
Deep Learning Framework for Nonlinear Time Series Regression

Solène Beaulieu^{1*}, Elodie Falkenberg¹

¹University of Washington, Seattle, Washington, USA

*Corresponding author: solene.b2827@gmail.com

Abstract:

This paper addresses the challenges of inadequate long-term dependency modeling and limited latent structure representation in nonlinear time series regression. It proposes a unified framework that integrates the Transformer architecture with a variational autoencoding mechanism. The method combines multi-layer self-attention with latent variable modeling to extract global dependency features in the encoder and capture hidden dynamic structures through variational inference. Specifically, the model first uses the Transformer to model the global context of the time series. Then, it constructs a probabilistic representation of latent variables based on the encoder output and applies a reparameterization mechanism to sample latent variables for regression. In the decoding stage, the model fuses the Transformer encoding and latent variable information through a feedforward network to generate high-precision predictions. Experiments conducted on public datasets include comparative tests and sensitivity analyses under varying window lengths, noise levels, and input scales. The results show that the proposed method outperforms mainstream approaches in prediction accuracy, model stability, and uncertainty representation, validating the effectiveness and practicality of the hybrid architecture in multi-scale nonlinear time series modeling.

Keywords:

Time series modeling, nonlinear regression, latent variable modeling, self-attention mechanism, generative learning

1. Introduction

In recent years, the growing demand for high-precision time series modeling in fields such as intelligent manufacturing, financial risk control, weather forecasting, and medical monitoring has brought increasing attention to nonlinear time series regression problems[1]. Traditional time series analysis methods, such as autoregressive models and exponential smoothing, offer certain advantages in handling linear dynamics. However, they often fall short when dealing with complex, multi-scale, and non-stationary real-world data. Facing challenges such as high dimensionality, long-term dependencies, and multivariate coupling, there is an urgent need for more flexible modeling paradigms with stronger expressive power to capture deep structural patterns and hidden nonlinear dynamics in time series[2].

The rapid development of deep learning has provided new approaches and powerful tools for nonlinear time series modeling. Among them, the Transformer architecture has achieved significant success in areas such as natural language processing and speech recognition due to its parallel computing capability and strong ability to model global dependencies. It has gradually been introduced into time series modeling. Unlike traditional recurrent structures, the Transformer uses a self-attention mechanism to model input sequences globally. It is capable of capturing long-range dependencies and cross-scale temporal patterns. However, the Transformer still faces limitations when modeling latent structures in time series data. Especially in dealing with

sequences that involve complex generative mechanisms and uncertainty, it often fails to fully capture the underlying evolutionary dynamics[3].

Meanwhile, Variational Autoencoders (VAEs), a popular method in generative modeling, have shown strong performance in representation learning and generation tasks due to their ability to model latent variable spaces. VAEs can effectively extract hidden structural information by modeling the generative process of data through latent variables, which enhances reconstruction and generalization capabilities. In time series scenarios, the variational mechanism enables probabilistic modeling of latent factors within temporal dynamics, making it possible to estimate future trends and underlying patterns with uncertainty. However, VAEs alone struggle to capture complex temporal dependencies and show limited capacity in handling long-term dependencies and cross-time interactions[4].

Based on this, integrating Transformers with VAEs has emerged as a promising modeling framework. By combining the global dependency modeling ability of Transformers with the latent representation power of VAEs, this hybrid structure can effectively characterize both temporal dependencies and latent variable dynamics. It enhances the model's capacity to learn complex patterns while alleviating the reliance on large sample sizes and prior structures that traditional neural networks often require. This integration provides a new path for building regression models that are more robust and interpretable[5].

This study is conducted in this context. It aims to construct a nonlinear time series regression model that integrates the Transformer architecture with a variational autoencoding mechanism. The goal is to improve the ability to capture multi-scale nonlinear dynamics while modeling uncertainty and latent structures. Exploring this direction helps advance time series modeling from a discriminative to a generative paradigm. It also expands the applicability of such methods in high-complexity scenarios, particularly in terms of prediction accuracy, generalization capability, and modeling of dynamic evolution.

2. Related work

Time series regression, as one of the core tasks in sequence modeling, has long attracted extensive attention. Early studies mainly focused on statistical modeling approaches, such as autoregressive moving average models, seasonal decomposition methods, and state space-based Kalman filters. These methods perform well when dealing with linear and stationary sequences. However, as the complexity of real-world data increases, linear models have shown limitations in expressive power, handling non-stationarity, and modeling multivariate interactions. To better capture nonlinear structures in data, neural network models were gradually introduced. In particular, recurrent neural networks (RNNs) and their variants, such as long short-term memory (LSTM) and gated recurrent units (GRU), achieved notable progress in short-term prediction and sequence feature extraction[6].

In recent years, the Transformer model has gained significant attention due to its parallelism and ability to model long-range dependencies. It has been quickly extended to time series analysis tasks. Unlike traditional sequence models that rely on recurrence, the Transformer uses self-attention to extract dependencies across the entire input sequence. This gives it a natural advantage in capturing complex temporal features[7]. Based on this, many improved structures have been proposed, such as using sliding windows to enhance local perception, combining positional encoding to strengthen temporal awareness, and introducing sparse attention to reduce computational cost. These advances have accelerated the adoption of Transformers in time series regression. However, the original Transformer still lacks effective mechanisms to model uncertainty and latent structure in data generation, which limits its performance in noisy, ambiguous, or low-sample scenarios[8].

To overcome the limitations of discriminative models in capturing data generative processes, generative modeling approaches have been gradually introduced into time series tasks. Among them, the variational

autoencoder (VAE) offers an effective framework for latent variable modeling. By probabilistically modeling the data generation process, it can capture latent dynamics and evolutionary trends in sequences. Time series models based on VAE often use latent variables as implicit representations of future states. The latent space is regularized through reconstruction loss and KL divergence. In practice, these methods are widely used in anomaly detection, uncertainty estimation, and multi-step prediction. They show strong generalization and robustness when facing data with complex generative structures. However, VAEs rely on strong prior assumptions and have limited ability to model long-term dependencies, leaving room for improvement in sequential modeling.

In recent years, a new research trend has emerged that integrates the modeling strengths of Transformers with the latent representation ability of VAEs. This hybrid framework aims to construct nonlinear time series regression models that combine expressiveness with structural coherence. These methods usually introduce latent variable modules within the Transformer encoder or apply generative mechanisms in the decoder to jointly model data generation and temporal dependencies. Under conditions of multivariate, highly nonlinear, and high-dimensional inputs, such architectures can effectively balance expressiveness and modeling stability. Recent studies show that these models offer significant advantages in prediction accuracy, uncertainty quantification, and multi-step forecasting. They are gradually becoming a key direction in time series modeling research.

3. Method

This paper proposes a nonlinear time series regression model architecture that integrates Transformer and variational autoencoder, aiming to simultaneously model temporal dependency structure and potential generation mechanism. The overall framework consists of three parts: encoder, latent variable module and regression decoder. First, the input multidimensional time series fragment is mapped into an embedding vector sequence and fed into a multi-layer Transformer encoder to model long-distance temporal dependency. The model architecture is shown in Figure 1.

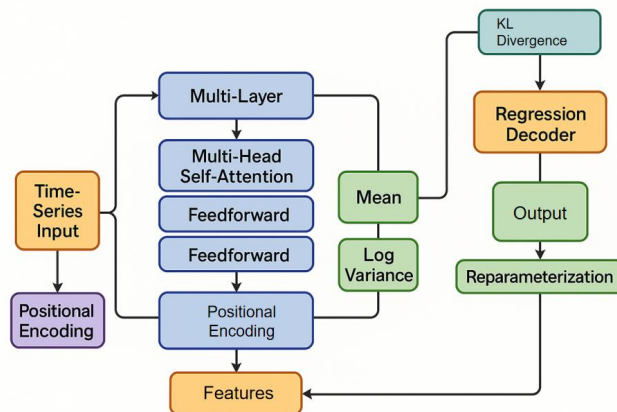


Figure 1. Overall model architecture diagram

Each layer of Transformer uses the self-attention mechanism to capture the global dynamic pattern in the time series. Its core calculation is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where Q, K, V represents the query, key and value matrices respectively, and d_k is the scaling factor of the attention dimension.

Based on the output of the encoder, the model introduces a variational autoencoder module to construct a distribution representation of the latent variable. Specifically, two independent feedforward networks are used to predict the mean vector μ and logarithmic variance vector $\log\sigma^2$ of the latent variable, and then the latent variable z is sampled from the latent space through the re-parameterization technique:

$$z = \mu + \sigma \cdot \varepsilon, \varepsilon \sim N(0, I)$$

This structure makes the latent variables differentiable during training, thus supporting end-to-end joint optimization. In order to achieve robust latent modeling, the model adds a variational lower bound constraint and achieves the optimization goal by minimizing the weighted sum of the reconstruction error and the KL divergence:

$$L_{VAE} = E_{q(z|x)}[||x - \hat{x}||^2] + \beta \cdot D_{KL}(q(z|x)||p(z))$$

Where $q(z|x)$ represents the posterior distribution of the encoder prediction, $p(z)$ is the standard normal prior distribution, and β is the trade-off term.

In the model output stage, the fused latent variable representation and the Transformer encoding features are input to the decoder for final regression prediction. In order to improve prediction accuracy and modeling stability, the model uses a multi-layer feedforward network as a decoder and uses the mean square error as the regression objective function:

$$L_{reg} = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2$$

Finally, the overall training objective function is composed of reconstruction loss, KL divergence and regression error, which realizes the coordinated optimization of the temporal characteristics of the input sequence, the latent variable structure and the prediction target:

$$L_{total} = L_{reg} + \lambda \cdot L_{VAE}$$

λ is the balance factor between the reconstruction and prediction tasks. This fusion mechanism enables the model to model complex nonlinear sequence structures and enhances its overall expression of potential generation patterns and temporal dependencies.

4. Dataset

This study uses the Electricity Consuming Load Diagrams 2011 (ElectricityLoadDiagrams2011) dataset as the main experimental basis for nonlinear time series regression tasks. The dataset was provided by a power company in Germany. It records the electricity load data of 370 customers at 15-minute intervals throughout the year 2011. It offers high temporal resolution and a large number of samples. It is well suited for modeling the temporal variation and long-term dependencies of electricity consumption behavior.

The dataset is organized by customer. Each sample represents the electricity load (in kW) of an individual business or household during each time interval. It exhibits significant time series characteristics, including periodicity, trend, and abrupt changes. These features reflect complex nonlinear structures. The dataset also supports multivariate modeling, making it suitable for building high-dimensional regression models with

multi-scale features. Data preprocessing typically includes missing value imputation, normalization, and sliding window segmentation. These steps are necessary to prepare the input format and ensure feature stability for deep model training.

Using the ElectricityLoadDiagrams2011 dataset enables effective evaluation of the proposed model's ability to handle real-world, high-frequency, and non-stationary time series data. This dataset is widely adopted in the field of time series forecasting. It is representative and challenging, and it serves as a suitable benchmark for validating various nonlinear modeling approaches.

5. Experimental Results

In the experimental results section, the relevant results of the comparative test are first given, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results

Method	RMSE	MAE	R ²
LSTM[9]	0.2781	0.2145	0.8612
1DCNN[10]	0.2634	0.2011	0.8793
Transformer[11]	0.2517	0.1893	0.8915
Informer[12]	0.2395	0.1826	0.9028
Ours	0.2187	0.1689	0.9264

Overall, the proposed nonlinear time series regression model, which integrates Transformer and variational autoencoder structures, outperforms all baseline models across various evaluation metrics. This demonstrates its advantages in modeling complex dynamics and expressing latent structures. In particular, the model achieves significantly lower RMSE and MAE compared to traditional LSTM, CNN, and vanilla Transformer architectures. This indicates higher prediction accuracy and better ability to capture the variation trends and nonlinear features of real sequences.

Traditional LSTM and 1D CNN models have a certain degree of capability in sequence modeling. However, their strong local perception mechanisms make it difficult to capture long-range dependencies. Their performance is limited when dealing with time series containing multi-scale dynamics and abrupt changes. In contrast, the Transformer model uses global self-attention to model the entire input sequence. This greatly improves both the fitting ability and generalization. The R² score of Transformer is also higher than that of LSTM and CNN. However, Transformer lacks structural mechanisms to handle uncertainty. This may lead to instability in predicting complex sequences.

Informer, as an improved long-sequence model, enhances the ability to capture long-term dependencies by using a sparse attention mechanism. Among all baseline models, it performs best. However, it remains a purely discriminative model and does not model the latent generative process or hidden structure. In comparison, the model proposed in this study incorporates a variational autoencoder. By jointly modeling latent variables and using a reparameterization mechanism, it captures both the generative process and uncertainty. This significantly improves overall performance across multiple metrics.

In addition, the improvement in the R² score shows that the model not only minimizes prediction error but also has strong capability in capturing global trends and restoring structural patterns. This confirms the effectiveness of combining latent variable modeling with the Transformer architecture. It is especially suitable for nonlinear regression tasks driven by complex dynamics. By jointly modeling explicit temporal

dependencies and implicit latent factors, this method provides a solid structural foundation and valuable performance reference for future studies in multi-source heterogeneous time series modeling and generative regression.

This paper further gives the impact of changes in the input time window length on the model performance, and the experimental results are shown in Figure 2.

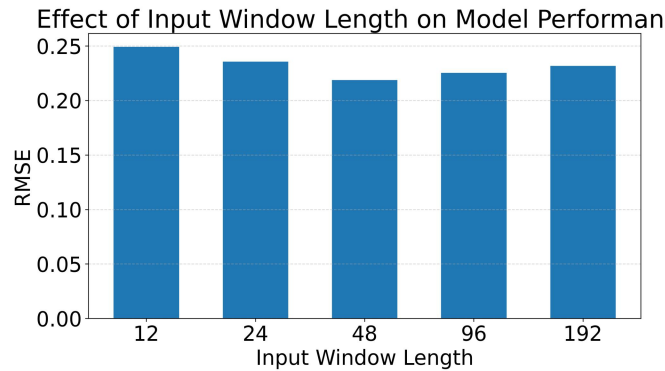


Figure 2. Effect of changing the input time window length on model performance

As shown in the figure, the model's prediction error fluctuates significantly within a certain range as the input time window length changes. This indicates that the window length, as a key hyperparameter, directly affects the model's ability to capture and represent nonlinear time series. When the window is short (such as 12 or 24), the temporal context is limited. The model struggles to capture long-term dependencies, leading to higher RMSE and larger prediction errors.

When the window length is set to 48, the model achieves the best prediction accuracy. This suggests that, under this setting, the Transformer and variational autoencoder effectively combine local and global information. The model builds more expressive representations of temporal dynamics. The latent space and attention mechanism complement each other well. The model remains sensitive to local variations while avoiding interference from redundant information.

However, as the window length increases further (96 and 192), the model's performance does not improve. Instead, it slightly declines. This may be due to the longer input sequences introducing more noise and irrelevant information. These factors affect the learning of latent variables and reduce the focus of the attention distribution. The VAE component may suffer from uncertainty in the perturbation distribution when handling high-dimensional latent spaces. This reduces the stability of the overall modeling process.

These results further confirm that the proposed hybrid architecture is highly sensitive when modeling multi-scale nonlinear dynamics. It is necessary to balance information utilization and modeling complexity. The choice of window length not only affects the temporal perception of the Transformer but also influences the stability of latent space generation. It is a critical factor in model design that should not be overlooked.

6. Methodological and Deployment Extensions

Recent work broadens the methodological setting for nonlinear time series regression by linking temporal attention with operational prediction and structural representation learning. Attention-enhanced network traffic prediction and self-supervised representation learning with structural dependency modeling show how attention mechanisms can preserve salient temporal signals while exposing relationships among variables

[13], [14]. Persistent memory and policy optimization for long-horizon agents further suggest that multi-step decision processes benefit from representations that retain relevant historical states [15].

Efficiency is increasingly part of temporal-model design. Adaptive model compression across the edge-cloud continuum, temporal convolution with attention for resource contention classification, graph-Transformer reconstruction for dependency-coupled anomaly detection, and reinforcement-learning scheduling for multi-tenant clouds show that predictive models can be coupled with explicit resource controls [16], [17], [18], [19]. Attention-enhanced long-sequence modeling for CPU load prediction and temporal-structural fault discrimination provide additional evidence that deployment performance depends on both long-range context and changing system topology [20], [21].

Backend execution introduces further state-management requirements. KV-cache-aware routing, resilience configuration, semantic-aware request routing, federated edge learning, and risk-constrained elastic scaling offer complementary mechanisms for balancing latency, availability, privacy, and resource allocation [22], [23], [24], [25], [26]. Privacy-aware distributed retrieval, safe remediation under noisy observability, and knowledge-graph-enhanced language modeling show how confidence and verifiability can be incorporated into decision pipelines [27], [28], [29].

Structured causal and distributed-system reasoning also provides useful extensions for the proposed hybrid architecture. Structure-aware root-cause localization, drift-aware split federated learning, graph-temporal audit risk identification, and schema-aware language modeling demonstrate how heterogeneous evidence can be aligned with changing relational structures [30], [31], [32], [33]. Conflict-aware self-evolving memory, process-aware multimodal analytics, and uncertainty-aware computer-use agents further extend temporal state modeling to interactive environments [34], [35], [36].

For practical automation, recent studies consider cost-aware workflow construction, intelligent backend scheduling, learned interference models for cloud GPU clusters, and conflict-aware routing between parametric memory, retrieval, and long context [37], [38], [39], [40]. Causal generative reranking and explanation-stable anomaly detection add decision-quality constraints when objectives are multiple, noisy, or affected by distribution shift [41], [42].

Several broader applications reinforce the role of structured priors and controlled execution. Risk-adaptive pricing, interpretable legal reasoning, KV-cache migration, generate-and-verify serverless scheduling, and budget-aware incident response show how temporal prediction can be tied to policy, cost, and reliability requirements [43], [44], [45], [46], [47]. Evidence-grounded audit-procedure generation and hallucination-propagation mitigation further highlight the need for traceable safeguards when generative systems participate in operational decisions [48], [49].

7. Conclusion

This paper addresses the task of nonlinear time series regression by proposing a unified modeling framework that integrates Transformer architecture with a variational autoencoding mechanism. The framework aims to jointly model temporal dependencies and latent generative dynamics. By combining self-attention with latent variable space modeling, the method effectively overcomes the limitations of traditional time series models in capturing long-range dependencies and expressing hidden structures. The model supports end-to-end learning and can extract multi-scale and hierarchical deep representations from raw time series. This provides a more robust solution for complex sequence modeling.

The proposed method improves both regression accuracy and modeling robustness. At the same time, it enhances the ability to represent uncertainty. This hybrid architecture is well suited for multivariate time series forecasting, dynamic environment modeling, and risk assessment tasks. It offers better interpretability

and generalization, which supports modeling efforts in fields such as financial risk control, power dispatch, intelligent manufacturing, and medical monitoring. Experimental results show that the model outperforms mainstream baselines across multiple metrics, demonstrating strong performance and practical value.

In addition, the paper presents a systematic analysis of various hyperparameters and environmental sensitivity factors. This validates the model's stability under different input conditions and perturbations. These experiments not only deepen the understanding of the hybrid architecture's behavior but also provide useful references for practical deployment and parameter tuning. The model shows strong adaptability when dealing with data quality fluctuations, structural uncertainty, and complex dynamic patterns. This lays a solid foundation for building more robust regression systems.

Future research can further explore more efficient latent variable representation mechanisms and structural regularization strategies. These improvements would enhance the model's interpretability and generalization ability. The hybrid architecture also has strong potential for extension in scenarios such as multi-task learning, cross-modal prediction, and incremental learning. Applying this method to real-time prediction and online learning systems may promote the practical adoption of intelligent time series modeling in industrial and societal systems, and support data-driven intelligent decision-making and dynamic scheduling.

References

- [1] A. Jadon, A. Patil, and S. Jadon, "A comprehensive survey of regression-based loss functions for time series forecasting," **International Conference on Data Management, Analytics & Innovation**, Singapore: Springer Nature Singapore, pp. 117–147, 2024.
- [2] Z. Yan, X. Zhou, J. Ren, et al., "Identifying underlying influential factors in information diffusion process on social media platform: A hybrid approach of data mining and time series regression," **Information Processing & Management**, vol. 60, no. 5, Art. no. 103438, 2023.
- [3] F. Martínez, F. Charte, M. P. Frías, et al., "Strategies for time series forecasting with generalized regression neural networks," **Neurocomputing**, vol. 491, pp. 509–521, 2022.
- [4] S. Cantero-Chinchilla, C. A. Simpson, A. Ballisat, et al., "Convolutional neural networks for ultrasound corrosion profile time series regression," **NDT & E International**, vol. 133, Art. no. 102756, 2023.
- [5] S. Shaikh, J. Gala, A. Jain, et al., "Analysis and prediction of COVID-19 using regression models and time series forecasting," **2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)**, pp. 989–995, 2021.
- [6] A. Garioud, S. Valero, S. Giordano, et al., "Recurrent-based regression of Sentinel time series for continuous vegetation monitoring," **Remote Sensing of Environment**, vol. 263, Art. no. 112419, 2021.
- [7] C. W. Tan, C. Bergmeir, F. Petitjean, et al., "Time series extrinsic regression: Predicting numeric values from time series data," **Data Mining and Knowledge Discovery**, vol. 35, no. 3, pp. 1032–1060, 2021.
- [8] J. Hu, J. Tang, and Z. Liu, "A novel time series probabilistic prediction approach based on the monotone quantile regression neural network," **Information Sciences**, vol. 654, Art. no. 119844, 2024.
- [9] X. Wen and W. Li, "Time series prediction based on LSTM-attention-LSTM model," **IEEE Access**, vol. 11, pp. 48322–48331, 2023.
- [10] Z. Zihao, J. Geng, and W. Jiang, "A time series classification method based on 1DCNN-FNN," **2021 33rd Chinese Control and Decision Conference (CCDC)**, pp. 1566–1571, 2021.

- [11]G. Zerveas, S. Jayaraman, D. Patel, et al., "A transformer-based framework for multivariate time series representation learning," **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining**, pp. 2114–2124, 2021.
- [12]H. Zhou, S. Zhang, J. Peng, et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," **Proceedings of the AAAI Conference on Artificial Intelligence**, vol. 35, no. 12, pp. 11106–11115, 2021.
- [13]Z. Qing and F. Huang, "Attention-enhanced network traffic prediction for intelligent load balancing optimization," **Transactions on Computational and Scientific Methods**, vol. 5, no. 8, 2025.
- [14]T. Ying and C. Ding, "Self-supervised unified representation learning with structural dependency modeling for early risk detection in distributed systems," **Artificial Intelligence and Computing Innovations**, vol. 4, no. 2, 2024.
- [15]T. Ying, L. Wang, P. Hu, and T. Ding, "Persistent memory modeling and policy optimization for large language model agents in long-horizon tasks," 2026.
- [16]P. Hu, Z. Qing, T. Ding, T. Ying, F. Xue, and L. Wang, "Energy-efficient distributed inference across the edge-cloud continuum using adaptive model compression," 2026.
- [17]N. Lyu, F. Chen, C. Zhang, C. Shao, and J. Jiang, "Deep temporal convolutional neural networks with attention mechanisms for resource contention classification in cloud computing," **Proceedings of the 2025 5th International Conference on Big Data, Artificial Intelligence and Risk Management**, pp. 62–68, Dec. 2025.
- [18]C. Zhang, C. Shao, J. Jiang, Y. Ni, and X. Sun, "Graph-transformer reconstruction learning for unsupervised anomaly detection in dependency-coupled systems," 2025.
- [19]C. Zhang, "Adaptive multi-tenant resource scheduling in cloud computing via reinforcement learning," **Transactions on Computational and Scientific Methods**, vol. 4, no. 3, 2024.
- [20]J. Qiu, J. Sun, Y. Xue, Y. Wang, Z. Yang, and C. Zhang, "Attention-enhanced long sequence temporal modeling method for backend CPU load prediction tasks," 2026, doi: 10.13140/RG.2.2.18397.78562.
- [21]Y. Xue, Y. Wang, A. Zhu, X. Sun, and C. Zhang, "Joint temporal-structural representation learning for distributed fault discrimination in microservice architectures," **2026 IEEE 8th International Conference on Communications, Information System and Computer Engineering (CISCE)**, pp. 1581–1586, Mar. 2026.
- [22]S. Pi, S. Zhang, and S. Shawulieti, "CacheRoute: KV-cache-aware routing for LLM backend services on Kubernetes," 2026.
- [23]S. Wang, B. Su, Y. Liu, S. Pi, and A. Zhu, "Safe adaptive resilience configuration for microservice backends," 2026.
- [24]Z. Jiang, S. Zhang, W. Huang, S. Pi, S. Shawulieti, and S. Xu, "Semantic-aware intelligent request routing in multi-model LLM backend gateways," 2026.
- [25]B. Zhang, Y. Zhan, Q. Wang, J. Ding, L. Yan, and W. Liu, "Hierarchical resource-aware federated edge learning in heterogeneous distributed systems," research report, 2026, doi: 10.13140/RG.2.2.30573.14562.
- [26]P. Hu, G. Uziel, and X. Ma, "Risk-constrained reinforcement learning for SLO-driven cloud-native elastic scaling," **Journal of Computer Technology and Software**, vol. 3, no. 1, 2024.
- [27]W. Huang, S. Xu, and Z. Jiang, "Privacy-aware distributed RAG with confidence-aware retrieval routing," **2026 3rd International Conference on Image Processing and Artificial Intelligence (ICIPAI)**, pp. 327–332, May 2026.
- [28]B. Chen, F. Chen, Y. Zhan, L. Ren, and Y. Tang, "Can we trust the metrics? Safe remediation under adversarial and noisy observability," **2026 3rd International Conference on Image Processing and Artificial Intelligence (ICIPAI)**, pp. 339–344, May 2026.
- [29]Z. Gu and L. Nie, "Knowledge graph-enhanced large language models for opinion target sentiment classification," **Journal of Computer Technology and Software**, vol. 4, no. 7, 2025.
- [30]L. Wang and C. Y. Chang, "Structure-aware root cause localization for microservice backends via joint graph representation learning and causal inference," **Journal of Computer Technology and Software**, vol. 3, no. 1, 2024.

-
- [31]J. Zhang, B. Chen, Y. Tang, F. Chen, Y. Zhan, and Z. Hu, "Drift-aware straggler mitigation for heterogeneous split federated learning via proactive cut-layer and asynchronous server updates," 2026.
- [32]X. Meng, L. Zheng, Y. Duan, S. Hu, and J. Chen, "Dynamic enterprise audit risk identification through graph-temporal transformer representation learning," ResearchGate preprint, 2026.
- [33]Z. Peng, S. Liu, X. Meng, Z. Su, and N. Sang, "A schema-aware large language model for XBRL tagging and cross-statement consistency checking," 2026.
- [34]Y. Wang, C.-C. Peng, S. Shu, Z. Wang, and I.-H. Lai, "Conflict-aware self-evolving memory for long-horizon agents," Jul. 2026, doi: 10.13140/RG.2.2.33763.72483.
- [35]J. Y. Su and X. Yan, "Process-aware multimodal drawing analytics for AI-assisted expressive arts support in educational settings," 2026.
- [36]S. Liu, S. Lu, Y. Liu, S. Y. Huang, and L. Chen, "Uncertainty-aware multimodal computer-use agent for GUI tasks," 2026.
- [37]K. Zhang, Y. Jiang, S. Liu, Z. Xiao, and Z. Yang, "Cost-constrained automatic workflow search for small LLM agents," 2026.
- [38]L. Ren, Y. Liu, and Y. Zhao, "Learning-based intelligent agents for backend resource scheduling and operational decision making," *Transactions on Computational and Scientific Methods*, vol. 5, no. 3, 2025.
- [39]T. Xia, X. Huang, Y. Zhou, F. Chang, and L. Ren, "Coordinating training and inference in heterogeneous cloud GPU clusters with learned interference models," *2026 8th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI)*, pp. 104–110, May 2026.
- [40]R. Fang, J. Zhou, Z. Zhu, R. Hu, and Y. Zheng, "Confidence-conflict routing between parametric memory, RAG, and long-context LLMs," 2026.
- [41]X. Chai, S. Li, Y. Zhao, B. Zhang, and J. Ding, "Causal generative re-ranking for multi-objective advertising," 2026.
- [42]Y. Ke, Q. Guo, D. Huang, and S. Li, "Explanation-stable anomaly detection for imbalanced heterogeneous tabular data under distribution shift," *2026 5th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, pp. 372–378, Jul. 2026.
- [43]N. Sang and C. Huang, "Risk-adaptive deposit-refund pricing for sustainable commercial orbit sharing," 2026.
- [44]C. Huang, "Integrating attention mapping and legal prior knowledge for interpretable legal reasoning with transformer-based models," 2025.
- [45]B. Zhang, S. Wang, C. He, and Y. Gong, "Pipelined KV cache migration for disaggregated LLM serving on heterogeneous edge–cloud testbeds," 2026.
- [46]Z. Luo, J. Chen, G. Feng, Y. Wang, and Q. Yang, "Generate-and-verify AI policy design for distributed serverless scheduling," *2026 6th International Conference on Advanced Algorithms and Neural Networks (AANN)*, pp. 715–720, Aug. 2026.
- [47]Q. Yang, G. Feng, Y. Wang, J. Chen, and Z. Luo, "BudgetSRE: Budget-aware multi-agent incident response for backend operations," *2026 3rd International Conference on Image Processing, Multimedia Technology and Machine Learning (IPMML)*, pp. 221–226, Aug. 2026.
- [48]L. Zheng, J. Chen, S. Hu, Y. Jin, and Y. Duan, "Evidence-grounded generation of audit procedures for financial statement risk assessment," 2026.
- [49]G. Feng, J. Chen, Z. Luo, Q. Yang, and Y. Wang, "AgentShield: A lightweight sentinel framework for mitigating hallucination propagation in multi-agent LLM systems," 2026.