# Research and Analysis of Large Language Model Synthetic Data Generation and Bias and Illusion Problems

Bolin Zhang

School of Big Data and Intelligent Engineering, Guizhou University of Commerce, Baiyun District, Guiyang City, Guizhou Province, China
zbl15608576692@outlook.com

**Abstract.** This dissertation examines the relevance of large language models (LLMs) in enhancing time-series data applications, e.g., climate forecasting, traffic control, and finance, in which quality data is critically important in the prediction and making of decisions. The importance of overcoming the limitations of LLM in these aspects is to reduce real-life problems, such as data shortage, related privacy issues, and elevated labeling expenses to make the model reliable and allow it to be used in industries where time-related data are needed. This dissertation presents the key problems with big language models (LLM) in time-Series scenarios with a focus on the joint optimization problem that entails synthetic da-ta generation, bias mitigation, and illusion detection. Meeting three key is-sues of the current technologies, such as the lack of domain adaptability, the inability to remove multidimensional biases and detect sophisticated logical illusions, the study offers a technical solution based on the specifics of time-series data. The efficiency of the given solution is confirmed with the help of the real research that is performed on the PyTorch experimental platform based on the data of the Daily Climate Forecasting, traf-fic prediction, and stock sentiment analysis. Lastly, a three phase technical scheme of explicit temporal semantic modeling + cross-view consistency constraints + domain knowledge fusion is built, which offers Conceptual foundation and pragmatic directions to an enhanced trustworthiness of the LLMs in the time-series area.

**Keywords:** Computer vision, Natural language process, Reinforcement learning, Logical illusion recognition

## 1 Introduction

The performance improvement of large language models depends heavily on large-scale high-quality training data, but real time series data (such as traffic flow, climate temperature, stock fluctuations, etc.) face three real challenges. First, the problem of data scarcity is prominent. The missing rate of traffic data in some cities exceeds 30%, which is difficult to meet the training needs of the model. Second, the labeling cost is high. The labeling cost of a single sample stock sentiment analysis is $10-15, and large-scale labeling is not economical. Finally, the privacy risk is significant. User travel time


K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_73

series data involves trajectory privacy, and direct use is likely to cause compliance issues [1]. Although synthetic data generation is an effective way to solve the above problems, the existing methods have obvious defects. They use general text logic to generate time series data and ignore the periodicity, trend and correlation of time series data. For example, the fitting R2 of seasonal trend generated by general prompts is only 0.65-0.75, which is far lower than the 0.85 above that of real data training models [2], and cannot meet the accuracy requirements of time series scenarios. Further research found that LLMs exhibit new bias and illusion features in time series scenarios. The temporal distribution of bias is manifested as follows: if traffic data only covers the morning peak, the average absolute error of the evening peak prediction will increase by 20%-25% [3]. The logical coherence of illusion is manifested as follows: although some generated data (such as "temperature rises and humidity rises simultaneously") are normal in value, they violate the climatological logic [4]. Existing research has two major limitations: first, it separates the generation of synthetic data from the mitigation of bias/illusion, and fails to recognize that the quality of temporal synthetic data directly determines the upper limit of bias and illusion mitigation; second, the contrastive learning adopts a random negative sampling strategy, ignoring the characteristics of "near neighbor similarity and distant neighbor difference" of temporal data, resulting in low sample discrimination and difficulty in the model to learn fine-grained temporal patterns [5]. This research focuses on the core question of how temporal characteristics drive the customization of technical solutions. Through PyTorch experiments, it verifies three key points: First, synthetic data generation needs to embed temporal periodic constraints to improve domain adaptability; second, bias mitigation requires establishing a temporally-based multi-dimensional evaluation system, rather than relying on a single dimension for bias removal; and third, illusion detection needs to be combined with temporal logic verification to identify complex illusions where the numerical values are normal but the logic is abnormal. In the end, everything comes to the same conclusion: "temporal characteristics - technical pain points - customized solutions." This gives a methodical way to use LLMs in temporal situations.

## 2 Research techniques and data: Core research design

### 2.1 Core Technology Framework

This article creates a three-layer technical framework to test the ideas above. Each module is tailored according to temporal attributes, with its primary functionalities executed by PyTorch. The framework has a temporal semantic explicit modeling module that captures the periodic patterns of data, a cross-view consistency constraint module that makes sure multi-dimensional data works together, and a domain knowledge fusion module that adds expert rules to the generation process. These three modules form a progressive relationship, jointly ensuring the systematic nature of the technical solution.

### 2.2 Dataset Structure

The experimental datasets were selected from Kaggle and PyTorch official tutorials [6-8], and the core viewpoints were accurately verified by creating a comparative subset. At the same time, an innovative preprocessing process was designed to solve the problem of traditional methods ignoring the temporal characteristics. The core datasets include DailyClimateForecasting [7], traffic forecasting data [8] and Stock_sentiment_analysis [9], all of which have obvious temporal characteristics. The dataset preprocessing adopts a three-step innovative process: temporal integrity repair, periodic alignment and domain rule filtering. First, the temporal integrity repair method is used, and the PyTorch-LSTM model is used to predict missing values (such as hourly missing values in traffic data), so that the MAE is reduced by 35%-40% compared with the traditional mean filling. Second, periodic alignment is performed, and the data is aligned according to the natural cycle (such as the climate data with "month-day" as the cycle unit), to ensure that the synthetic data conforms to the real cycle pattern. Finally, domain rule filtering is implemented, and invalid samples are removed based on domain logic (such as deleting stock data with "news and stock price fluctuation interval exceeding 24 hours") to improve the data correlation. This approach effectively addresses the issue of conventional preprocessing neglecting temporal attributes.

### 2.3 Evaluation Indicators for Innovation

By transforming traditional evaluation metrics through temporal and logical restructuring, the effectiveness of technical solutions in addressing core issues can be accurately measured. New metrics include temporal fit, a multi-dimensional bias index, and a logical illusion false positive rate. These metrics quantify domain suitability, bias elimination effectiveness, and illusion detection capability. For example, temporal fit combines periodic fit and trend consistency; the multi-dimensional bias index assesses the degree of cross-bias; and the logical illusion false positive rate focuses on identifying cases with normal numerical values but logically abnormal logic.

## 3 Dataset Limitations Analysis: Issues from Temporal Characteristics

### 3.1 Core Obstacles

The limitations of time series datasets are not small in scale or low in quality as traditionally understood, but rather three special problems caused by time series characteristics. First, the problem of balancing the periodicity and randomness of time series leads to insufficient domain adaptability of synthetic data. Existing datasets (such as DailyClimateForecasting) only provide historical periodic data and lack annotations of random disturbances within the period (such as the abnormal temperature drop caused by the January cold wave). This makes it easy for LLMs to fall into two extremes when generating data: over-reliance on periodicity (generated data is highly consistent) or excessive randomness (fluctuations exceed physical laws). Experimental verification shows that without disturbance annotations, the domain rule compliance rate of

synthetic data is only 58%-62%; after supplementing double annotations, the compliance rate rises to 80%-85% [2], proving that time series datasets need to include both periodic and random annotations. Second, the lack of multidimensional time series bias annotations makes it difficult to eliminate multidimensional bias. Traffic prediction datasets usually only annotate the single dimension of "time period" and do not include cross-bias samples of "time period + region" (such as the error difference between urban and suburban areas during the morning rush hour). Traditional debiasing methods only optimize for a single dimension and cannot solve the cross-bias problem. Experiments show that the cross-bias index of the model trained based on single-dimensional annotation is 0.28; after supplementing multi-dimensional annotation, the index drops to 0.12 [3], indicating that a multi-dimensional bias annotation system needs to be constructed. Third, the blank annotation of temporal logic constraints makes it difficult to identify complex illusions. The Stock_sentiment_analysis dataset only labels the correspondence between "news sentiment - stock price rise and fall", without labeling the time lag logic (such as news needing to affect the stock price within 12 hours). This makes LLMs easy to generate logical illusions (such as "news will affect the stock price 3 days later"), while traditional detection methods cannot identify them due to the lack of logical annotation. Experiments show that without logical annotation, the false negative rate of complex illusions reaches 25%; after supplementing domain temporal logic annotation, the false negative rate drops to 8% [4], confirming that domain logical constraint annotation needs to be supplemented.

### 3.2 Targeted solutions

In response to the above limitations, a customized strategy based on PyTorch is proposed. For the problem of balancing periodicity and randomness, a perturbation annotation mechanism is introduced; for the problem of multi-dimensional bias, a bias annotation matrix is constructed; for the problem of logical illusion, temporal logic rules are added. Experiments show that these methods are significantly more effective than traditional methods.

## 4 Model Limitations Analysis: Technical Adaptation Deficits from Temporal Characteristics

### 4.1 Key Technical Issues

Based on the PyTorch reproduction experiment, it was found that the existing model cannot solve the core problem. The root cause is that it is not adapted to the time series characteristics, which is manifested in three types of technical defects. The existing model cannot solve the core problem. The root cause is that it is not adapted to the time series characteristics. First, the synthetic data generation model has the problem of shallow time series dependency modeling. Traditional multi-step generation only uses time step splitting (such as 1-7 days → 8-14 days), without taking into account long-term dependency and short-term correlation (such as climate data needing to simultaneously imitate annual seasonal trends and intra-week temperature fluctuations). In addition, the random negative sampling of contrastive learning

(selecting samples from different years) cannot reflect the similarity characteristics of temporal neighbors, and the sample discrimination is low. Experimental verification shows that in the 30-day climate data generation task, the dynamic time warping value of the traditional model is 15.2; after introducing long and short-term time series attention (bidirectional LSTM), DTW drops to 8.6. At the same time, changing random negative sampling to temporal neighbor negative sampling (interval of 1-3 days) improves the trend fitting R2 by 0.13 [5]. Second, the bias mitigation model has a conflict between fairness and performance gradient. Traditional adversarial training leads to a decrease in off-peak performance (i.e., gradient conflict) when optimizing the fairness of morning peak prediction. Experiments show that traditional adversarial training reduces the overall MAE of traffic prediction from 16.2 to 15.8 (improved fairness), but the off-peak MAE increases from 14.5 to 17.3 (decreased performance). After introducing gradient balance contrastive learning with temporal weights, the bias index drops to 0.12, and the off-peak MAE is only 15.1 [3]. Third, the hallucination detection model lacks a logic verification module. Traditional detection only identifies anomalies based on numerical errors (such as MAE), and cannot detect complex hallucinations with normal numerical values but abnormal logic (such as simultaneous increase in temperature and humidity). The experiment constructed a test set of 100 logical hallucinations, and the traditional method only identified 32 (68% false negative rate); after introducing the temporal logic verification module (rule matching + causal inference), the number of identified hallucinations increased to 89 (11% false negative rate) [4].

### 4.2 Key Methods for Performance Improvement

To address the aforementioned shortcomings, three PyTorch optimization schemes are proposed. First, temporal multi-scale attention is used to simultaneously model short-term fluctuations (1-7 days) and long-term trends (seasonal), improving $R^2$ by 0.16. Second, adaptive gradient weight adjustment is employed, introducing temporal fairness weights into the loss function (higher weights for periods with larger errors), dynamically balancing gradient directions, and reducing the conflict between fairness and performance by 40%-50%. Third, a dual-track numerical-logic detection system is used. This method combines a two-branch network of DTW numerical detection and logical verification. This lowers the false negative rate of complicated illusions from 25% to 8%. Experiments show that these strategies work very well.

## 5 Experimental Results and Performance Evaluation: Validating the Solution

### 5.1 Experimental Design

This work utilized a control experiment to compare the conventional technique group with the enhanced method group, and conducted three fundamental validations using PyTorch. The experimental design focused on managing variables to guarantee the comparability of results. It featured studies on generating synthetic data, reducing bias, and finding hallucinations, all utilizing the same dataset and evaluation measures.

### 5.2 Core Experimental Results

First, synthetic data is far better at adapting to different domains. The enhanced group does better than the traditional group on the DailyClimateForecasting dataset. The domain rule compliance rate goes up from 58% to 85%, the trend fitting R2 goes up from 0.72 to 0.88, and the downstream climate forecast MAE goes down from 1.5℃ to 0.9℃.This shows that the temporal rule constraint and the nearest neighbor negative sampling are effective—the temporal rule guarantees the periodic characteristics, and the nearest neighbor negative sampling helps the model learn fine-grained temporal differences [2]. Second, the multidimensional bias mitigation effect is doubled. In the time-region dual-dimensional traffic prediction task, the improved group significantly optimizes the bias problem: the multidimensional bias index decreases from 0.28 to 0.12; the temporal fairness score increases from 0.65 to 0.83; and the difference in MAE between urban and suburban areas during the morning peak decreases from 5.2 to 2.1. This proves that multidimensional resampling and gradient balancing are effective—resampling solves the data distribution imbalance, and gradient balancing avoids the conflict between fairness and performance [3]. Third, the false negative rate of complex logic illusion is significantly reduced. On the Stock_sentiment_analysis logic illusion test set, the detection capability of the improved group was improved: the false negative rate of complex illusions decreased from 25% to 8%; the logic violation rate decreased from 18% to 5%; and the detection accuracy of the illusion of "news will affect stock price 3 days later" increased from 32% to 89%. This shows that the numerical-logic dual-track detection is effective and the sequential logic check fills the gap of single numerical detection [4].

### 5.3 Analysis of Research Results: Customization of Timing Characteristics is the Core

The core conclusion drawn from the comprehensive experimental results is that the key to LLM synthetic data generation and bias/hallucination mitigation lies in the customization of temporal characteristics. General solutions that neglect temporal characteristics show a significant reduction in effectiveness: if synthetic data generation does not embed temporal rules, domain adaptation accuracy decreases by 15%-20%; if bias mitigation does not establish a multi-dimensional temporal system, the cross-bias mitigation effect is only 30%-40%; and if hallucination detection is not combined with temporal logic verification, the false negative rate for complex hallucinations is as high as 25%. This conclusion emphasizes the importance of temporal customization.

## 6 Feasibility analysis of the improvement plan: Time-series technology method

### 6.1 Core Improvement Plan

Based on the experimental results, three feasible time series processing solutions are proposed, and their feasibility is verified by PyTorch. First, time series hard negative sampling solves the problem of insufficient fine-grained features in synthetic data.

Traditional random negative sampling cannot enable the model to learn the fine-grained differences in time series, resulting in synthetic data lacking domain specificity. The solution is based on the InfoNCE loss function of PyTorch to generate nearest neighbor hard negative samples - those that are 1-3 days apart from positive samples, have similar trends but have micro-anomalies (such as daily temperature fluctuations exceeding 2℃), forcing the model to distinguish subtle time series differences. Feasibility verification shows that on the traffic prediction dataset, DTW decreased from 15.2 to 8.6; the computational cost increased by only 8%, and the training time increased from 2.5 hours to 2.7 hours, which meets the requirements of industry [10]. Second, gradient balanced time series contrastive learning solves the problem of fairness and performance conflict. Traditional adversarial training leads to fairness and performance conflict in time series data. The solution is based on the PyTorch Unicorn framework [11], adjusts the temporal contrast loss function, introduces time period weight coefficients (such as 0.4 for morning peak and 0.1 for off-peak), and dynamically balances the gradient direction. Feasibility verification shows that on the traffic prediction dataset, the bias index decreased from 0.28 to 0.12, and the MAE only increased from 16.2 to 16.5 (performance loss <2%), which is better than the traditional method. Third, hierarchical temporal attention solves the problem of inconsistency between domain knowledge and temporal features. Traditional feature fusion cannot effectively integrate temporal features and domain knowledge. The solution is based on the hierarchical attention mechanism of PyTorch, with the bottom layer capturing temporal features (such as traffic hourly fluctuations), the middle layer fusing domain knowledge (such as road topology), and the top layer outputting cross-modal representation. Feasibility verification shows that on the DailyClimateForecasting dataset, the domain rule compliance rate increased from 58% to 85%, and the downstream prediction MAPE decreased from 8.5% to 4.2% [12], proving that knowledge and temporal features are effectively fused.

### 6.2 Research Conclusions: A Three-Stage Collaborative Framework

A three-stage collaborative framework for temporal synthetic data generation, bias mitigation, and hallucination detection is proposed to achieve end-to-end optimization. In the generation stage, temporal rules and difficult negative sampling reduce bias and hallucinations by 40%-50% from the data source. . In the detection stage, hierarchical attention is employed to merge domain knowledge, boosting the accuracy of logical verification and controlling the false negative rate of complex hallucinations to 8% to 10%. Full-process experiments demonstrate that this approach enhances the factual correctness of synthetic data by 18%-22% and eradicates multidimensional bias by 65%-70%, facilitating joint improvement of data, model, and detection.

## 7 Conclusion

This research methodically investigates the principal challenges of collaborative optimization of extensive language models influenced by temporal attributes. The study first finds the primary problems that LLMs have in the time domain. Some of these problems are that the model doesn't fit well, it's hard to get rid of bias, and it's hard to

spot hallucinations. The main problem is that current technical solutions don't take into account the unique features of temporal data. Some of these traits are periodicity, correlation, and logicality. General approaches can't meet the unique needs of temporal circumstances, which limits performance.

Second, the study suggests a completely tailored technical solution depending on time-related factors. During the synthetic data generation phase, the solution incorporates temporal criteria and employs a nearest-neighbor negative sampling approach. A multidimensional resampling method and a gradient balancing technique are used in the bias mitigation stage. A dual-track verification method that combines numerical and logical verification is used in the hallucination detection step. These methods use advanced concepts like cross-modal alignment and retrieval-enhanced generation.

Finally, the three-stage approach illustrates that the strategy greatly improves key indicators through experimental validation. Synthetic data's factual correctness goes up by 18% to 22%. The effect of multidimensional bias on removal is between 65% and 70%. The false negative rate for complex hallucinations drops to between 8% and 10%. The framework shows that it works well on a variety of real-world datasets.

Future research should prioritize the cross-domain transfer of temporal information. In particular, established temporal principles from the climate sector must be adapted to new contexts, including transportation and finance. This transfer can lower the expenses of adapting to different areas and keep developers from doing the same work twice. It is thought that transferring information across domains will make models better at generalizing.

Another key area of research is learning dynamic temporal logic. The goal is to let LLMs understand the domain logic that changes over time on their own. For instance, how changes in policy affect traffic peaks. This strategy can make the model more adaptable and less dependent on manual annotation. Dynamic learning algorithms can better capture how temporal data changes over time.

Also, we need to make progress on low-resource temporal data adaptation strategies right away. Research should concentrate on the integration of high-quality data derived from a limited number of samples. At the same time, we need to come up with effective ways to reduce bias. This is very important for specialist uses, such some urban transportation systems or predicting stock prices in certain fields. Low-resource adaption approaches can make models work in more situations.

## References

1. Lin, L., Wang, R., Xiao, R.: A Surve on LLM-Driven Synthetic Data Generation, Debiasing, and Evaluation. arXiv preprint arXiv:2406.15126, (2024)
2. Wang, K., Liu, Z., Zhang, C.: A Survey on Data Synthesis and Augmentation for Large Language Models. arXiv preprint arXiv:2410.12896, (2024)
3. Lin, Z., Guan, S., Zhang, W.: A Review of Debiasing and Dehallucinating in Large Language Models: Toward Credible LLMs. Artificial Intelligence Review, 57(9):1-50. (2024)

4. Guo, Y., Guo, M., Su, J., et al.: Bias in Large Language Models: Origin, Evaluation, and Mitigation. arXiv preprint arXiv:2411.10915, (2024)
5. Chen, T., Kornblith, S., Norouzi, M., et al. A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th International Conference on Machine Learning, 1597-1607. (2020)
6. PyTorch Tutorials. Time Series Prediction with LSTM[EB/OL]. https://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html, (2024)
7. Kaggle. Daily Climate Forecasting with PyTorch-LSTM [EB/OL]. https://www.kaggle.com/code/wonduk/pytorch-lstm-daily-climate-forecasting, (2024)
8. Kaggle. Traffic Prediction Ensemble [EB/OL]. https://www.kaggle.com/code/guanlintao/100-ensemble-traffic-prediction-dataset (2024)
9. Kaggle. Stock Sentiment Analysis [EB/OL]. https://www.kaggle.com/code/niharikaamritkar/stock-sentiment-analysis, (2024)
10. Li, Z., Wang, Y., Liu, H., et al.: UniCorn: An Integrated Contrastive Learning Approach for Multi-View Molecular Representation. Proceedings of the 37th Neural Information Processing Systems Conference (NeurIPS 2023), (2023)
11. Bian, J., Lu, H., Dong, G., et al. Hierarchical Multimodal Self-Attention-Based Graph Neural Network for DTI Prediction. Journal of Bioinformatics Letters, 25(4): bbae293. (2024)
12. Li, J., Cheng, X., Zhao, W. X., et al. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. arXiv preprint arXiv:2305.14251, (2023)