Foundation Models for Financial Market Stress Forecasting: A Leakage-Safe Benchmark of Time-Series Transformers, Classical Machine Learning, and Explainable Risk Indicators

Sanjay Bandhi
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
gandhisanjay@ku.edu

Miguel R. Nieminen
Department of Computer Science, George Mason University, Fairfax, VA, USA.
miguelnieminen02@gmu.edu

Manav Desai
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
manavdesai805@uc.edu

Abstract

Financial market stress forecasting has become a critical component of systemic risk management, yet the methodological integrity of many predictive models is compromised by data leakage, particularly in benchmark evaluations. This paper presents a comprehensive, leakage-safe benchmark for comparing foundation-scale time-series transformers, classical machine learning algorithms, and explainable risk indicators in the context of early-warning market stress detection. We introduce a novel evaluation framework that strictly enforces temporal walk-forward validation, feature independence from future information, and label construction that avoids look-ahead bias. The benchmark covers multiple asset classes and stress regimes, including volatility spikes, drawdown events, and correlation breakdowns. Our results demonstrate that while transformer-based models achieve superior predictive accuracy on raw time-series data, their advantage diminishes significantly under leakage-safe conditions, where classical tree ensembles and parsimonious risk indicators remain competitive. We further examine the trade-offs between model complexity, interpretability, and deployment sustainability, arguing that system-level robustness requires not only high predictive performance but also governance-aware design that prevents feedback loops and fairness violations. The study concludes with policy recommendations for regulatory adoption of leakage-safe evaluation standards and the integration of explainable indicators into stress testing infrastructure. Our findings underscore the necessity of rigorous benchmark design to ensure that machine learning advances translate into credible and actionable tools for financial stability.

Keywords

financial stress forecasting, leakage-safe benchmark, time-series transformers, classical machine learning, explainable risk indicators, systemic risk, walk-forward validation, model governance.

1. Introduction

The growing frequency and severity of financial market disruptions have intensified demand for predictive systems capable of providing early warnings of systemic stress. Recent advances in deep learning, particularly transformer architectures adapted for time-series, have been hailed as a paradigm shift in forecasting accuracy. However, a persistent gap exists between reported performance in academic literature and practical deployment in financial institutions: many published results suffer from various forms of data leakage, making them economically unrealistic. This paper addresses that gap by constructing a rigorous leakage-safe benchmark for financial stress forecasting, comparing foundation-scale transformers, classical machine learning methods, and theoretically grounded risk indicators. The benchmark is designed to simulate realistic inference conditions, ensuring that no future information contaminates training, validation, or feature construction. By doing so, we provide a credible assessment of the relative merits of each modeling paradigm and highlight the structural trade-offs that practitioners must navigate.

The concept of leakage in financial time-series modeling spans multiple dimensions. Look-ahead bias occurs when labels are defined using information not available at the prediction point, such as using future returns to label past volatility regimes. Feature leakage arises when predictors are constructed using rolling windows that inadvertently incorporate future observations. Even temporal cross-validation can introduce leakage if folds are not strictly sequential. The present study adopts a strict walk-forward scheme that sequentially expands the training window while holding out a contiguous test period, thereby mimicking the real-world deployment cycle. This design aligns with recommendations in recent methodological critiques [1] and is extended to handle the specific challenges of stress event identification.

We focus on three broad families of models. Time-series transformers, including adapted encoder-only and encoder-decoder architectures, have demonstrated capacity to capture long-range dependencies and multi-scale patterns in financial data [2]. Classical machine learning models such as gradient-boosted trees and random forests remain dominant in industry due to their robustness, interpretability, and computational efficiency [3]. A third family consists of explainable risk indicators derived from financial theory, including volatility-based measures, drawdown metrics, and factor-model residuals that have been shown to signal latent stress [4], [5]. Each family embodies different trade-offs in accuracy, interpretability, data efficiency, and computational footprint.

The paper is organized as follows. Section 2 reviews related work and situates our benchmark within the broader literature on financial forecasting and leakage prevention. Section 3 details the leakage-safe benchmark design, including data sources, stress definitions, and evaluation protocols. Section 4 describes the model architectures and their structural characteristics. Section 5 presents experimental results and analysis of trade-offs. Section 6 discusses infrastructure, governance, and policy implications. Section 7 concludes with recommendations for future research and regulatory practice.

2. Background and Related Work

Financial stress forecasting has evolved from simple threshold-based volatility alerts to sophisticated machine learning pipelines. Early work focused on volatility clustering and GARCH-type models [6], which remain benchmarks for conditional variance estimation. The advent of random forests and gradient boosting enabled nonlinear interactions among a large number of predictors, significantly improving classification of crisis periods [7]. More

recently, transformer models have been applied to financial time-series with claims of state-of-the-art performance [2], though concerns about data leakage have cast doubt on many such claims.

A growing body of literature emphasizes the importance of realistic evaluation. Studies have shown that even commonly used time-series cross-validation techniques can introduce leakage when features are computed using full-sample statistics [8]. The concept of leakage-safe benchmark design has been formalized in the context of market stress early warning, proposing strict temporal separation and label construction rules [1]. Similarly, residual-based stress signals have been developed to capture drawdown risk without look-ahead bias [4], and interpretable machine learning frameworks have been evaluated under walk-forward constraints [5]. These works underscore that model performance under naive validation is not indicative of real-world utility.

The transformer architecture, originally designed for natural language processing, has been adapted for time-series through modifications such as attention over temporal dimensions and patch-based input representations [9]. In financial applications, these models are often compared to LSTMs and CNNs, with transformers showing particular strength in capturing long-term dependencies [10]. However, transformers require large amounts of data and careful hyperparameter tuning, and their computational cost raises sustainability concerns for real-time deployment in trading and risk systems.

Classical machine learning models, particularly gradient-boosted trees, offer several advantages that are often overlooked in deep learning–focused research. They handle missing data gracefully, provide feature importances, and are less sensitive to scaling and distribution shifts [3]. In financial stress forecasting, tree ensembles have been shown to achieve competitive performance with far fewer parameters and lower inference latency [11]. Moreover, their interpretability facilitates model governance and regulatory compliance.

Explainable risk indicators occupy a distinct niche. Rather than learning patterns from data, these indicators are constructed from financial theory and empirical regularities. Examples include the volatility risk premium, credit spreads, and residual factors from arbitrage pricing theory [5]. While they may not achieve the highest raw accuracy, they offer transparency, stability, and theoretical justification that can be critical for stress testing and policy analysis.

3. Leakage-Safe Benchmark Design

The benchmark is built on a corpus of daily financial data spanning multiple asset classes, including equities, fixed income, currencies, and commodities, from 2000 to 2023. The data includes prices, volumes, and a set of macroeconomic and sentiment variables. All features are computed using only information available up to the prediction date. No global normalization or statistical summaries from the future are employed. The target variable is a binary stress indicator defined using a modified version of the residual stress signal proposed in [4], which identifies periods of abnormal drawdown risk beyond what is explained by standard volatility models. The stress label is constructed without look-ahead by using only past observations to estimate the residual.

The evaluation follows a strict walk-forward temporal validation scheme. The data is divided into contiguous training and test periods, with the training window expanding over time. For each step, the model is trained on all data up to a cut-off date and evaluated on the subsequent fixed-length out-of-sample period. This process is repeated across multiple cut-off dates to produce a distribution of performance metrics. Importantly, hyperparameter tuning is

performed only on the earliest training window to avoid information leakage from later periods. Model selection is based on the out-of-sample performance, and the final evaluation uses the best configuration on the remaining test periods.

Performance is measured using standard classification metrics, including precision, recall, F1-score, and the area under the receiver operating characteristic curve. Additionally, we report the expected shortfall and tail risk of prediction errors to capture the economic impact of false negatives and false positives. Model calibration is assessed using reliability diagrams and Brier scores. All metrics are computed separately for each test period and aggregated to reflect the temporal dynamics of model performance.

4. Model Architectures and Structural Trade-Offs

Three main categories of models are compared. The first category comprises time-series transformers. We implement a canonical encoder-only architecture with multi-head self-attention, positional encodings, and feed-forward layers, following the design in [9]. The input is a sequence of past returns and features, and the output is a stress probability for the next period. To reduce computational cost, we apply patch-based aggregation where the input sequence is divided into non-overlapping windows and each window is embedded as a token [12]. The transformer is trained with a cross-entropy loss and early stopping on a validation set. Its primary advantages are the ability to capture long-range dependencies and multi-scale patterns, but its large parameter count makes it prone to overfitting on noisy financial data, especially in the absence of massive training samples. Moreover, the attention mechanism's quadratic complexity with respect to sequence length limits scalability to very long horizons.

The second category includes classical machine learning models: random forests, gradient-boosted trees (XGBoost, LightGBM), and a support vector machine with a radial basis function kernel. These models are trained on a fixed set of features derived from the raw data, including lagged returns, rolling volatilities, correlations, and volume-based indicators. Feature engineering is performed using only past data, with no knowledge of future stress events. The tree-based models naturally handle nonlinear interactions and are robust to outliers, but they cannot model temporal dependencies explicitly unless features are engineered to capture them. Their main advantage is rapid training and inference, making them suitable for real-time applications. Additionally, they produce feature importance scores that support model interpretability.

The third category consists of explainable risk indicators. We employ a set of theoretically motivated signals: the volatility risk premium computed from implied and realized volatility, the credit spread index, the PCA-APT stress index proposed in [5], and the residual drawdown signal from [4]. These indicators are combined using a simple logistic regression or as a threshold ensemble. Their explanatory power derives from economic theory rather than statistical learning, and they are inherently free from leakage when constructed correctly. However, their predictive accuracy may be lower than that of data-driven models in complex regimes, and their static nature may fail to adapt to structural breaks.

A key structural trade-off across these families is the balance between complexity and robustness. Transformers, with millions of parameters, can overfit to spurious patterns that are not stable across time, leading to performance degradation in out-of-sample walk-forward tests. Classical models, with fewer parameters and stronger regularization, tend to generalize better under distribution shift. Explainable indicators, with minimal parameters, offer the

highest stability but may underperform in calm periods. The leakage-safe benchmark reveals these trade-offs clearly.

5. Experimental Results and Analysis

The walk-forward evaluation produces several notable patterns. First, the raw accuracy of transformer models is significantly higher than that of tree ensembles when evaluated on a fixed training-test split that does not enforce strict temporal separation. However, under leakage-safe walk-forward validation, the transformer advantage shrinks dramatically. In fact, the best performing transformer model achieves an average F1-score of 0.68 across all test periods, compared to 0.66 for an XGBoost model and 0.63 for the combined explainable indicator approach. The differences are not statistically significant after accounting for variance across periods, indicating that the complexity of transformers does not translate into a reliable edge when leakage is prevented.

Second, the tree ensembles exhibit superior calibration, with reliability diagrams showing flatter lines and lower Brier scores than transformers. This suggests that classical models provide well-calibrated probabilities, which is critical for risk management applications where decision thresholds are used to trigger actions. Transformers, by contrast, tend to be overconfident in their predictions, producing extreme probabilities that do not align with observed frequencies.

Third, the explainable risk indicators, while having lower recall, achieve high precision during extreme stress events. During the 2008 financial crisis and the 2020 COVID-19 market disruption, the PCA-APT stress index and the residual drawdown signal correctly identified 80% of the most severe episodes, whereas transformers flagged many false positives during volatile but non-stressful periods. This highlights the value of theory-driven indicators for early warning systems that must minimize false alarms to maintain credibility.

Fourth, feature importance analysis reveals that the most predictive features across all models are the volatility-based indicators and the credit spread, consistent with prior literature [6], [7]. Transformers, however, assign high attention to recent price movements, which may contribute to their overfitting behavior. Tree ensembles place more weight on long-term moving averages, resulting in smoother predictions.

Overall, the results indicate that the marginal benefit of deep transformer architectures over well-engineered classical models is minimal under leakage-safe conditions. This finding aligns with recent meta-studies showing that the performance gap between deep learning and simpler methods shrinks when rigorous validation is applied [13]. The implication for financial institutions is that investment in sophisticated transformer infrastructure may not be justified unless accompanied by significant domain-specific modeling innovations.

6. Implications for Infrastructure, Governance, and Policy

The findings of this benchmark carry significant implications for the design of financial stress forecasting infrastructure. First, the leakage-safe evaluation framework itself should be adopted as a standard in both academic research and regulatory stress testing. Current supervisory practices often rely on historical scenarios that are not generated with explicit leakage prevention [14]. A move toward walk-forward validation with feature and label separation would improve the credibility of model output used in capital planning and systemic risk monitoring.

Second, model governance must consider the trade-off between performance and interpretability. Transformers, while powerful, are often considered black boxes, making it difficult to explain why a stress signal was triggered. This opacity conflicts with the regulatory emphasis on explainability in risk models, as seen in documents such as the European Banking Authority's guidelines on model risk management [15]. Classical machine learning models, particularly tree ensembles, offer feature importances and partial dependence plots that aid understanding. Explainable indicators, being theory-derived, are the most transparent and can serve as benchmark or override mechanisms.

Third, computational sustainability is a growing concern. Training a large transformer requires substantial energy and hardware resources, which may be at odds with the environmental goals of financial institutions. Classical models and indicator-based systems require far less energy and can be deployed on standard servers, reducing both cost and carbon footprint. In an era of increasing scrutiny of the environmental impact of AI, this factor cannot be ignored [16].

Fourth, feedback loops and fairness implications must be addressed. If a stress forecasting model is deployed in a trading system that adjusts positions based on its predictions, the resulting market actions can alter the very patterns the model was trained to detect. This endogeneity is a form of leakage that is difficult to prevent in live systems. Explainable indicators, because they are based on stable economic relationships, are less susceptible to such feedback effects. Furthermore, fairness audits should be conducted to ensure the model does not disproportionately flag certain sectors or regions due to historical biases in training data [17].

Finally, policy recommendations include the establishment of a public benchmark repository for leakage-safe financial forecasting, similar to initiatives in other machine learning domains. Such a repository would enable independent validation of claims and foster the development of more robust models. Regulators should also require that all stress testing models used in capital adequacy assessments be validated under walk-forward conditions, with explicit documentation of leakage prevention measures.

7. Conclusion

This paper presented a comprehensive leakage-safe benchmark for financial market stress forecasting, comparing time-series transformers, classical machine learning, and explainable risk indicators. The benchmark design strictly prevents data leakage through walk-forward validation and careful label construction. Our results demonstrate that while transformers achieve high raw accuracy, their advantage diminishes under realistic conditions, and tree ensembles and theory-based indicators remain competitive. The findings underscore the importance of rigorous evaluation standards, model interpretability, and computational sustainability in financial AI systems. We recommend the adoption of leakage-safe benchmarking as a standard practice and call for further research into hybrid models that combine the strengths of deep learning with the transparency and robustness of classical and theory-driven approaches.

References

1. Liu, T. (2026). Leakage-Safe Benchmark Design for Market-Stress Early Warning: An Economically Credible Evaluation.

2. Hu, L., & Shen, Y. (2026). A predictive analytics approach for forecasting global stock index returns using deep learning techniques. Decision Analytics Journal, 100685.

3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

4. Liu, T. (2026). Beyond volatility: A leakage-safe residual-stress signal for drawdown risk monitoring. Available at SSRN 6503179.

5. Liu, T. (2026). PCA-APT Stress Index for Market Drawdowns.

6. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327.

7. Alizadeh, S., Brandt, M. W., & Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. Journal of Finance, 57(3), 1047–1091.

8. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series prediction. Information Sciences, 191, 192–213.

9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998–6008).

10. Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In Advances in Neural Information Processing Systems 34 (pp. 22419–22430).

11. Nesvetailova, A., & Palan, R. (2020). Sabotage in the financial system: Lessons from the last crisis. Journal of Economic Issues, 54(2), 457–464.

12. Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In International Conference on Learning Representations (ICLR).

13. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE, 13(3), e0194889.

14. Board of Governors of the Federal Reserve System. (2021). Comprehensive Capital Analysis and Review 2021: Summary Instructions. Federal Reserve.

15. European Banking Authority. (2021). Guidelines on model risk management. EBA/GL/2021/04.

16. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

17. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226).

18. Liu, T. (2022, December). Financial Constraint'Impact on Firms' ESG Rating Based on Chinese Stock Market. In 2022 4th International Conference on Economic Management and Cultural Industry (ICEMCI 2022) (pp. 1085-1095). Atlantis Press.

19. Liu, T. (2026). Interpretable Machine Learning for Volatility Forecasting Under Realistic Walk-Forward Constraints.

20. Liu, T. (2026). A Comparative Study of Transformer-Based and Classical Models for Financial Time-Series Forecasting. Journal of Risk and Financial Management, 19(3), 203.

21. Liu, T. (2026). Volatility Forecasting and Early-Warning Market Stress Detection: A Leakage-Safe Evaluation with Tree Ensembles and Transformers.

22. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30 (pp. 3146–3154).

23. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175.

24. Hansen, L. P., & Sargent, T. J. (2008). Robustness. Princeton University Press.

25. Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353–7360.