Carbon Price Forecasting with LLM-based Refinement and Transfer-Learning

Haiqi Jiang[1], Ying Ding[1], Rui Chen[1], and Chenyou Fan[1][0000−0002−9835−8507]

South China Normal University, Guangdong, China
fanchenyou@scnu.edu.cn

Abstract. We propose a unified forecasting framework for accurately predicting carbon markets of EU Emission Trading Scheme (EU ETS) and Chinese Emission Allowance (CEA). Our framework utilizes a Time-Series Model (TSM) for initial prediction followed by applying a Large Language Model (LLM) to refine the forecasts. We prompt the LLM to refine the TSM forecasts by demonstrating an example pair of past TSM predictions and their corresponding true future prices to the LLM as a chain-of-thought. The in-context learning capacity of the LLM allows the LLM to rectify inaccurate predictions to reflect on TSM predictions and refine the forecasts. To further reduce the prompting delays and expenses involving LLMs, we innovate a post-finetuning approach to train a Gated Linear Unit (GLU) model to condense the LLM's in-context learning capability. This enables direct fine-tuning of TSM outputs without the need for explicit prompting LLM during inference. Experimental results show that our method can refine the TSM prediction by 10% to 40% in various zones, as well as enhance transfer learning by 10% to 21% through the inclusion of market context of the source zone when predicting the target zone. Remarkably, our GLU model achieves comparable, and in some cases superior, performance compared to LLM prompting. It effectively combines the short-term forecasting capability of classical Time Series Models with the long-term trend prediction ability typically associated with the LLMs.

Keywords: Carbon Future Market, Price Forecasts, Large Language Models, Transfer Learning, Time-Series Prediction, Gated Linear Unit, Post-Finetuning

1 Introduction

The recent proposal of carbon neutrality aims to eliminate net carbon emissions in the next 20 to 30 years. To regulate economic activities towards this goal, countries like China and the European Union have established carbon markets where emission allowances can be traded. Industrial manufacturers can either purchase more allowances or reduce their own emissions, promoting cleaner energy and encouraging innovation.

Accurately predicting carbon prices can help manufacturing companies minimize costs through effective planning, while also offering valuable insights to governments for regulating domestic industrial sectors. The challenge in predicting the carbon market arises from the lack of sufficient data in emerging novel markets, as well as the presence of non-linearity and volatility, rendering traditional prediction methods less effective.

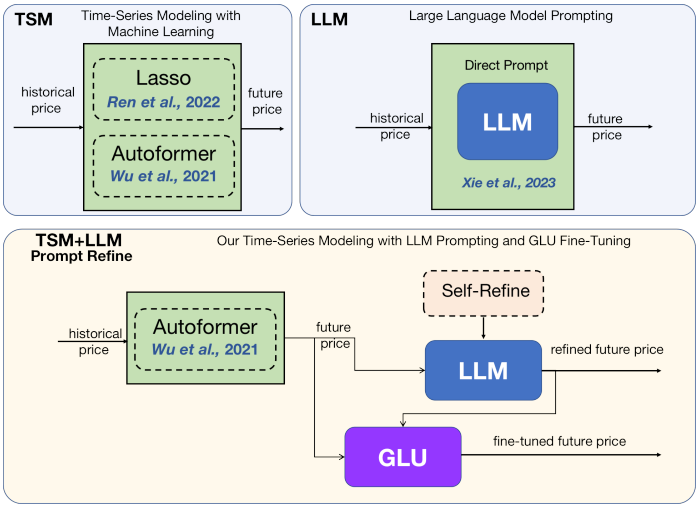


Fig. 1: Motivation of using both TSM (Time-Series Model) and LLM (Large Language Model) for carbon market forecasting. The top-left box shows studies of using TSM methods such as Lasso and Autoformer for time-series prediction. The top-right box shows recent works of prompting LLM with historical prices for future predictions. The bottom box shows our methods of using Autoformer for raw prediction followed by prompting the LLM to refine the model outputs. We also design fine-tuning the TSM output with a GLU (Gated Linear Unit) to replace LLM for efficient inference.

To tackle this challenge, we propose enhancing carbon price forecasting by integrating machine-learning-based Time-Series Models with recent advancements in Pre-trained Large Language Models in the field of AI.

The recent emerging Large Language Models (LLMs) have been proven to possess robust Few-Shot learning capability [2, 8]. Through pre-training on extensive human knowledge, LLMs acquire a rich understanding not only of human languages, but also in mathematical deduction and reasoning [17], as well as financial marketing [5]. Our paper tries to answer the following important research questions: Do LLMs understand the price trends of the EU Emission Trading Scheme and the Chinese carbon emission market? Can LLMs have abilities to refine the prediction results of the machine learning models? We address these questions by constructing a Time-Series Model and Large Language Model based two-stage time-series forecasting framework.

Firstly, we gather data of EU ETS carbon prices from 2009-2020, as well as Chinese Emission Allowance (CEA) prices from 2015-2022. Following [12], we also collect macro-economic influencing factors as predictors such as commodity prices of oil and coal, as well as stock indices regarding cleaner energy for both EU and China. We pre-process the data with missing data filling and feature selection using a traditional Lasso-based method [9].

Next, we train a state-of-the-art deep-learning-based Time-Series Model (TSM) to fit temporal patterns of historical carbon prices in a standard supervised way using historical data. To this end, we employ the state-of-the-art attention-based Autoformer [19]

as our TSM backbone Deep Neural Network and perform end-to-end training. Subsequently, we employ the well-trained TSM to generate raw future predictions based on the learned market context patterns condensed in the model parameters.

Secondly, we leverage the capabilities of the Large Language Models to incorporate market context, world knowledge, and associations of regional markets into predictions. To accomplish this, we incorporate advanced prompting techniques such as Chain-of-Thought (CoT) [17] and Self-Refine [7] into the prompting LLM procedures. The CoT technique enables successful predictions based on observable historical patterns, such as preceding time periods, thereby providing the LLM with relevant market context. The Self-Refine method encourages the LLM to actively reflect on past predictive inaccuracies and refine its current predictions accordingly.

Moreover, we consider a realistic case of predicting carbon price at a novel market even without supervised training on its historical data. We innovate an approach of demonstrating the price trend of a mature regional market in a global context followed by predicting the future of the novel market.

Finally, we aim to minimize communication and prompting costs associated with involving LLMs. We introduce a post-finetuning approach, condensing the LLM's in-context learning capability into a gated linear unit model. This enables direct fine-tuning of TSM outputs without the need for explicit LLM prompting during inference. We will demonstrate that this TSM post-finetuning approach can effectively match the capacity of the LLM refinement process and achieve comparable or superior prediction accuracy.

Overall, we conduct a comprehensive series of empirical analyses to show that the LLM could significantly refine the TSM prediction by 9.6% to 31% by merely one demonstration, and improve transfer learning by 14% to 20% without any historical training data of a novel market. Moreover, our post-finetuning process with designed GLU model achieves comparable, and in some cases superior, performance compared to LLM prompting.

In summary, the main contributions of our work include:

1. We are among the first to study the critical topic of predicting EU Emission Trading Scheme and Chinese Emission Allowances for social good. We leverage advanced AI models to incorporate diverse economic influencers as market contexts effectively into our forecasting methodology.
2. We propose a novel two-stage framework that utilizes a capable Time-Series Model (TSM) for initial prediction, then prompting the state-of-the-art LLMs to enhance the forecasts with demonstrated market contexts by learning from past deviations.
3. We also show to utilize the LLM to transfer future predictions generated by TSM from one mature region to an emerging market, thereby improving prediction accuracy even in the absence of historical data.
4. We further condense the in-context learning capability of the LLM into a GLU model, allowing for direct fine-tuning of TSM outputs without prompting the LLM during inference.
5. Our methods improve the prediction accuracy of carbon market prices by 2%-57% in movement trend classification and 9.6%-40% decrease in regression MSE. In transfer learning, our method shows 3-22% and 10-21% improvement in trend classifica-

tion and regression, respectively. Our GLU model also achieves comparable performance in both ETS and CEA markets without need of prompting the LLMs.

2 Related Work

Our research is based on the following aspects and tries to explain whether the LLMs can understand and accurately predict the price of EU-ETS and China CEA markets.

2.1 Time-Series Modeling (TSM) and Carbon Market Prediction

As the problem of carbon emissions has become increasingly prominent, the literature begins to study how to predict carbon prices more accurately through advanced artificial intelligence algorithms. Previous studies focus on using time-series models to predict the price and they find that advanced machine learning methods can improve the prediction accuracy. [24] developed a general carbon price prediction framework based on decomposition-synthesis. [23] proposed to decompose multi-dimensional data to capture both long-term trends and short-term fluctuations. [12] recently proposed to use Quantile Group Lasso for feature selection and carbon futures price prediction in EU ETS market. [9] further proposed to use of adaptive sparse Quantile Group Lasso for more robust price predictions. However, the research on the price prediction of China's carbon emission market is not comprehensive, and some of the studies are only aimed at the earliest carbon emission markets in Beijing and Guangdong. Our study uses data from carbon markets in four regions of China and studies the potential correlation between different markets.

2.2 Large Language Models (LLMs)

Recently, LLMs such as the GPT family [2, 8] and LLaMA [15] have shown great advantages in modeling language tasks, such as arithmetic reasoning and question answering. Several studies propose to simulate the human thinking process by LLMs, such as thinking step-by-step with Chain-of-Thought [17], reflecting on past experience [14] and making further refinement over past decisions such as Self-refine [7]. Some works also studied utilizing LLMs for financial tasks such as finance-related content generation and question answering [5, 20], extracting information from corporate policy [4], as well as mining trading signals or factors [16]. Due to the lack of specific domain knowledge, the LLMs could under-perform on specific queries such as medical QA and time-series forecasting. Researchers have adopted methods such as low-parameter fine-tuning [3, 6], external knowledge retrieval augmentation [10, 13], post-pretraining [18] and prompt-based in-context learning [1, 17, 22] to improve the output of LLMs in vertical domains and make them more professional and precise.

3 Data and Methodology

We define the carbon price forecasting task as follows. By observing the past T_h steps $\mathcal{T}_h = \{1, 2, ..., T_h\}$, we predict the subsequent T_f future steps $\mathcal{T}_f = \{T_h + 1, ..., T_h +$

Table 1: CEA Feature List

Panel A: Numerical features with descriptions.	
p_{close}	The closing price of CEA.
vol/amt	The quantity and total amount of traded CEAs.
$\boldsymbol{f}_{ind}$	The features of carbon indices including price, vol. & amt.
VXFXI	The China ETF Volatility Index as global market sentiment.
p_{oil}	China Daqing crude oil spot price.
p_{coal}	China Qinhuangdao coal spot price.

Panel B: Selected Carbon Indices of various CEA Zones	
GD	Carbon Tech 30 & 60, Mainland L-C Index
HB	Carbon Tech 60, Mainland L-C Index
SZ	Carbon Tech 60, Mainland L-C Index
SH	Carbon Tech 30, Carbon Tech 60

$T_f\}$. For the market of EU-ETS, we obverse the past 20 monthly prices and predict the future 12 monthly prices. For China CEA markets, we observe the past 48 days and predict the next 36 days. Next, we describe the data features.

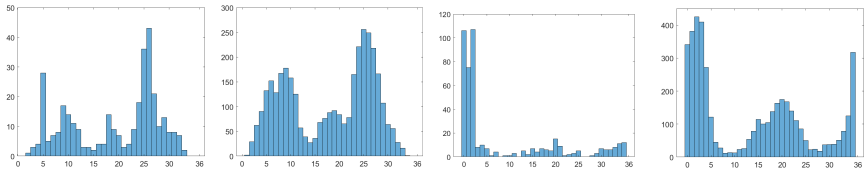
3.1 Carbon price Data and Feature Selection with Lasso-based Method

We adopt the EU carbon future price data from previous works [9, 12], which also includes 13 selected factors including crude oil and natural gas production, imports and exports of European countries, as well as economic indices such as FTSE100 index, M2 values, inflation rate and interest rates, etc. The factors have been extensively discussed in previous work [9].

Similarly, we self-collect the Chinese Carbon Emission Allowance (CEA) data from four CEA zones in China, including Hubei (HB), Shenzhen (SZ), Shanghai (SH), Guangdong (GD).

We collect from each CEA market the closing price, trade volume and amount at each day. In addition, we also collect auxiliary economic data, including the China ETF volatility index (VXFXI), China Daqing crude oil spot price, and China Qinhuangdao coal spot price. Detailed descriptions are presented in Table 1 Panel A. We also collect from China's stock market 25 carbon economy stock indices with values of closing price, volume, and amount. With this large of stock indices, we select the most relevant stock indices for predicting a specific CEA zone.

Following previous work [12], we utilize the classical Lasso method to perform stock index selection for each CEA zone. We select the top 2 or 3 most relevant features if the coefficient score is above a threshold of 0.1. In Table. 1 Panel B, we show the selected carbon indices for different CEA zones. We provide the details as follows.



(a) SH (top-1) (b) SH (top-10) (c) SZ (top-1) (d) SZ (top-10)

Fig. 2: Statistics of learned lags for SH and SZ.

Mainland L-C Index is the CSI mainland low-carbon economy index, which is composed of 50 stocks in China's A-share market involving clean energy power generation, energy conversion and storage, cleaner production and consumption, and waste disposal. Carbon Tech 60 is the CNI CIKD Carbon Neutral Technology Power Index, which selects 60 stocks from the Shenzhen Stock Exchange in the A-share market as sample stocks. Carbon Tech 30 is the SZSE ChiNext Carbon Neutral Technology Power Index, which selects 30 stocks from the Growth Enterprise Market of the Shenzhen Stock Exchange. The index compilation is based on the classification of the carbon-neutral technology service industry of the listed companies.

3.2 Time-Series Modeling (TSM) with Autoformer (AF)

We utilize Autoformer [19] as the state-of-the-art backbone Deep Neural Network for TSM to provide initial coarse carbon price forecasts. The Autoformer has two main components. The **Auto-Correlation** is a mechanism that can capture the period-based dependencies of historical price and factors X_h by computing the correlation of sub-series with different time delays and aggregating to a new series denoted by $\mathcal{X}$.

The **Series-Decomp**$(\mathcal{X})$ denotes the historical series decomposition which can separate the series into trend-cyclical $\mathcal{T}$ and seasonal parts $\mathcal{S}$ from the input series. Thus the two parts add to the future prediction $\hat{Y} \leftarrow \mathcal{S} + \mathcal{T}$.

We fit the Autoformer to the CEA data as TSM and denote this process as

$$\hat{Y}^{AF} \leftarrow AF(X_h)\ . \tag{1}$$

We can visualize the cyclic trends of the CEA price. A lag τ reflects the time-delay similarity between $\mathcal{X}_t$ and its τ lag series $\mathcal{X}_{t-\tau}$. We iterate $\tau \in [1, 2, ..., L]$ and show the top-1 and top-10 most correlated lags in Shanghai and Shenzhen CEA markets, respectively, in Fig. 2. The X-axis is τ, while the Y-axis is the count of history periods that match that specific lag. In Fig. 2 (a), we observe two prominent peaks in the top-1 lags across all test sequences, namely the 5-th day (weekly) and the 26-th day (monthly). We further explore this pattern in (b) by considering the top-10 lags and finding consistent results. Analyzing (c) and (d), we observe that the top-1 lags in Shenzhen are concentrated within the first few days, while the top-10 lags exhibit either short-term or long-term (three weeks) trends. This suggests the presence of volatility and non-linearity in the CEA markets.

4 Apply Large Language Model for Forecast Refining

In this study, we utilize GPT-3.5 as the underlying LLM for refining the forecasts of CEA markets based on in-context learning capacities.

4.1 DP - Direct Prompting LLM Methodology

We directly utilize the LLM to predict the carbon price for future steps in the future without using TSM. To this end, we prompt the LLM with "Give you historical carbon price: [...], please predict the price for next 48 days." We extract the predicted sequence from the LLM's response and denote this methodology by *DP*.

This method is also known as "zero-shot learning" which relies entirely on the world-knowledge of the LLM to predict the future [2]. We will take this as a basic baseline. We will show that our observation is consistent with previous work [21] that current LLMs are not capable of precisely predicting the stock or market price without giving enough market contexts.

4.2 CoT-RF - Joint Time-Series and Large Language Modeling

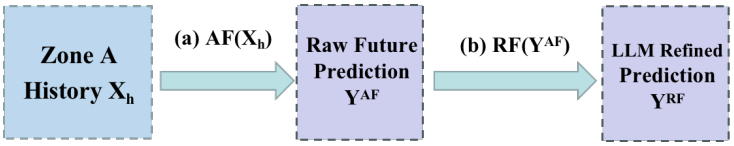


Fig. 3: Flowchart of refining Autoformer predictions with LLM by CoT-Refine. Step (a): we apply trained AF to predict the future prices. Step (b): we apply LLM to enhance AF and obtain refined future predictions.

In contrast to direct prompting methods, our approach involves utilizing TSM to generate initial predictions with one selected period of past steps, along with the corresponding true prices as references. We then employ the LLM to refine these predictions for future time steps. This methodology is similar to Chain-of-Thought (CoT) prompting, as described by Wei et al. [17]. We utilize CoT prompting to leverage the contextual learning capabilities of the LLM to extrapolate from limited examples (past predictions) and generalize to new scenarios (future predictions). We give a concrete example below.

We design the CoT prompting template as "We give you some historical carbon prices [...] and AF's predicted prices. Then I need you to improve AF's predictions and give you the real prices and let you reflect on your predictions. Then we also give you AF's predictions for next 48 days [...]. Please improve AF's next 48 days prediction."

We call this method as *CoT-RF* since we feed the step-wise future raw prediction of Autoformer $\hat{Y}^{TSM}$ to the LLM to refine. This is denoted as:

$$\hat{Y}^{RF} \leftarrow \textit{CoT-RF}(\hat{Y}^{AF}) \ . \tag{2}$$

We demonstrate this pipeline in Fig. 3. Step (a) corresponds to Eq.(1) which applies trained AF to predict the future prices. Step (b) corresponds to Eq.(2) which applies LLM to enhance AF and obtain refined future predictions.

4.3 FT - Efficient Low-Data Fine-tuning with Gated Linear Unit

The CoT-RF described above requires feeding the Autoformer predictions to the LLM for refinement, thus causing potential high communication delays and high computational costs. Additionally, utilizing commercial LLMs like ChatGPT can be prohibitively expensive due to the large number of tokens required for each prompt. Furthermore, there are considerable risks and legal issues associated with uploading sensitive private data and proprietary features to LLM providers.

We introduce an innovative post-finetuning approach to mitigate the computational expenses and privacy issues while upholding the efficacy of the LLM. We train a supplementary GLU (Gated Linear Unit) model to fine-tune the Autoformer outputs with GPT refinement results efficiently. Consequently, during inference, the GLU incurs minimal computational overhead and entirely removes the necessity for frequent communications with a cloud-deployed LLM like ChatGPT.

Inspired by the recent embedding fine-tuning of LLMs, we design a two-layer GLU with Input-Linear-Swish-Linear-Output architecture. The Swish activation function [11] is defined as:

$$\text{Swish}_{\beta_1,\beta_2}(\boldsymbol{x}) = (\beta_1^\top \boldsymbol{x}) \cdot \sigma(\beta_2{}^\top \boldsymbol{x}) \,. \tag{3}$$

Thus our designed GLU model with trainable parameters β_1, β_2, W_1 and W_2 (biases omitted) can be formulated as:

$$\mathit{GLU}(\boldsymbol{x}) = W_2 \cdot \text{Swish}_{\beta_1,\beta_2}(W_1 \boldsymbol{x}) \tag{4}$$

We fine-tune with GLU unit by learning to transform the AF raw predictions to LLM refined predictions. This post-finetuning process is as follows:

$$\begin{aligned} \hat{Y}^{AF} &\leftarrow \mathit{AF}(X_h) \\ \hat{Y}^{FT} &\leftarrow \mathit{GLU}(\hat{Y}^{AF}) \end{aligned} \tag{5}$$

Given $\hat{Y}^{RF}$ is from LLM refinement of Eq.(2), the loss function is to minimize the fine-tuning output with LLM responses, such that:

$$\mathcal{L} \leftarrow ||\hat{Y}^{FT} - \hat{Y}^{RF}||_2^2 \,. \tag{6}$$

We call this process as post-finetuning as the GLU unit refines the Autoformer without altering its heavy parameters. The compact GLU unit has only 27952 parameters, amounting to only 0.3% compared with 10587153 parameters of the Autoformer. We will demonstrate in experiments that our designed post-finetuning can quickly adapt the GLU to the LLM capacity with just a few hundred training examples.

5 Empirical Results

We compare the following methods of predicting the future CEA in three different regions, Hubei (HB), Shenzhen (SZ), and Guangdong (GD). There is a total number of 960 test steps with timestamps from 2021-03 to 2022-02.

Lasso [12] uses Lasso regression to fit historical data. **AF** (Sec. 3.2) trains the Autoformer to make future predictions with supervised learning. **DP** (Sec. 4.1) relies on the LLM to directly prompt the future price given the history. We take AF and DP as two baselines. **CoT-RF** (Sec. 4.2) is based on LLM Refinement which prompts the LLM to refine AF predictions by demonstrating its predictions and true price sequences over past steps. **GLU** (Sec. 4.3) is our proposed post-finetuning process which trains a compact GLU to simulate LLM refinement process of enhancing the AF predictions.

5.1 Evaluation tasks and metrics

We evaluate all methods with the future regression task and the trend classification task.

MSE. The regression task is measured with Mean Squared Error (MSE, lower is better) averaged over all future 30 predicted steps.

Accuracy. The 3-way future trend classification task evaluates the predicted price on Day-10, 20, and 30 as up, neutral or down, indicating the relative position at a future step compared with the mean observed price p over the historical 18 steps. We define the neutrality class as a price range within $[(1-\tau)p, (1+\tau)p]$. The range of upward and downward classes are $((1+\tau)p, \infty)$ and $(-\infty, (1-\tau)p)$, respectively. We choose τ to be 2%. For example, if the price on Day-10 is 36.0 and the mean historical price over the past window of 18 steps is 35.0, the increase is approximately 2.8%, indicating an upward trend.

5.2 EU Emission Trading Scheme (EU ETS) forecasting result analysis

We show the EU-ETS results in Table 2. Since the ETS carbon future data consists only monthly prices from 2009-03 to 2020-12, there is a total number of 182 data samples. Following [9], we split the data split to training/validation samples from 2009- 03 to 2019-12 and test on 12 monthly prices from 2020-01 to 2020-12. Due to the lack of sufficient data, we failed to train a proper deep-learning based Autoformer model. Thus we did not provide AF results.

Instead, we take the **Lasso** as the baseline and report relative increases or decreases in MSE in Table 2 of LLM-based methods. The trend classification labels are converted from the regressed price and compared with actual trend labels (up, neutral or down).

CoT-RF refines Lasso predictions with LLM in-context learning ability. **GLU** performs post-finetuning over Lasso prediction by using LLM refined results. This is achieved by minimizing the disparity between GLU predictions and CoT-RF predictions across training timesteps. Subsequently, we employ the trained GLU model to make inferences on test timesteps.

We have observed the following trends in our analysis: The GLU model exhibits the lowest MSE, with CoT-RF following closely in second place. Both GLU and CoT-RF demonstrate a highest 67% accuracy rate in predicting trend classes. Notably, GLU,

Table 2: EU-ETS carbon future price forecasting results.

Method	MSE ↓	Accuracy ↑
Lasso ([12])	7.46 (0%)	58%
DP ([17])	10.16 (36% ↑)	33%
CoT-RF (ours)	6.50 (13% ↓)	**67%**
GLU (ours)	**6.41** (14% ↓)	**67%**

This table presents the Mean Squared Error (MSE) for predicting the subsequent 12 months of European Union Emission Trading Scheme (EU-ETS) data for year 2022, along with the mean Accuracy of 3-way trend classification. The benchmark model is **Lasso** [12], which uses Lasso as the baseline which regresses on historical data. **DP** relies on the LLM to directly prompt the future price given the history. **CoT-RF** uses LLM to learn patterns from a complete Lasso prediction example on past steps as chain-of-thought then improves the AF predictions over new observed history. **GLU** performs post-finetuning over Lasso prediction by using LLM refined results.

which is post-finetuned from CoT-RF results across training steps, exhibits very similar performance to CoT-RF. In fact, GLU even surpasses CoT-RF in predicting future steps, indicating its superior generalizability to unobserved future data. On the other hand, DP achieves the lowest performance among the models considered.

5.3 The Chinese Emission Allowance (CEA) forecasting result analysis

We show the CEA price predictions over 3 zones in Table 3. We take the AF as the baseline and report relative increases or decreases in MSE of other methods. We observe trends as follows.

Firstly, the **Lasso** method exhibits significantly larger MSE when compared to other methods, revealing its struggle in accurately learning highly non-linear CEA prices. Secondly, the advanced machine-learning method **AF** outperforms **DP** with $21\% - 43\%$ decrease in MSE. This shows that AF is more effective in learning temporal patterns from long-term sequences than a general LLM, due to AF's time-series modeling architecture and multi-period learning [19].

Thirdly, the **CoT-RF** applies the LLM to refine the AF predictions, significantly outperforming the AF. Compared to AF, CoT-RF gives 9.6% (6.94 vs. 7.68) MSE reduction in HB, 22% (31.72 vs. 40.72) in SZ, and 31% (5.37 vs. 7.80) in GD. Compared to AF, GLU gives 10% (6.88 vs. 7.68) MSE reduction in HB, 20% (32.41 vs. 40.72) in SZ, and 40% (4.71 vs. 7.80) in GD. That is, CoT-RF consistently achieves state-of-the-art performance. These results show that LLMs have the capable in-context learning ability to learn from past AF mistakes and refine AF future predictions.

Finally, our **GLU** method performs closely to CoT-RF in HB and SZ in MSE, achieving 0.4% lower MSE in HB (6.88 vs. 6.94) and 2% higher MSE in SZ (32.41 vs. 31.72). However, GLU significantly outperforms in GD, with 9% less in MSE of CoT-RF (4.71 vs. 5.37). By checking the Day-10 and Day-30 accuracy metrics, the GLU demonstrates superior performance in both short-term prediction, akin to AF, and long-term prediction, akin to CoT-RF. This observation suggests that post-finetuning

Table 3: CEA forecasting results in regions of HB, SZ and GD.

Hubei (HB)

Method	MSE	Accuracy		
		Day-10	Day-20	Day-30
Lasso ([12])	13.83 (664% ↑)	25%	30%	25%
AF ([19])	7.68 (0%)	**64%**	53%	50%
DP ([17])	10.12 (32% ↑)	42%	42%	40%
CoT-RF (ours)	6.94 (9.6% ↓)	63%	**65%**	52%
GLU (ours)	**6.88 (10% ↓)**	61%	61%	**55%**

Shenzhen (SZ)

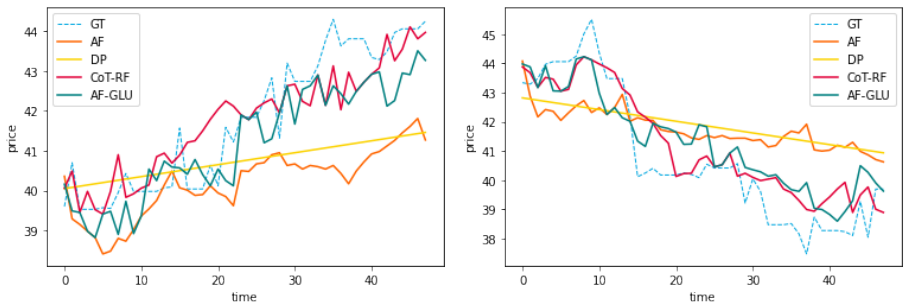
Method	MSE	Accuracy		
		Day-10	Day-20	Day-30
Lasso ([12])	82.35 (183% ↑)	40%	35%	40%
AF ([19])	40.72 (0%)	35%	31%	28%
DP ([17])	71.73 (43% ↑)	21%	22%	21%
CoT-RF (ours)	**31.72 (22% ↓)**	**74%**	**78%**	**85%**
GLU (ours)	32.41 (20% ↓)	68%	71%	82%

Guangdong (GD)

Method	MSE	Accuracy		
		Day-10	Day-20	Day-30
Lasso ([12])	14.03 (81% ↑)	45%	35%	40%
AF ([19])	7.80 (0%)	38%	27%	26%
DP ([17])	9.44 (21% ↑)	31%	22%	25%
CoT-RF (ours)	5.37 (31% ↓)	34%	**65%**	**67%**
GLU (ours)	**4.71 (40% ↓)**	**49%**	62%	**67%**

This table presents the Mean Squared Error (MSE) for each method over the subsequent 30 trading days, along with the Accuracy of 3-way trend classification at Day-10, Day-20, and Day-30. The benchmark model is **AF**, which trains the Autoformer to make future predictions with supervised learning. **Lasso** uses Lasso regression to fit historical data. **DP** relies on the LLM to directly prompt the future price given the history. **CoT-RF** leverages the LLM to learn patterns from a comprehensive AF prediction instance over past steps. It then refines AF predictions based on newly observed historical data. **GLU** performs post-finetuning over AF prediction by using CoT-RF refined results.

effectively inherits the predictive capabilities of both AF and CoT-RF, encompassing both short-term and long-term forecasting capacities.



(a) Rise (b) Fall

Fig. 4: Four methods for CEA price forecasting

We observe the prediction results in Fig. 4. We visualize two randomly selected testing sequences in which (a) exhibits a rising trend and (b) exhibits a declining trend. The CoT-RF (red) and GLU (green) fit the ground-truth (dashed blue) best. The AF (orange) has large deviations in long-term future, while DP (yellow) can only provide a straight line as an estimation of the trend. Therefore, the GLU inherits both the short-term fitting capacity of AF but also the long-term trend prediction capacity of CoT-RF.

5.4 Transfer learning result analysis

We show the transfer-learning results of three paired zones in Table 4. For example, the SZ-GD panel shows results of transferring predictions from SZ (source zone) to GD (target zone) market. For **AF**, we directly use the Autoformer trained on source zone to infer the target zone's future prices, which takes as the baseline. For **DP**, we directly ask the LLM to predict on target zone's future prices, without using information on source zone. **CoT-RF** demonstrates the LLM with AF predictions on source zone and true prices on both source and target zone over aligned past timesteps. This gives contexts of the market differences of source and target zones. Then it asks the LLM to transfer and refine AF predictions on source zone of future timesteps to target zone. **GLU** post-finetunes AF predictions on source zone with transferred predictions of CoT-RF on target zone. This integration enables GLU to simultaneously refine AF predictions and transfer market knowledge.

We observe the following trends. Firstly, **CoT-RF** still consistently performs the best among all methods. In SZ-GD case, where we transfer the market context from Shenzhen to Guangdong, CoT-RF outperforms DP by a 23% (17.63 vs. 27.87) decrease in MSE and 10% higher in movement trend classification accuracy. In SH-HB, we observe a 34% (7.56 vs. 14.53) decrease in MSE, and $43-47\%$ increase in Accuracy. A similar trend has also appeared in HB-SZ case. Secondly, CoT-RF also significantly outperforms the baseline AF. Compared to AF, CoT-RF gives 14%(17.63 vs. 20.41)

Table 4: Transfer-learning of paired CEA regional markets.

Shenzhen → Guangdong (SZ-GD)

Method	MSE	Accuracy		
		Day-10	Day-20	Day-30
Lasso ([12])	94.1 (992% ↑)	10%	15%	15%
AF ([19])	20.41 (0%)	**46%**	40%	24%
DP ([17])	27.87 (37% ↑)	23%	26%	21%
CoT-RF (ours)	**17.63 (14% ↓)**	35%	40%	**40%**
GLU (ours)	18.30 (10% ↓)	33%	**41%**	38%

Shanghai → Hubei (SH-HB)

Method	MSE	Accuracy		
		Day-10	Day-20	Day-30
Lasso ([12])	35.3 (697% ↑)	20%	15%	30%
AF ([19])	9.43 (0%)	57%	57%	61%
DP ([17])	14.53 (54% ↑)	21%	20%	17%
CoT-RF (ours)	7.56 (20% ↓)	64%	64%	64%
GLU (ours)	**7.49 (21% ↓)**	**65%**	**65%**	**67%**

Hubei → Shenzhen (HB-SZ)

Method	MSE	Accuracy		
		Day-10	Day-20	Day-30
Lasso ([12])	181.1 (201% ↑)	10%	5%	20%
AF ([19])	84.73 (0%)	44%	58%	51%
DP ([17])	131.42 (55% ↑)	12%	23%	30%
CoT-RF (ours)	**69.29 (18% ↓)**	**66%**	65%	**59%**
GLU (ours)	69.97 (17% ↓)	65%	**66%**	55%

This table presents the Mean Squared Error (MSE) for each transferring market pair over the subsequent 30 trading days, along with the Accuracy of 3-way trend classification at Day-10, Day-20, and Day-30. The benchmark model is **AF**, which trains the Autoformer on source zone while making future predictions on target zone. **Lasso** fits historical data of source zone while predicting on target zone. **DP** uses the LLM to directly prompt the future price given the history. **CoT-RF** demonstrates the LLM with AF predictions on source zone and true prices on both source and target zones. The LLM learns from the market contexts and refines AF predictions on source zone of future timesteps to target zone. **GLU** post-finetunes AF predictions on source zone with transferred predictions of CoT-RF on target zone.

MSE decrease for SZ-GD, 20% (7.56 vs. 9.43) decrease for SH-HB, and 18% (69.29 vs. 84.73) for HB-SZ.

Furthermore, our **GLU** can also achieve reasonably good results. GLU underperforms the best CoT-RF in 4% in SZ-GD case (18.30 vs. 17.63), while comes very close $\pm 1\%$ in SH-HB and HB-SZ in MSE. The above observation shows that for SZ-GD and SH-HB cases, the source zone actually provides a general trend of the market which can be effectively utilized as extra information to boost target zone prediction.

While for the HB-SZ case, the MSE is as large as over 69, indicating a substantial market difference between these two zones. Indeed, the SZ market exhibits significant daily price fluctuations around 3-10 CNY, whereas the HB market remains notably more stable, with daily fluctuations around just 1 CNY. Despite these disparities, our transferring technique effectively captures the overarching trend, resulting in a respectable trend accuracy of over 60%.

Lastly, the performance of Lasso is highly abnormal with an exceptionally high MSE, which strongly indicates that using Lasso for fitting on one zone and predicting on another zone is inadequate.

6 Conclusion

We introduce a framework that integrates supervised Time-series Modeling (TSM) with LLMs prompting to enhance the accuracy of future forecasts for CEA markets. By presenting TSM predictions alongside relevant market contexts to the LLM through textual prompts, we effectively improve forecasting outcomes by leveraging the LLM's inherent few-shot learning capabilities. Additionally, we exhibit the LLM's capacity to learn from past errors and refine its predictive abilities through self-reflection. Furthermore, we confirm the LLM's capability to extrapolate global market information from a source market to predict future trends in another target market, even in the absence of TSM data. Finally, we innovate a post-finetuning process which distills the refinement capacity of the LLM into a compact GLU model. As a result, we can bypass the LLM prompting phase and effectively mitigate the expenses associated with utilizing commercial LLMs.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Projects 62106156), and the South China Normal University, China. We also thank Tianqi Pang for providing the implementations of Lasso methods on EU ETS forecasting.

References

1. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., et al.: Graph of thoughts: Solving elaborate problems with large language models. arXiv preprint arXiv:2308.09687 (2023)

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
4. Jha, M., Qian, J., Weber, M., Yang, B.: Chatgpt and corporate policies. Tech. rep., National Bureau of Economic Research (2024)
5. Li, Y., Wang, S., Ding, H., Chen, H.: Large language models in finance: A survey. In: Proceedings of the Fourth ACM International Conference on AI in Finance. pp. 374–382 (2023)
6. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021)
7. Madaan, A., Tandon, N., et al.: Self-refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651 (2023)
8. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155 (2022)
9. Pang, T., Tan, K., Fan, C.: Carbon price forecasting with quantile regression and feature selection. In: 2023 7th International Symposium on Computer Science and Intelligent Control (ISCSIC) (2023)
10. Pang, T., Tan, K., Yao, Y., Liu, X., Meng, F., Fan, C., Zhang, X.: Remed: Retrieval-augmented medical document query responding with embedding fine-tuning. IJCNN (2024)
11. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017)
12. Ren, X., Duan, K., Tao, L., Shi, Y., Yan, C.: Carbon prices forecasting in quantiles. Energy Economics **108**, 105862 (2022)
13. Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.t.: Replug: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652 (2023)
14. Shinn, N., Labash, B., Gopinath, A.: Reflexion: an autonomous agent with dynamic memory and self-reflection. arXiv preprint arXiv:2303.11366 (2023)
15. Touvron, H., Lavril, T., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
16. Wang, S., Yuan, H., Zhou, L., Ni, L.M., Shum, H.Y., Guo, J.: Alpha-GPT: Human-AI interactive alpha mining for quantitative investment. arXiv preprint arXiv:2308.00016 (2023)
17. Wei, J., et al.: Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 (2022)
18. Wu, C., Gan, Y., Ge, Y., Lu, Z., Wang, J., Feng, Y., Luo, P., Shan, Y.: Llama pro: Progressive llama with block expansion. arXiv preprint arXiv:2401.02415 (2024)
19. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems **34**, 22419–22430 (2021)
20. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
21. Xie, Q., Han, W., Lai, Y., Peng, M., Huang, J.: The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. arXiv preprint arXiv:2304.05351 (2023)
22. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601 (2023)

23. Zhou, F., Huang, Z., Zhang, C.: Carbon price forecasting based on ceemdan and lstm. Applied Energy **311**, 118601 (2022)
24. Zhu, B.: A novel multiscale ensemble carbon price prediction model integrating empirical mode decomposition, genetic algorithm and artificial neural network. Energies **5**(2), 355–370 (2012)