPITAL AND STOCK FACTOR BASED ON LARGE LANGUAGE MODELS

JIANMING LI¹, TIANWEN FENG¹ AND XIANGPEI HU²

¹School of Computer Science

²School of Management and Economics

Dalian University of Technology

No. 2, Linggong Road, Ganjingzi District, Dalian 116023, P. R. China

{lijm; fengtw; drhxp}@dlut.edu.cn

Received December 2025; revised March 2026

ABSTRACT. *This study addresses the insufficient domain adaptability of Time-LLM (Time-series Large Language Model) in financial time-series forecasting by proposing a stock prediction algorithm that integrates cognitive enhancement and dynamic calibration. Targeting the characteristics of financial time-series data, high sensitivity, overlapping periodicities, and strong noise, we first reconstruct the prompting using Chain-of-Thought techniques to decompose complex prediction tasks into interpretable reasoning steps, effectively guiding hierarchical decision-making. Subsequently, a dynamic calibration mechanism is designed, incorporating a sliding-window-based “Deep Breath” mechanism error correction module to dynamically suppress error propagation in long-horizon forecasting. Finally, a multimodal fusion architecture is developed, employing a dual-channel framework to model Hong Kong Stock Connect capital flows and market technical indicators separately, with GPT-4 synthesizing multi-source information and generating decision explanations. Experimental results demonstrate significant improvements in prediction accuracy and interpretability, offering a novel technical pathway for financial time-series forecasting.*

Keywords: Large language model, Time series, Hong Kong capital, Technical data, Market prediction

1. Introduction. As a pivotal offshore investor in the A-share market, Hong Kong-based capital demonstrates acute sensitivity to market dynamics, consistently injecting liquidity through the Stock Connect programs while exerting significant pricing influence in core sectors such as consumer goods and pharmaceuticals. Leveraging Hong Kong’s mature regulatory framework and institutional investor-dominated positioning structure, it has evolved into a forward-looking “smart money” entity by integrating global macro strategies with quantitative models. The transparent daily disclosures by HKEX further provide a robust foundation for high-frequency research. Current academic methodologies predominantly focus on three dimensions: short-term capital flow predictions [1], sector rotation predictions driven by positioning shifts [2], and machine learning-enhanced integrated frameworks [3]. Notably, as the concentration of Hong Kong-based capital intensifies, markets must vigilantly monitor its potential spillover effects on specific sectors [4].

Technologically, the Transformer architecture proposed by Vaswani et al. in 2017 revolutionized natural language processing. Its core innovations, self-attention mechanisms eliminating sequence length constraints, parallelizable training for enhanced efficiency, multi-head attention for multi-dimensional information extraction, and explicit positional

encoding for temporal sequencing, have demonstrated exceptional performance in noise-prone, multi-periodic, and highly sensitive financial time-series forecasting tasks while excelling in long-context modeling [5]. This breakthrough laid the foundation for subsequent Large Language Models (hereafter LLMs), serving as the architectural backbone for GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), and related frameworks [6,7]. With continuous advancements in LLMs technology, these models now extend beyond natural language to diverse domains including film, music, artistic design, finance, and healthcare [8-10].

Through comparative experiments with classical portfolio optimization models, researchers have found that LLMs not only exhibit remarkable stock selection capabilities but also generate enhanced portfolio performance when integrated with traditional quantitative investment methodologies. These findings illuminate a promising avenue for intelligent decision-making through the integration of LLMs and quantitative finance. Current applications include frameworks such as the open-source projects FinGPT [11] and PIXIU [12], which construct social sentiment dynamics and financial news datasets paired with corresponding equity price movements. This approach enables LLMs to predict subsequent equity price impacts through real-time analysis of corporate news and market narratives.

Therefore, against the backdrop of ongoing research on Hong Kong-based capital and the rapid development of large language model technology [13,14], this study focuses on utilizing LLMs to accomplish financial time-series data prediction tasks. Recognizing the characteristics of financial time-series forecasting: high noise, overlapping cycles, and high sensitivity, this research will adopt the following approach: decoupling the prediction task into two dimensions, namely Hong Kong-based capital behavior prediction and technical data prediction, to obtain more comprehensive dual-dimensional predictions. Movements of Hong Kong-based capital can partially reflect the impacts of sensitive events such as international dynamics and geopolitical factors, while technical data emphasizes data-driven influences from market sentiment and trading behaviors. By integrating these differentiated prediction tasks, the study aims to reflect market conditions from multiple perspectives, thereby addressing the technical challenges of financial time-series forecasting. Subsequently, leveraging LLMs' semantic comprehension capabilities combined with tailored prompts, consolidated prediction results synthesizing insights from both Hong Kong-based capital and technical data dimensions will be generated.

In brief, the prediction framework proposed in this study is as follows: first, taking Hong Kong-based capital and technical data as the research objects from a dual-dimensional perspective; second, optimizing LLMs to adapt to time-series prediction tasks; finally, utilizing conversational LLMs for summarization, refinement, and logical integration to generate interpretable comprehensive prediction conclusions.

2. Related Work.

2.1. Related work in LLMs for finance. The development of Financial Large Language Models (FinLLMs) originated from the success of general-domain Pre-trained Language Models (PLMs), including architectures such as the GPT series, BERT, and ELECTRA. These models acquired robust language comprehension capabilities through pre-training on massive text corpora, demonstrating emergent properties like in-context learning as their parameter scales expanded. To address the specialized demands of the financial industry, researchers began exploring methods to adapt these general models into domain-specific FinLLMs. For instance, FinBERT, one of the earliest financial domain-specific models, underwent iterative refinements through continuous pre-training on finan-

cial corpora, followed by task-specific fine-tuning for financial NLP applications, thereby enhancing its comprehension of financial texts [16,17].

Subsequently, technologies such as fine-tuning and RAG provide a reliable approach with low computing power requirements for LLMs research. Researchers advanced this paradigm by scaling up language models for financial applications, giving rise to modern FinLLMs [18]. These models not only inherit the foundational strengths of PLMs but also achieve superior performance in financial tasks through increased parameter scales and sophisticated training strategies. For example, BloombergGPT, the first FinLLM developed using the BLOOM architecture, was trained on a hybrid corpus comprising 345 billion tokens of general-domain data and 363 billion tokens of specialized financial data (FinFile), enabling versatile NLP capabilities [19]. In contrast, FinGPT adopts an open-source, data-driven framework optimized via Low-Rank Adaptation (LoRA), ensuring competitive performance under resource-constrained conditions [11].

Currently, there are relatively few representative studies on financial time-series data. Yu et al. constructed a custom instruction set using the Nasdaq dataset and financial text data, and fine-tuned Llama to endow it with predictive capabilities [20].

2.2. Related work in LLMs for time series. Time-series data are widely present in multiple domains such as climate, the Internet of Things, healthcare, transportation management, and finance. Traditional time-series analysis methods include classical signal processing techniques, such as frequency analysis and decomposition methods, as well as recently emerged deep learning approaches like convolutional neural networks. Although LLMs were initially designed for training on textual data, they possess broad potential for analyzing numerical time-series data. Researchers are now exploring the feasibility of applying these models to time-series analysis. To bridge the modality gap between LLMs' original text-based training and time-series data, researchers have attempted various strategies to transfer and distill LLMs' knowledge into time-series analysis. For example, Xue and Salim proposed a prompt-based time-series forecasting method that predicts time series in a sentence-to-sentence format [22]; Chung et al. utilized a Vector Quantized Variational Autoencoder (VQ-VAE) to discretize ECG time-series data and generated 12-lead ECGs based on clinical text reports [23]; Liu et al. adopted the concept of alignment to embed ECG time-series data into the semantic space of language models for cross-modal learning [24].

These methods not only demonstrate the extensive application prospects of LLMs in time-series tasks but also reveal potential future research directions. We can anticipate that LLMs will bring further breakthroughs in the field of time-series analysis.

2.3. Our contributions. As elaborated in Section 2.1, current research on large models in the financial domain represents an evolution from conventional PLMs to more complex and large-scale LLMs. These models integrate advanced techniques such as cross-domain pre-training, instruction fine-tuning, and prompt engineering, thereby enhancing their analytical capabilities across various types of financial texts [21].

As the core carrier of stock information, time series directly reflect market information, which is more direct and relevant than textual information. Meanwhile, this also constitutes a gap in the current application of LLMs in the financial field. Therefore, based on the foundation of FinLLMs research and inspired by the concept of LLMs for time series, we propose a novel stock prediction algorithm – using LLMs to predict stock time-series data. Unlike existing FinLLMs cases that use textual data such as news and social media to enable LLMs to analyze text and predict stock trends through fine-tuning, our approach employs time-series data from two dimensions: Hong Kong-based capital

investment behaviors and daily technical data of stocks, to comprehensively reflect market trading dynamics.

For the prediction algorithm, we improve the Time-LLM model to enhance its domain adaptability for stock forecasting tasks, and propose a new prediction method based on the modified model. Specifically, we utilize Chain-of-Thought prompting to guide task decomposition, thereby improving task completion accuracy, and implement a “Deep Breath” mechanism to reduce error accumulation, making it more suitable for vertical applications in stock time-series prediction. The predictive capabilities of the optimized model are quantified through comparative experiments. The model generates multi-dimensional prediction results. To synthesize intuitive forecasts that integrate both capital flow and technical data perspectives, we introduce GPT-4, leveraging its semantic comprehension and reasoning capabilities to summarize time-series results from capital flow and technical indicator dimensions, ultimately producing direct and interpretable future market predictions. Detailed implementation will be elaborated in Chapter 3.

3. Algorithm Design.

3.1. Original Time-LLM model. The original architecture of Time-LLM is illustrated in Figure 1. Given an input sequence $X \in R^{N \times T}$ containing N variables of length T , the model predicts future values $Y \in R^{N \times H}$ for H sampling points. Its primary innovation lies in the modality conversion process for input transformation [15].

To enable language models to better comprehend time-series data, Time-LLM aligns time-series data with word embeddings. Pre-trained word representations $E \in R^{V \times D}$ are

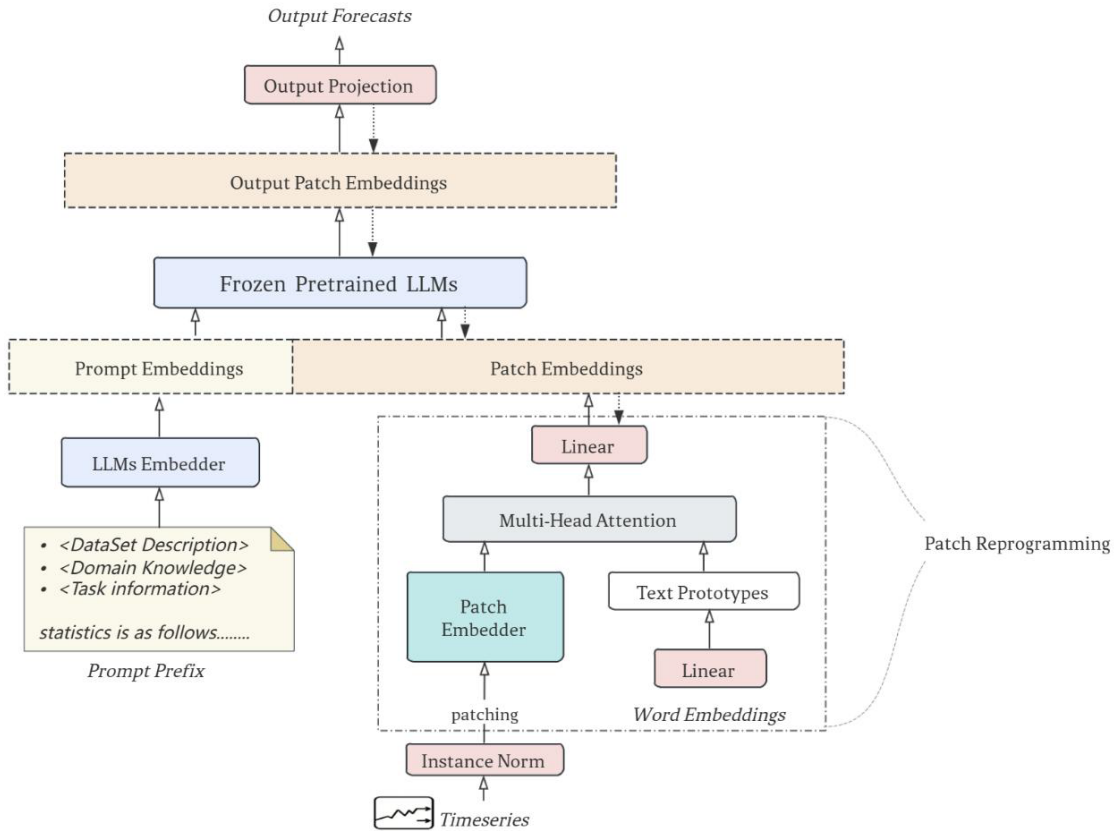


FIGURE 1. Original Time-LLM model reprogramming the input time series into text prototype representations that are more naturally suited to language models’ capabilities

linearly combined to obtain text prototypes $E' \in R^{V' \times D}$ where $V' \ll V$, which capture the variational characteristics of time-series data. Through a multi-head self-attention mechanism, the model adaptively derives text descriptions for each input channel's segmented patches: $Z_k^{(i)} = \text{ATTENTION}(Q_k^{(i)}, K_k^{(i)}, V_k^{(i)})$ where $Q_k^{(i)} = \hat{X}_p^{(i)} W_k^Q$, $K_k^{(i)} = E' W_k^K$, $V_k^{(i)} = E' W_k^V$. The outputs from multiple heads are concatenated to form a single-channel time-series representation. This representation is then prefixed with prompt tokens and fed into the frozen LLM, with outputs parsed analogously.

3.2. Optimized Time-LLM model.

3.2.1. Optimization overview. To enhance the original model's real-world applicability in stock time-series prediction, we have optimized the relatively simplistic Prompt-as-Prefix stage while retaining the Patch Reprogramming procedure. To improve prediction accuracy, we design distinct Chain-of-Thought paths for different prediction subtasks through Task Decomposition. To mitigate the continuous accumulation of errors in long-term tasks, we adopt the Deep Breath mechanism. The improved architecture, illustrated in Figure 2, will be detailed in subsequent chapters.

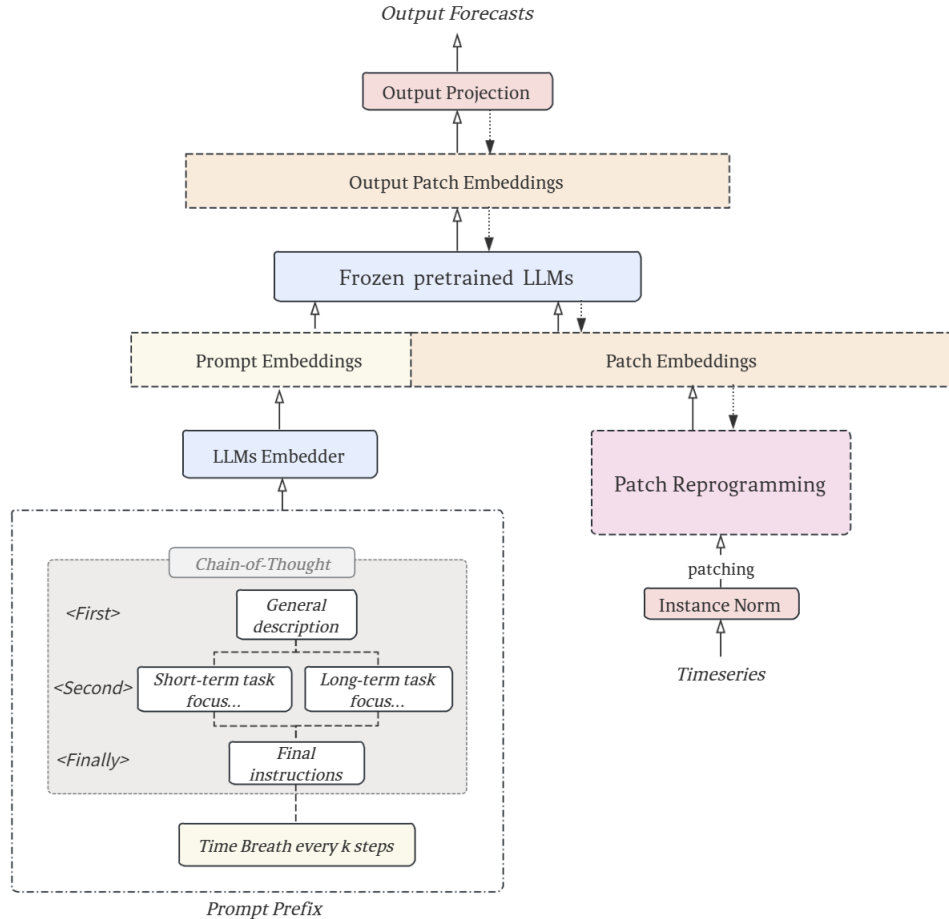


FIGURE 2. Our model optimized the prompt-prefix strategy in Time-LLM by breaking down the major task into multiple smaller tasks with Chain-of-Thought and incorporating Deep Breath mechanisms. To simplify the visualization, we consolidated the patch reprogramming section of Figure 1 into a unified Patch Reprogramming module while preserving its functionality.

3.2.2. Task decomposition. Although Time-LLM demonstrates robust generalization capabilities in generic scenarios, its domain adaptability still requires optimization for high-noise, multi-periodic, and highly sensitive financial time-series forecasting tasks. To enhance the model’s domain-specific proficiency in stock market prediction, we innovatively design a Chain-of-Thought prompting.

Compared to requiring LLMs to directly comprehend complex prompts and complete intricate tasks, decomposing tasks into ordered, simpler subtasks enables superior performance by helping LLMs construct coherent reasoning pathways [26]. Following this principle, we partition the prediction task into long-term and short-term mechanisms based on temporal spans, and then guide LLMs to address distinct subtasks for different dimensions of stock time-series inputs, thereby focusing on diverse temporal characteristics. For short-term predictions, the model prioritizes technical data anomalies such as price-volume divergence and abnormal daily trading volumes of Hong Kong-based capital to detect transient market sentiment impacts. For long-term predictions, it emphasizes quarterly moving averages and latent seasonal fluctuations in technical data, along with position stability and trend continuity in Hong Kong-based capital flows, which reflect institutional confidence in long-term stock valuations and provide critical references for investors. By progressively guiding LLMs through differentiated reasoning chains tailored to task-specific temporal features, the model adapts more effectively to technical data analysis and Hong Kong-based capital dynamics prediction tasks.

3.2.3. Deep Breath mechanism. In our research, we observed that as the prediction window expands, the deviation between predicted and actual values progressively increases. This phenomenon may stem from implicit temporal dependencies within model-generated predictions, leading to cumulative error propagation that renders long-term forecasts increasingly divergent.

To enable LLMs to maintain sustained high reasoning accuracy and avoid cognitive collapse induced by excessively long prediction contexts, we implement the “Deep Breath” mechanism, a prompting technique empirically demonstrated to enhance LLM performance [27]. Integrated into our prompting, this mechanism guides LLMs to perform rhythmically timed pauses during sequential time-series reasoning. For instance, in time-series forecasting tasks with an H -step prediction horizon, LLMs are instructed to “breathe” at every k -step interval, where k is a hyperparameter determining the respiration frequency (detailed in the experiments section). Through this approach, LLMs periodically reassess their prediction mechanisms, gradually attenuating the influence of temporally distant steps that minimally impact the current reasoning window. This prevents interference from irrelevant prior inferences and enhances model adaptability to evolving prediction contexts. Particularly in technical indicator analysis, strategically inserting corrections at critical junctures such as resistance or support yields measurable performance improvements.

3.3. Stock prediction algorithm based on Hong Kong capital and factors. Our prediction algorithm structure is illustrated in Figure 3 and consists of two phases: the prediction phase and the summarization phase. The prediction phase includes the Hong Kong-based capital behavior predictor and the technical data predictor. The input sequences for both predictors must ensure identical time spans and the same number of trading days. We obtain Hong Kong-based capital flow data from HKEX and crawl daily stock technical indicator (High, Low, Open, Close, Vol) using Tushare as technical data. After normalizing both datasets, they are fed into the optimized Time-LLM prediction model described earlier. The model performs modality conversion and generates prediction

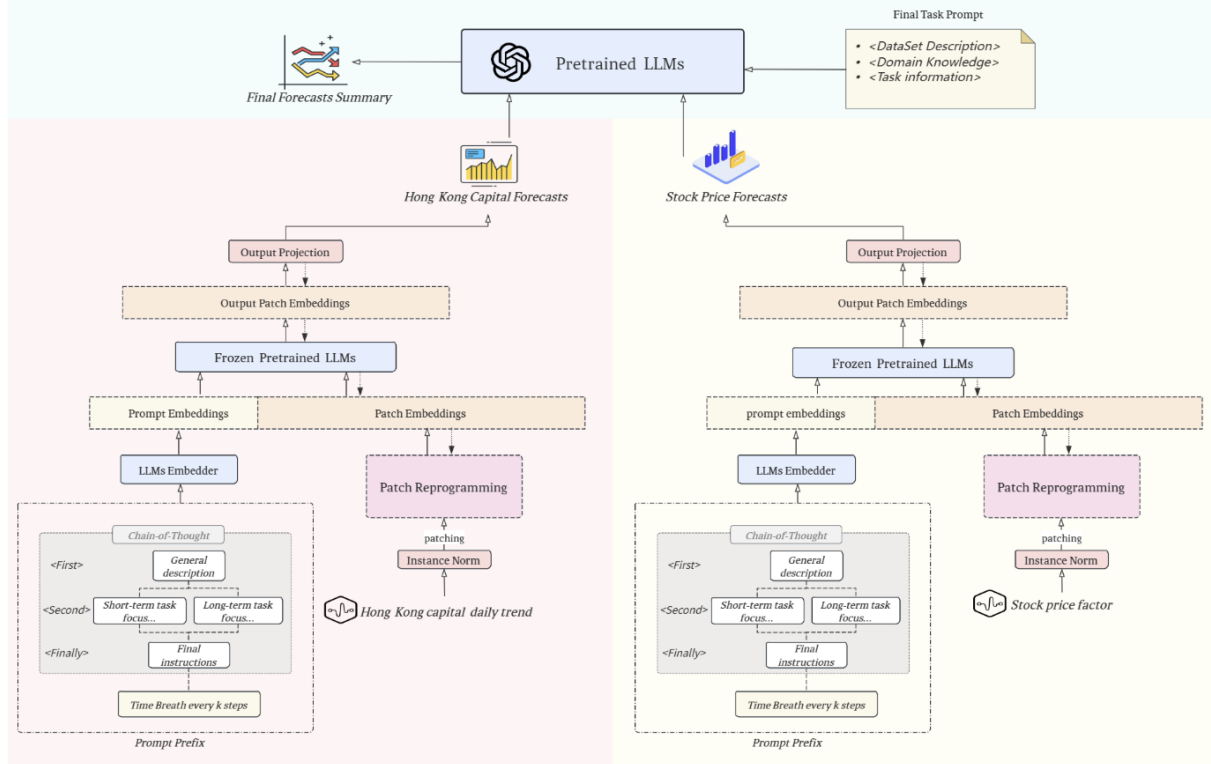


FIGURE 3. Conceptual architecture of our algorithm consisting of prediction phase which uses our model and summary phase which used pretrained LLM

outputs, ultimately mapping the textual modality back to the time-series modality to obtain individual predictor results.

In the summarization phase, the multi-dimensional outputs from both predictors are integrated. The Hong Kong-based capital prediction results and technical data prediction results are embedded into a Chain-of-Thought prompt designed to generate the final market trend conclusion. GPT-4 synthesizes these inputs to produce a consolidated stock future trajectory that combines insights from both Hong Kong-based capital and technical data predictions. To ensure the model's final output reflects comprehensive multi-factor reasoning, the results are structured in a "Conclusion + Rationale" format, as exemplified in Figure 4.

4. Experimental Section.

4.1. Experiments in prediction phase. The objective of this experimental section is to validate the predictive performance of the Hong Kong-based capital predictor and the technical data predictor. To ensure uniformity in input sequence lengths, we selected Hong Kong-based capital flow data and corresponding daily technical indicators spanning 252 trading days from June 25, 2023, to June 25, 2024, as the standardized input sequences. Predictions were generated for distinct short-term and long-term horizons: short-term outputs involved sequence lengths of [32, 48] time steps, while long-term outputs used lengths of [96, 128] time steps. To quantify the disparity between predicted results and ground-truth historical data, we adopted Mean Squared Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics. All reported experimental results represent averages calculated across different output lengths.

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

The user receives two input time series of length 252: one representing Hong Kong capital institutions' trading activity (e.g., buy/sell volume) for a single stock over 252 trading days, and the other containing daily price factors (e.g., open, close, high, low, volume) for the same stock during the period. Rolling forecasts with a step size of 48 are performed on both series. To synthesize the Northbound capital trading behavior and price factor volatility, the following considerations apply:

• **Hong Kong Institutional Trading Activity:**

Forecast data indicates institutional trading volume peaked at a 7.2% increase during this period. Despite irregular declines, the overall trend shows a stable, moderate upward trajectory. However, the absence of sustained buying activity in the final 11 trading days suggests institutions no longer expect the stock's gradual appreciation to persist. This implies a potential minor correction in the near term.

• **Price Factors & Volume Analysis:**

In the initial 15 trading days, closing prices slightly exceeded opening prices more frequently, coupled with an expanding trading volume trend. This signals that upward momentum was likely to persist. However, subsequent sessions showed no clear dominance of closing-over-opening price relationships, indicating the upward momentum gradually weakened. The price eventually entered prolonged consolidation, suggesting a modest upside potential within a range of $^{**}+0.5\%$ to $+5\%^{**}$.

• **A-Share Market Volatility:**

Given the relatively high volatility of China's A-share market compared to global peers, forecast accuracy is inherently constrained. Prudence dictates adopting a conservative strategy to account for market instability and unpredictable swings.

Synthesis of Factors:

While both time series forecasts (institutional flows and price factors) hint at modest upside potential, the high volatility of A-shares and weakening institutional conviction in the latter phase necessitate tempering optimism. The final consolidated recommendation is: **Range-bound movement (-0.5% to +0.5%)**.

FIGURE 4. Output example of interpretable stock prediction conclusion, demonstrating the final forecast with explicit rationale for each analytical component

In model selection, we categorize the selected models into three classes: Pretrained, End-to-end, and LLM Cross-modal. Pretrained models are pre-trained on large-scale datasets, demonstrating strong cross-domain generalization capabilities, representative models include TimeGPT; End-to-end models are trained directly on the target dataset without pre-training, and representative models include Dlinear, N-Beats and AutoFormer; while LLM Cross-modal models adapt existing LLMs to time-series tasks through input/output format adjustments, and our model and TimeLLM are included. By comparing these representative models across heterogeneous domains, we obtained comprehensive experimental results.

The hyperparameter configurations for the experiments are detailed in the following table. Notably, the setting of the Deep Breath frequency k was informed by conclusions from its original study [27].

TABLE 1. Hyperparameter settings for experimental setup

seq_len	pred_len	batch_size	learning_rate	train_epochs	breathK
252	96, 128	4	0.005	100	5

For the Hong Kong-based capital behavior predictor, through an analysis of trading behaviors across multiple Hong Kong-based institutions, we categorize their investment activities into two primary types: position-holding and allocation-oriented. We selected The Hongkong and Shanghai Banking Corporation (HSBC) as a representative of allocation-oriented capital and Interactive Brokers Hong Kong Limited as a proxy for

trading-oriented capital. A stock pool comprising equities with long-term purchase records from both institutions was established to conduct predictions for each investment style.

Experimental results in Table 2, averaged across multiple trials for both short-term and long-term predictions, demonstrate that our model achieves superior error metrics. In short-term forecasting tasks, it outperformed baseline models by at least 3% in MAE and 8.2% in MSE. For long-term forecasting, the improvements reached 9.1% in MAE and 7.5% in MSE. However, due to the inherent complexity of long-term forecasting and market volatility, all models exhibited increased prediction errors compared to short-term tasks.

TABLE 2. Short & long-term time-series forecasting results

Model	MAE		MSE		RMSE		Type
	<i>stm</i>	<i>ltm</i>	<i>stm</i>	<i>ltm</i>	<i>stm</i>	<i>ltm</i>	
DLinear	0.991	1.524	1.110	1.892	1.053	1.375	End-to-end
N-Beats	0.939	1.339	1.482	2.010	1.217	1.418	End-to-end
TimeGPT	0.774	1.005	0.964	1.057	0.982	1.028	Pretrained
AutoFormer	0.856	1.410	1.134	1.883	1.065	1.372	End-to-end
Time-LLM	0.820	0.986	0.850	1.091	0.922	1.044	LLM Cross-modal
Our model	0.772	0.977	0.843	1.049	0.918	1.024	LLM Cross-modal

For the technical data predictor, in experiments evaluating its predictive performance, we used the same stock pool and conducted multiple trials for both short-term and long-term tasks, with results averaged as shown in Table 3. The experimental results indicate that the performance gaps between models narrowed in technical data forecasting, and errors in long-term tasks were effectively controlled. Our model still demonstrates superior performance in both short-term and long-term technical data prediction tasks. In short-term tasks, the MSE (0.711) slightly exceeds that of the pre-trained large time-series model TimeGPT (0.709), but the MAE outperforms the second-ranked TimeGPT by 7.2%. In long-term tasks, our model surpasses other baseline models by at least 1.2% in MAE and 7.0% in MSE.

TABLE 3. Short & long-term time-series forecasting results

Model	MAE		MSE		RMSE		Type
	<i>stm</i>	<i>ltm</i>	<i>stm</i>	<i>ltm</i>	<i>stm</i>	<i>ltm</i>	
DLinear	0.826	1.091	1.031	1.108	1.015	1.053	End-to-end
N-Beats	0.887	1.030	0.910	1.092	0.954	1.045	End-to-end
TimeGPT	0.686	0.804	0.709	0.898	0.842	0.948	Pretrained
AutoFormer	0.742	0.931	0.797	0.994	0.893	0.997	End-to-end
Time-LLM	0.690	0.811	0.776	0.901	0.881	0.949	LLM Cross-modal
Our model	0.681	0.803	0.711	0.892	0.843	0.945	LLM Cross-modal

4.2. Experiments in summarization phase. In the summarization phase, we embed the outputs from both predictors into prompts and submit them to GPT-4 for generating final investment recommendations. The desired output format, exemplified in Figure 4, follows a “Conclusion + Rationale” structure where the output specifies the anticipated price movement magnitude over a future period. We adopt a categorical prediction approach, segmenting individual stock volatility ranges based on the CSI 300 Index’s volatility σ . Considering the significant market impact on individual stocks and the gradual weakening of the correlation between individual stock volatility and market volatility

TABLE 4. Volatility threshold set

CSI 300 volatility	Stock volatility threshold
Ultra-low volatility ($\sigma < 15\%$)	Stock $\sigma < 20\%$
Low volatility ($15\% \leq \sigma < 25\%$)	$20\% \leq \text{Stock } \sigma < 35\%$
Moderate volatility ($25\% \leq \sigma < 35\%$)	$35\% \leq \text{Stock } \sigma < 50\%$
High volatility ($35\% \leq \sigma < 45\%$)	Stock $\sigma \geq 50\%$
Extreme volatility ($\sigma \geq 45\%$)	Stock $\sigma \geq 50\%$

TABLE 5. Summarization phase forecasting result

Model	ACC	Precision	Recall	Marco-F1	Kappa
OnlyGPT	0.577	0.519	0.523	0.495	0.397
Our model	0.695	0.622	0.701	0.680	0.483

as market volatility rises, we have appropriately increased the volatility thresholds for individual stocks, with the interval settings shown in Table 4. By comparing the predicted movement categories against actual market price trajectories, we compute classification metrics to quantify our model’s categorical accuracy. To ensure result reliability, we conducted multiple experiments across diverse equities, contrasting our outputs with baseline predictions where GPT-4 directly generates forecasts from historical trends without predictor inputs. As shown in Table 5, our predictions significantly outperform the baseline, validating the method’s efficacy.

5. Conclusion. This paper first improves LLMs which are originally specialized in natural language processing, by incorporating the principles of modality alignment and prompt engineering. By combining traditional deep learning with prompting techniques, we enhance their performance in time-series forecasting tasks. The refined model is then applied to predicting target stocks across two dimensions: Hong Kong-based capital dynamics and technical data. Finally, advanced LLMs synthesize predictions from both dimensions to forecast stock volatility magnitudes within preset future time windows, providing actionable references for investors. To validate the method’s effectiveness, we conducted experiments in both the prediction and summarization phases. In the prediction phase, the model achieved state-of-the-art performance compared to traditional neural networks and novel time-series LLMs. In the summarization phase, our model demonstrated superior accuracy in predicting volatility magnitudes compared to results generated by GPT-4 without leveraging the predictors.

The enhanced model can be embedded into quantitative hedging strategy frameworks to enable dynamic multi-factor portfolio optimization. Particularly in predicting Hong Kong Stock Connect capital flows, the model can detect anomalies in northbound capital movements in advance, offering high-frequency signal support for cross-border arbitrage strategies.

It should be noted that the model has limitations in responding to extreme events outside historical data distributions, such as sudden regulatory changes or black swan events triggering market circuit-breaker scenarios. When technical indicators diverge directionally from capital flow trends, the probabilistic reasoning mechanisms of LLMs may amplify misjudgment risks. Additionally, under special market conditions with liquidity shortages, systematic deviations may occur in the mapping between time-series data and language modalities.

In the future, using LLMs for stock trading may become feasible, highlighting the significance of this method in integrating LLMs with quantitative trading. Future research could extend to a three-dimensional prediction algorithm, augmenting current technical and capital flow analyses with multimodal alignment techniques to incorporate unstructured data such as corporate financial reports and industry policy interpretations. Leveraging LLMs' semantic comprehension capabilities, a more comprehensive stock prediction system could be constructed.

REFERENCES

- [1] H. Xu and S. Li, What impacts foreign capital flows to China's stock markets? Evidence from financial risk spillover networks, *International Review of Economics & Finance*, vol.85, pp.559-577, 2023.
- [2] Y. Meng, L. Xiong, L. Xiao and M. Bai, The effect of overseas investors on local market efficiency: Evidence from the Shanghai/Shenzhen-Hong Kong Stock Connect, *Financial Innovation*, vol.9, no.1, 42, 2023.
- [3] A. Akusta, Analysis of the relationship between cross capital flows and stock exchange index with machine learning, *Abant Sosyal Bilimler Dergisi*, vol.24, no.1, pp.244-263, 2024.
- [4] C. Liao, G. Tang and X. Xu, Smart money or chasing stars: Evidence from northbound trading in China, *International Journal of Finance & Economics*, vol.29, no.2, pp.1781-1803, 2024.
- [5] A. Vaswani, N. Shazeer, N. Parmar et al., Attention is all you need, *Advances in Neural Information Processing Systems*, vol.30, 2017.
- [6] J. D. M. W. C. Kenton and L. K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of NAACL-HLT*, vol.1, 2, 2019.
- [7] A. Radford and K. Narasimhan, Improving language understanding by generative pre-training, *North American Chapter of the Association for Computational Linguistics*, 2018.
- [8] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi and C. Rizzo, ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health, *Frontiers in Public Health*, vol.11, 1166120, 2023.
- [9] X. Zhang and Q. Yang, Xuanyuan 2.0: A large Chinese financial chat model with hundreds of billions parameters, *Proc. of the 32nd ACM International Conference on Information and Knowledge Management*, pp.4435-4439, 2023.
- [10] Y. Zeng, G. Wei, J. Zheng, J. Zou, Y. Wei, Y. Zhang and H. Li, Make pixels dance: High-dynamic video generation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8850-8860, 2024.
- [11] H. Yang, X. Y. Liu and C. D. Wang, FinGPT: Open-source financial large language models, *arXiv Preprint*, arXiv: 2306.06031, 2023.
- [12] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira and J. Huang, PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance, *Advances in Neural Information Processing Systems*, vol.36, pp.33469-33484, 2023.
- [13] F. B. Oriani and G. P. Coelho, Evaluating the impact of technical indicators on stock forecasting, *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp.1-8, 2016.
- [14] J. Ye, W. Zhang, K. Yi, Y. Yu, Z. Li, J. Li and F. Tsung, A survey of time series foundation models: Generalizing time series representation with large language model, *arXiv Preprint*, arXiv: 2405.02358, 2024.
- [15] M. Jin, S. Wang, L. Ma et al., Time-LLM: Time series forecasting by reprogramming large language models, *arXiv Preprint*, arXiv: 2310.01728, 2023.
- [16] A. H. Huang, H. Wang and Y. Yang, FinBERT: A large language model for extracting information from financial text, *Contemporary Accounting Research*, vol.40, no.2, pp.806-841, 2023.
- [17] Z. Liu, D. Huang, K. Huang, Z. Li and J. Zhao, FinBERT: A pre-trained financial language representation model for financial text mining, *Proc. of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, pp.4513-4519, 2021.
- [18] Y. Hong and D. Kim, Innovative applications of RAG-enhanced small LLM for closed-domain Q&A, *International Journal of Innovative Computing, Information and Control*, vol.21, no.2, pp.481-490, 2025.
- [19] S. Wu, O. Irsoy, S. Lu et al., BloombergGPT: A large language model for finance, *arXiv Preprint*, arXiv: 2303.17564, 2023.

- [20] X. Yu, Z. Chen, Y. Ling et al., Temporal data meets LLM-explainable financial time series forecasting, *CoRR*, 2023.
- [21] Y. Li, S. Wang, H. Ding and H. Chen, Large language models in finance: A survey, *Proc. of the 4th ACM International Conference on AI in Finance*, pp.374-382, 2023.
- [22] H. Xue and F. D. Salim, PromptCast: A new prompt-based learning paradigm for time series forecasting, *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- [23] H. Chung, J. Kim, J. M. Kwon, K. H. Jeon, M. S. Lee and E. Choi, Text-to-ECG: 12-lead electrocardiogram synthesis conditioned on clinical text reports, *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1-5, 2023.
- [24] C. Liu, Z. Wan, S. Cheng, M. Zhang and R. Arcucci, ETP: Learning transferable ECG representations via ECG-text pre-training, *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.8230-8234, 2024.
- [25] J. Wei, X. Wang, D. Schuurmans et al., Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems*, vol.35, pp.24824-24837, 2022.
- [26] H. Liu, Z. Zhao, J. Wang, H. Kamarathi and B. A. Prakash, LSTPrompt: Large language models as zero-shot time series forecasters by long-short-term prompting, *arXiv Preprint*, arXiv: 2402.16132, 2024.
- [27] S. Liu, C. Chen, X. Qu, K. Tang and Y. S. Ong, Large language models as evolutionary optimizers, *2024 IEEE Congress on Evolutionary Computation (CEC)*, pp.1-8, 2024.

Author Biography



Jianming Li received the bachelor's degree in Ship Engineering from Dalian University of Technology, China, in 1999; the M.Sc. degree in Computer Application Technology from Dalian University of Technology, in 2002; the Ph.D. degree in Computer Application Technology from Dalian University of Technology, in 2007.

Dr. Li is currently a full-time associate professor at the Dalian University of Technology. His main research interests include the machine learning, classification and prediction algorithms of deep learning, software automation, and quantitative analysis and strategy research in the financial field. He has published over 50 papers in journals and conferences.



Tianwen Feng received the bachelor's degree in engineering degree, majoring in Software Engineering from Yunnan University, China, in 2023.

Mr. Feng is currently studying for a master's degree in Dalian University of Technology, China. Deep learning, large language models, time series forecasting, and the study of retrieval augmented generation technologies are included in his primary research areas.



Xiangpei Hu received his B.S. (1983), M.S. (1987) and Ph.D. (1996) degrees from Harbin Institute of Technology, China, respectively. He is a professor of Management Science at Dalian University of Technology, China, “Distinguished Young Scholars” of National Natural Science Foundation of China (NNSFC), “Chang-jiang Scholars Distinguished Professor” of Ministry of Education (MOE) of China.

His research and teaching interests are electronic commerce, supply chain and logistics management, intelligent operations research and the real-time optimization control for dynamic systems. He has published over 200 scholarly papers in refereed journals.