# The Application of Transformer in Financial Management

**Zeying Yu**
*Department of CS & IT, Beijing University of Technology, Beijing, China*

**Abstract: This paper focuses on internet finance scenarios and sorts out the application of Transformer models in financial risk management. Internet finance has developed rapidly due to its characteristics of convenience and real-time performance, but it also faces various risks such as credit, operation, liquidity, and compliance. Traditional risk management tools have obvious shortcomings in integrating big data, achieving data sharing, and making models easy to understand. However, the Transformer model, with its unique self-attention mechanism, shows advantages in dealing with associated risks, systemic risks, and extreme risks in financial markets. It functions through methods such as hierarchical encoding, event-aware modeling, and multi-period feature fusion. Comparative studies have found that in tasks such as stock price prediction, exchange rate prediction, and financial report risk grading, Transformer performs better than traditional models such as ARIMA and GARCH, as well as other deep learning models such as LSTM and GRU, which can reduce errors and improve prediction accuracy. To solve the problem of high computing cost, Transformer is optimized through lightweight designs such as streamlined architecture and reconstructed attention; to cope with unstable data, it enhances its adaptability to different scenarios by means of event embedding and cycle fusion; to meet regulatory requirements, it makes its decision-making process more understandable through mechanisms such as semantic visualization and feature attribution. However, there are some contradictions in current research: simplifying the model structure may affect the expression effect; enhancing adaptability to dynamic changes may bring the risk of overfitting; deeper interpretation of the model may affect computational efficiency.**

**Keywords: Transformer; Financial Risk Management; Lightweight Design; Interpretability; Scenario Adaptation**

## 1. Introduction

Internet finance is a business model. It is dominated by traditional financial institutions or internet enterprises. These institutions or enterprises rely on internet technology. The services they provide include financial services such as financing, payment, and investment. Internet finance has entered a stage of rapid development since 2012. It has three significant characteristics, namely convenience, real-time performance, and no geographical restrictions. These characteristics have enabled it to rapidly expand its market scale and attract a large number of participants. Its service scope is very wide, covering P2P lending (Peer-to-Peer Online Lending), online payment, intelligent investment consulting and many other fields. These services have lowered the threshold of financial services and improved the efficiency of financial services. However, internet finance also faces higher risk challenges for three reasons: low access threshold, fast transaction rhythm, and loose audit mechanism [11].

The core risks and challenges of Internet finance are mainly reflected in three key areas. Fraud risks are widespread and rapidly evolving, stemming from the anonymity and instantaneity of transactions. They are manifested in identity theft, account takeover, and loan document forgery. The core difficulty lies in the continuous iteration of fraud patterns, which forces frequent updates to risk control models. Model and algorithm risks are particularly prominent, as core businesses rely on data models. Specific challenges include discriminatory outputs due to data bias, misjudgments caused by model design flaws, and the black box nature that weakens transparency, thereby affecting risk control and trust [11]. Compliance and regulatory arbitrage risks are highly complex, as the pace of innovation far exceeds regulation and often involves cross-border activities. Specific manifestations include regulatory lags creating gray areas, platforms seeking regulatory arbitrage spaces, high compliance costs in response to

 

multiple requirements, and cross-border regulatory conflicts. These risks are interwoven and collectively form the core challenges distinct from traditional finance.

From the perspective of research and practice, China is a core region for internet financial risk research. The research mainly focuses on some scenarios, including P2P platforms and credit evaluation [11]. At the same time, internet finance is closely linked to business model innovation. When enterprises explore new service models, they need to balance innovation and risk control and avoid amplifying risks due to defects in model design [3].

The development of P2P online lending industry in our country indicates that the main challenges faced by Internet finance are fraud risk, algorithm risk and compliance risk. The fraud risk manifests as collusion between the platform and the borrowers. For instance, ppdai.com assisted borrowers in quickly forging credit ratings through technical means and induced lenders to invest. More seriously, there was also the multi-head borrowing fraud, where borrowers took advantage of the platform's information isolation to borrow from multiple sources and maliciously default. In 2014, 136 platforms across the country went bankrupt within just 7 months, accounting for 11.3%, reflecting the continuous evolution of the fraud pattern. The algorithm risk is typical, such as the adverse selection phenomenon on the Renren Loan platform: high-risk borrowers provided a 13% high interest rate but were fully funded within 6 minutes, while high-quality borrowers only received 2% financing. This reveals serious flaws in the risk control model - over-reliance on surface data, lack of in-depth verification, and failure to establish a mutual prevention mechanism. Eventually, high-risk borrowers exploited the interest rate game to occupy the market. The compliance risk can be exemplified by the “Wangjinbao” case. This platform fabricated the deposit guarantee and guarantee commitment from the central bank, misappropriated funds for self-funding and then fled, exposing the regulatory lag and arbitrage space caused by the lack of regulatory authority for third-party payment and the loopholes in the system [14].

Financial risk management is a management process. This process is realized through functions such as planning, organizing, leading, coordinating, and controlling. Its role is to take measures against various risks that may occur in financial activities, aiming to reduce the negative impact of these risks [13]. Its management content has diversified characteristics, including risks in traditional financial scenarios such as credit risk, market risk, and operational risk, as well as risks unique to platforms in the internet financial environment [11,13].

In terms of management methods, the prevention and control system are mainly built through some ways, including risk retention, transfer, avoidance, and loss control. Risk retention refers to accepting part of the risk in order to obtain potential benefits. Risk transfer is to transfer risks to third parties through tools such as insurance and derivatives. Risk avoidance is to avoid participating in high-risk activities. Loss control is to reduce losses after risks occur through process optimization, monitoring and other means [13].

However, financial risk management still faces multiple challenges and problems. One of them is the insufficient adaptability between the traditional financial system and big data technology. The second problem is the lag in the knowledge structure of professional talents. The third problem is the limited data sharing among enterprises. These problems restrict the effectiveness of risk management [13]. In the field of internet finance, China's research has limitations: one is the concentration of research regions, and the other is the insufficient integration of external environment data [11]. Moreover, model interpretability and regulatory compliance are obstacles to the large-scale application of enterprises. In business model innovation, enterprises need to find a balance between business expansion and risk control and avoid excessive innovation, which may ignore potential risks [3].

Transformer is a sequence transduction model. It is completely based on the attention mechanism and does not rely on recurrent or convolutional neural networks [12]. Its core is to capture the dependencies between different positions in the sequence through the self-attention mechanism, so as to realize efficient modeling of input and output sequences.

The self-attention mechanism has a function that allows the model to pay attention to information at all positions in the sequence at the same time. It does not need to process data in sequence like recurrent neural networks, so it has stronger parallelism, which can significantly shorten the training time [12].

 

However, Transformer also faces challenges. Its self-attention mechanism leads to high computing cost and long training and inference time [2,12]. This limits its application, mainly in scenarios such as high-frequency trading, which have high requirements for real-time performance. In addition, in the field of internet financial risk management, the application research of Transformer is still relatively limited, and it has not yet become a mainstream model [11].

## 2. Literature Review

### 2.1 Macro and Scenario: Expansion of Transformer in Analyzing Systemic Risks

Financial risks not only come from fluctuations within the market but also are closely related to changes in macroeconomic indicators. The ability of Transformer to model long-sequence dependencies provides a new perspective for such analysis. BILSTM model is based on the bidirectional long short-term memory network, but it introduces a multi-head attention mechanism into the model, which provides ideas for analyzing the connection between macroeconomic factors and financial risks [6]. For example, it can capture the long-term dependence between indicators such as the unemployment rate and residents' income and the inflation rate. This analysis helps to early warn of potential liquidity risks. BILSTM model is similar to Ruan's "correlation modeling" idea, but the difference is that BILSTM model expands the scope of correlation analysis from within the market to the macroeconomic field [6,8].

In the field of risk simulation, The time fusion transformer (TFT) and large language models (LLMs) have been combined together [1]. They fuddsed quantitative data and qualitative scenario description data, where qualitative scenarios include major events such as geopolitical conflicts. This combination improves the realism of extreme risk simulation scenarios. This method extends Liang's focus on "extreme risks". Mathematical models captures extreme market fluctuations, while combining time fusion transformers (TFTs) with large language models (LLMs) enrich the specific manifestations of extreme risks by generating scenarios [1,5]. Both promote the development of risk analysis methods, transforming risk analysis from a single numerical "prediction" to a more comprehensive "scenario simulation".

### 2.2 Model Fusion and Method Reference: Expansion of Transformer in Risk Applications

The combined use of Transformer and other models further expands its application range in risk management. For example, GRU-Transformer hybrid model, which combines the Gated Recurrent Unit (GRU) and Transformer. GRU is good at capturing short-term fluctuation characteristics in time series, while Transformer is good at modeling long-term dependencies. This combination achieves a good balance in high-frequency risk monitoring tasks, with the model performing well in both rapid response and grasping the overall trend. This hybrid model makes up for the deficiency of using Transformer alone in adapting to different time scales [7].

In addition, Neural network quantile regression method did not directly apply the Transformer model, but the quantile loss function designed in their research provides an important reference for risk modeling. This loss function focuses on the prediction error distribution at different quantiles, which helps to more carefully characterize the characteristics of extreme tail risks. This idea is similar to the goal of the t-distribution loss function used by Liang. Both provide methodologies for dealing with "tail risks", which can provide reference for Transformer models to deal with the tail problems of financial risk data [5,15].

## 3. Methods

### 3.1 Literature Research Method

This study needs to systematically capture the technological evolution of Transformer technology in the field of financial risk management and pay attention to the practical application of this technology. Therefore, retrieval work will be carried out through authoritative databases, including Web of Science, IEEE Xplore, and China National Knowledge Infrastructure (CNKI). The search keywords include several important terms, specifically "Transformer financial risk management", "Transformer financial application", "event perception", and "credit evaluation". These keywords can cover basic model research, derivative variant research, and specific application scenario research. The time range of the literature is limited to the past five

 

years (2020-2025) to ensure the timeliness of the research.

### 3.2 Comparative Research Method

This study needs to clarify the technical positioning of Transformer and define the boundary of its advantages. The first dimension is model performance comparison, which needs to horizontally compare the differences between two types of models. First, the comparison between Transformer and traditional models, including ARIMA and GARCH (Traditional Time Series and Volatility Model), in the scenario of market risk prediction, such as the comparison of exchange rate prediction errors. Second, the comparison between Transformer and other deep learning models, including LSTM and GRU-CNN (Deep Learning Sequence and Hybrid Model) in the scenario of credit risk assessment.

The second dimension is the comparison of applicable scenarios, which needs to sort out the scenario adaptation characteristics of different Transformer variants. For example, lightweight models are suitable for specific scenarios: models such as TinyBERT (Distilled and compressed miniature BERT) are suitable for text-based risk assessment scenarios, such as 10-K report analysis scenarios, while time-series fusion models are suitable for different scenarios: models such as TCN-Transformer (Temporal convolutional networks are integrated with Transformers) are more suitable for high-frequency trading risk prediction scenarios.

### 3.3 Inductive Analysis Method

After the completion of literature research and comparative analysis, this study will Extract the essence through the inductive method: The first finding is about the technical evolution path, which needs to sort out three stages according to the time context. The first stage is the development of basic architecture, such as the application of the original Transformer. The second stage is the development of financial-specific variants, such as the application of event-aware Transformer and TCN-Transformer. The third stage is the development of cross-technical integration, such as the combination of LAMFormer (An extreme risk early warning model combining reinforcement learning and multi-head attention) and reinforcement learning. Through this context sorting, the evolution logic of key breakthroughs, including lightweight design and multi-modal fusion, is clarified [4].

The second finding is about the core research directions, which need to summarize the current three research focuses. The first direction is lightweight optimization research, such as research on knowledge distillation to compress models. The second direction is interpretability enhancement research, such as attention visualization research. The third direction is multi-factor fusion research, such as research on macroeconomic indicator embedding. It is also necessary to analyze the existing technical bottlenecks in each direction.

The third finding is about unsolved problems, which need to extract common industry challenges. First, the problem of data timeliness: macroeconomic data usually lags behind by 2-3 quarters, which leads to insufficient real-time performance of the model. Second, the problem of computing power accessibility: traditional Transformer consumes too much computing power, so it is difficult to adapt to small and medium-sized financial institutions [10]. Finally, the problem of regulatory compliance: the model has the "black box" characteristic, which conflicts with the requirements of the EU AI Act.

## 4. Discussion and Result

### 4.1 Breakthrough Progress in Lightweight Design

Financial risk management scenarios have strict requirements on the real-time performance of the model. Traditional Transformer has encountered bottlenecks in deployment due to complex calculations. Lightweight innovation has made breakthroughs in three aspects.

Architecture Compression: The TinyBERT model proposed uses the distilled TinyBERT as the encoder. Its parameter count is reduced by 7.5 times. On devices with 11GB of memory, the model can process thousands of financial reports in real-time. Its training speed is 39% faster than the standard Transformer [10]. This design first proves that lightweight models can maintain prediction accuracy in resource-constrained environments, which provides the possibility for edge computing deployment.

Attention Mechanism Reconstruction: The Layer-Transformer model adopts a hierarchical attention structure [8]. The first layer processes the time-series features of a single stock, and the second layer learns the sector correlation between stocks. This structure reduces the

 

calculation of redundant information. In the prediction task of the A-share market, the inference efficiency of the model is improved by 40%, which verifies that it is feasible to improve efficiency while ensuring accuracy.

Dynamic Feature Screening: Shi combined XGBoost (Gradient boosting decision trees are used for feature screening and contribution analysis) with Bayesian optimization to screen 101 key features from 268 factors. This reduces the input dimension of TCN-Transformer by 42%. This factor compression strategy based on contribution rate significantly reduces the interference of noise. At the same time, it also improves training efficiency, reducing the number of iterations by 43% [9].

### 4.2 Multi-Dimensional Enhancement of Scenario Adaptability

Financial data has the characteristic of instability, which requires the model to have dynamic adaptation ability. Scenario adaptation technology improves the stability of the model through three mechanisms.

- Event-aware Modeling: This model embedded the encoding of macroeconomic events, including the Federal Reserve's policy adjustments, in exchange rate prediction. This enables Event-aware Transformer to capture sudden fluctuation patterns. On the 5-minute data of USD/JPY, this design reduces the predicted RMSE by 44.2%, which proves the role of external event information in explaining time-series mutations [16].
- Multi-period Feature Fusion: TinyBERT model reorganized historical features by week and month dimensions, which enhances the model's sensitivity to the seasonal laws of financial reports. In the 2023 test, this strategy made the directional accuracy (DA) of quarterly predictions reach 0.576, which is 9.7 percentage points higher than the benchmark model [10].
- Expansion of Risk Measurement System: Zeng & You broke through the limitation of traditional volatility, introduced Sortino ratio and skewness analysis, and constructed a composite risk indicator [15]. In the backtest of extreme markets, this scheme reduced the RMSE fluctuation of the QRNN model by 22%, which verifies the role of multi-dimensional risk characterization in improving model stability.

### 4.3 Decision Trust Driven by Interpretability

Black-box models affect the demand for regulatory compliance. The interpretability mechanism has established trust in decision-making through a three-layer architecture.

Semantic-level Visualization: TinyBERT model developed dynamic word cloud technology, generating risk hot word maps by normalizing attention weights. For example, a high weight of "debt dependence" indicates high risk. This mechanism makes the consistency between model decisions and analyst evaluations reach 89%, which significantly reduces regulatory doubts [10].

Feature Attribution Quantification: The contribution of technical indicators through factor contribution rate analysis and found that the closing price and trading volume account for more than 60% in stock price prediction, while the impact of factors such as turnover rate is less than 5%. This quantitative attribution provides an operable optimization path for feature engineering [9].

Hierarchical Interpretation Framework: The Layer-Transformer model realizes the visualization of sector correlation through the hierarchical design of stock coding and time coding. In the A-share industry analysis, this design increases the attention coefficient (IC) of stocks within the sector by 0.035 [8]. This fine-grained interpretation meets the compliance requirements of financial institutions for "traceability of prediction basis".

### 4.4 Performance Evaluation with Unified RMSE Measurement

To horizontally compare the accuracy of the models, we use RMSE as the core evaluation benchmark.

- Lightweight Contribution: TinyBERT reduces the RMSE to 17.12 in financial report risk assessment through architecture compression and triple loss optimization, which is 15.6% higher than the traditional Transformer.
- Time-series Adaptation Advantage: TCN-Transformer integrates the local feature extraction of TCN and the global modeling of Transformer [9]. In the prediction of Shanghai A50 stocks, its RMSE reaches 52.06, which is significantly better than the pure Transformer model.
- Event Response Value: Event-aware Transformer has an RMSE of only 0.0413 in the scenario of exchange rate mutations, which proves the adaptability of event coding to high-frequency data [15].

 

As shown in Table 1, the unified RMSE measurement provides a horizontal comparison of model performance across different financial risk scenarios that the collaborative optimization of lightweight and attention is the key to performance improvement. For example, TCN-Transformer reduces data noise through feature screening, improving the RMSE of Shanghai A50 prediction by 15.42% [9]. The hierarchical attention design of Layer-Transformer reduces the computational load by 28% while maintaining accuracy [8].

**Table 1. UnifIED the RMSE Measurement Model Performance**

| **Scenario Type** | Best Model | RMSE | **Baseline Model** | Improvement |
|---|---|---|---|---|
| Stock Price Prediction | TCN-Transformer | 52.06 | ARIMA-GARCH | +59.4% |
| FX Prediction (High-Frequency) | Event-Aware | 0.0397 | LSTM | +85.1% |
| Financial Risk Grading | TinyBERT | 17.12 | TF-IDF | +9.3% |
| Cross-Market Asset Correlation | Layer-Transformer | 0.218 | Correlation Matrix | +42.7% |
| Extreme Risk Warning | LAMFormer | 0.112 | GARCH | +68.9% |

### 4.5 Contradictory Findings and Research Challenges

There are three contradictions in current research that need to be solved. Balance between Lightweight and Expressiveness: Although TinyBERT improves efficiency, its performance fluctuates in small-cap stock prediction, with an RMSE fluctuation of ±2.3. This reflects the limitation of compressed models in capturing atypical patterns [10].

Dynamic Adaptation vs. Overfitting Risk: Event-aware Transformer excels during policy-intensive periods, but its RMSE increases by 12.7% during economically stable periods, revealing weaker generalization capabilities in event-dependent models [10,15].

Explanatory Depth vs. Computational Cost: Detailed attention visualization analysis adds 30% inference overhead, conflicting with real-time risk control requirements [8].

## 5. Conclusion

This survey systematically explores the advancement of Transformer models in financial risk management. The research focuses on four key directions: lightweight model design, scenario adaptability enhancement, explainability improvement, and unified performance evaluation using the RMSE metric. Analysis demonstrates that Transformer models, through specific technical refinements, can more effectively address complex challenges in financial risk management. However, challenges remain in specific internet finance scenarios such as P2P lending, where data sparsity, dynamic fraud patterns, and regulatory fragmentation constrain model efficacy-manifested in limited adaptability to evolving borrower collusion schemes and insufficient real-time monitoring capabilities under high-frequency transaction volumes.

Breakthroughs in Lightweight Design: Architecture compression strategies, attention mechanism restructuring methods, and dynamic feature filtering techniques collectively address the computational burden [8,9,10]. These advancements significantly reduce resource demands, enabling model deployment in resource-constrained environments-a crucial factor for small and medium-sized financial institutions.

Multidimensional Enhancement in Scenario Adaptability: Diverse Transformer improvements exhibit strong capabilities across varied financial risk scenarios. Event-aware modeling improves the accuracy in capturing sudden FX fluctuations [16]; multi-period feature fusion mechanisms enhance sensitivity to identifying seasonal patterns in financial statements; and expanded risk metric systems strengthen model stability under extreme market stress [10,16]. Collectively, these findings demonstrate the significant potential of Transformers in handling the non-stationary nature of financial data.

Multi-path Development of Explainability Mechanisms: Semantic-level visualization tools, feature attribution quantification methods, and hierarchical explanation frameworks synergistically reduce model opacity in the decision-making process [8, 9, 10]. This progress directly addresses the stringent transparency requirements of financial regulation, thereby effectively enhancing the credibility of model outputs.

Unified RMSE Evaluation Validates Performance Advantage: In tasks including stock volatility prediction, high-frequency FX forecasting, and financial statement risk grading, refined Transformer models consistently outperform traditional models and other deep

 

learning approaches [9, 10, 16]. This result systematically validates their predictive accuracy and engineering reliability.

However, unresolved contradictions remain: Lightweight design often compromises model expressiveness; enhanced dynamic adaptation can trigger overfitting risks; and increased explanatory depth amplifies computational costs. These contradictions outline critical paths for future research: developing adaptive lightweight architectures that balance efficiency and accuracy; improving the cross-scenario generalization ability of event-driven models; and constructing real-time explanation mechanisms with low computational overhead to meet high-frequency risk control demands.

**References**

[1] Aldridge, I., & Kim, D. (2024). Quantitative financial models with scenarios from llm: Temporal fusion transformers as alternative monte-carlo. Available at SSRN 4999492.

[2] Dong, Z., & Xu, L. (2025). Deep learning for financial forecasting and strategic business optimisation in enterprises. International Journal of Information and Communication Technology, 26(19), 79-101.

[3] Foss, N. J., & Saebi, T. (2017). Fifteen years of research on business model innovation: How far have we come, and where should we go?. Journal of management, 43(1), 200-227.

[4] Li, X., Bian, C., Li, X., Yu, S., & Jiang, B. (2025). Lamformer: LSTM-enhanced agent attention and mixture-of-experts transformer for efficient stock price prediction. International Journal of Machine Learning and Cybernetics, 1-14.

[5] Liang, C. Y. (2025). Research and Application of Stock Index Trend Prediction Based on Transformer (Master's Thesis, Beijing University of Chemical Technology). Master's https://link.cnki.net/doi/10.26939/d.cnki.gbhgu.2025.000541 doi:10.26939/d.cnki.gbhgu.2025.000541.

[6] Lyu, S. (2025, April). BILSTM Seq-to-Seq with Multi-Head Attention for Consumer Price Index Forecasting. In 2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID) (pp. 298-301). IEEE.

[7] Pradhan, R., Alazzam, M. B., Keswani, S., Bhasin, N. K., Jaff, N. A., & Muthuperumal, S. (2025, February). A Hybrid GRU-Transformer Model for Financial Forecasting and Risk Management. In 2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS) (pp. 1-5). IEEE.

[8] Ruan, J. H., Wang, K., Feng, T. Y., & Wang, Z. C. (2025). Stock Prediction Method Based on Transformer. Computer and Digital Engineering, 53(05), 1375-1380+1433.

[9] Shi, D. J. (2025). Research on Stock Price Prediction Based on TCN-Transformer with Multi-Factor Fusion (Master's Thesis, Jiangxi University of Finance and Economics). Master's https://kns.cnki.net/kcms2/article/abstract?v=IKKGlZ0AkeVpF1-euB3WRFZkORiNRFRI7E6CYGTSeAzLTItGbonRVv2zlXF1NPZWZHA8Bx0qaFH-FF_lpp3X4fQqicF5pkOFDS5qI9-WK63AeelKmhIzFHGuepU3IwwNvch6xvsM9GcpXmabOWy_f9Zw-sEvNnsiYo07B3VXQzdfxrH7V2__JQ==&uniplatform=NZKPT&language=CHS

[10] Tan, X. W., & Kok, S. (2025). Explainable AI for Comprehensive Risk Assessment for Financial Reports: A Lightweight Hierarchical Transformer Network Approach. arXiv preprint arXiv:2506.23767.

[11] Tian, X., Tian, Z., Khatib, S. F., & Wang, Y. (2024). Machine learning in internet financial risk management: A systematic literature review. Plos one, 19(4), e0300195.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[13] Wang, L. S. (2024). Research on Enterprise Financial Risk Management in the Era of Big Data. Business Exhibition Economy, (17), 185-188.

[14] Xie, C., & Wang, J. (2015). Research on credit risk of P2P network lending platform. Finance, 5(1), 1-5.

[15] Zeng, Z. F., & You, Y. (2020). Financial Risk Warning Based on Neural Network Quantile Regression. Statistics and Decision, 36(14), 137-140.

[16] Zhang, S., Che, T., Zhu, Z., Luo, G., & Feng, P. (2025). Forecasting of exchange rate time series based on event-aware transformer mode. Soft Computing, 1-11.

 