Temporal Fusion Transformers for S&P 500 Return Forecasting with Mixed-Frequency Macroeconomic Data

Sam Ehrle
Arizona State University
sehrle@asu.edu

Gourishankar Mahadeo Bansode
Arizona State University
gourishankar@asu.edu

Ojas Makarand Deodhar
Arizona State University
odeodhar@asu.edu

Aman Pandey
Arizona State University
apand105@asu.edu

Danyal Khorami
Arizona State University
dkhorami@asu.edu

Abstract

Predicting stock returns is hard: the efficient market hypothesis implies that prices already reflect available information, and daily returns have a signal-to-noise ratio around 30:1. We test whether Temporal Fusion Transformers (TFT) can forecast S&P 500 returns using mixed-frequency data, with some features such as VIX updating daily and others, such as macroeconomic indicators of CPI, only updating monthly. The standard approach of forward-filling treats 30-day-old inflation data the same as today's volatility measures, thereby ignoring how information gets stale. We compare TFT against ARIMAX and LSTM baselines to see how well attention mechanisms handle this heterogeneity. Our experiments reveal prediction collapse when models process data with different update frequencies and analyze how attention patterns shift across market regimes. We test modifications that explicitly weight data by freshness. Results show the presence of fundamental challenges in applying transformer architectures to financial forecasting when the inputs have mixed temporal resolution.

Introduction

Predicting the market is a considerable challenge, primarily due to the fundamental properties of how prices fluctuate. At its core, the efficient market hypothesis, formulated by Fama et al. (1970) [1], states that prices in capital markets fully reflect all available information, rendering financial markets "efficient" in the sense that we can not systematically exploit historical data and publicly known signals for excess returns. On the one hand, the random walk model is confirmed to be natural in light of the independence and identical distribution of asset prices, with returns being unpredictable based solely on prices.

In terms of financial market forecasting, any signal that can go beyond the 50% level of accuracy demonstrated by a random walk process can be considered significant in light of the inherent uncertainty of efficient markets. For instance, signals with even minimal advantages over the random walk, such as 52% accuracy, can be considered outstanding and may represent an opportunity for exploiting market inefficiencies. The efficient market theory proposed by Fama states that signals significantly deviating from the random walk model are unlikely to persist in the market.

Fama and others have documented that financial return data is just plain noisy. Economic news, the mindset of investors, and technological changes constantly and easily move prices and may obscure the signal. This is why most models tend to underperform simple statistical benchmarks. Identifying a potential signal is not enough. The actual problem is to extract a stable, implementable signal from all the random noise, particularly when the data is uncertain. Markets also show regime changes where model parameters shift between bullish and bearish periods [2, 3].

The main problem in time-series forecasting lies in capturing non-local temporal structures, the distant dependencies that conventional sequential models (like RNNs) often miss. The transformer-based architecture is becoming one of the state-of-the-art solutions, as it utilizes a self-attention mechanism, allowing for changes in weights based on historical observations and providing a framework to model complex, non-sequential relationships. This breakthrough has made it a leading approach, exemplified by influential models like the Temporal Fusion Transformer (TFT) [4]. TFT specifically addresses multi-horizon forecasting with heterogeneous inputs, utilizing attention to intelligently blend historical lags, external covariates, and contemporaneous signals—a necessity in mixed-frequency environments typical of financial prediction. Besides TFT, the

Autoformer's decomposition framework for better periodic modeling [5] and Informer's sparse self-attention for efficiency in long sequences [6], provide a strong dominance in transformer-based architecture models for time series forecasting. These reasons make the Transformer-based architecture one of the choices for modeling S&P 500 returns forecasting with mixed-frequency macroeconomic data.

Non-stationarity is a fundamental challenge in financial prediction. Regime changes, shifts in trends, and structural breaks in market data mean that historical patterns most often fail to forecast the future behavior of the market. This characteristic creates unstable dynamics where statistical properties of the data change over time. Recent work addresses non-stationarity through adapting segmentation of the input series [7].

Another challenge in predicting the stock market is data heterogeneity. This problem arises when features update at different frequencies, as some update daily (VIX volatility) while others update monthly or even quarterly (CPI inflation, GDP growth). The mixed-frequency in such data is not predictable by standard processing methods; for instance, forward-fill propagates the most recent value across time periods and, regardless of data freshness, treats 30-day-old CPI measurements identically to yesterday's stock price. Such an approach overlooks the temporal decay of information content, eventually creating an asymmetry where stale macroeconomic data receives equal weighting to fresh market signals. Ghysels, Santa-Clara, and Valkanov (2004) note that "the relevant information is high frequency data, whereas the variable of interest is sampled at a lower frequency," and introduced Mixed-Data Sampling (MIDAS) regressions to address this challenge through polynomial-weighting schemes that parsimoniously aggregate high-frequency observations using distributed lag polynomials[8].

This work makes the following three key contributions: (1) We systematically investigate the effectiveness at which Temporal Fusion Transformers can predict S&P 500 returns from mixed-frequency data, in comparison with classical ARIMAX and LSTM baselines. (2) We investigate how these models fail when being asked to process data with disparate update frequencies by analyzing prediction collapse and attention pattern behavior. (3) Architectural modifications are evaluated to address mixed-frequency forecasting challenges, with analysis of what components across attention mechanisms, output generation, or loss functions, can benefit the most from domain-specific constraints.

Related Work

Deep Learning for Financial Forecasting

To deal with these limitations, many of the researchers have increasingly turned to neural network approaches, most notably, LSTM networks. Fischer et al. (2018) [9] were among the first to use LSTM architectures in predicting the direction of movements and daily returns for constituents of S&P 500 from 1992-2015. They demonstrated that LSTMs have better generalization accuracy and performance per unit than methods that do not make use of memory. The immediate advantage of the LSTM was due to its processing of a sequential input, 240-day return sequences, and the capture of complex temporal dependencies through the memory cell architecture comprised of forget, input, and output gates, which control the flow of information across time steps. Notably, their models captured complex short-term reversal patterns, volatility clustering, high-beta stock characteristics, and regime shifts beyond the reach of traditional statistical methods.

Much of the deep learning research focuses only on price histories. It overlooks macroeconomic factors. Our work extends these approaches using attention-based architectures, aiming to capture better temporal dependencies and macro-market relationships than both classical and recurrent baselines.

Transformer Architectures for Time Series

Transformer architectures have been introduced as a strong alternative to recurrent models for time series forecasting. However, the usage comes with significant computational and architectural challenges that need to be addressed. For instance, the quadratic complexity of a standard transformer restricts the application of long sequences, and its point-wise attention operation may not necessarily be the most effective way to interact with the temporal nature of time series data. In order to address the issues, Lim et al. (2021) proposed the Temporal Fusion Transformer (TFT) that utilizes the power of recurrent layers to model local temporal patterns and applies self-attention to capture long-range dependencies. The variable selection network of the model dynamically selects different inputs; hence, it can manage mixed input types without requiring manual feature specification [4].

On the other hand, Zhou et al. (2021) brought up the Informer, which is equipped with a ProbSparse attention mechanism that locates the most informative queries for a selective focus to reduce the complexity to $O(L \log L)$ and also allows generative predictions for the entire forecast horizon; therefore, inference is significantly faster for long-term forecasting [6]. Wu et al. (2021) addressed different patterns and seasonal components of time series by directly integrating decomposition into the Autoformer architecture and using an Auto-Correlation mechanism to accurately capture periodic patterns [5]. We adapt to the TFT method to the financial constraints.

Mixed-Frequency Data

Mixed-frequency data challenges occur when high-frequency variables hold important information for a target that is observed at a lower frequency [8]. MIDAS regression solves this issue by using polynomial weighting schemes to efficiently combine high-frequency observations into low-frequency predictions without averaging everything [8]. We build on MIDAS's polynomial weighting concept by applying it to attention. This lets the model learn the temporal weights instead of depending on preset functional forms.

Data Quality Issues in Financial Machine Learning

The IFC (2025) asserts, "poor data quality can lead to unreliable models, which may impede decision-making and prediction," when handling "data sparsity, noisy data, and dynamic environments" [10]. More traditional preprocessing techniques, like forward-fill, solve ignored data problems by perpetually carrying forward the last available observation for every empty time space. The method gives equal weight to a macro release from two months ago as to a real-time observation of market volatility for predicting market direction. Ignoring the time reduction of informational value results in an imbalance where outdated economic indicators are weighted as heavily as current financial signals, which systematically divides prediction errors [10].

Attention Mechanisms in Finance

Applications of temporal fusion transformers (TFTs) in financial markets show their use for stock price forecasting and predicting market trends [11, 12]. Recent studies also look at attention-based models. These models combine structured price data and unstructured news sentiment to better classify financial market movements, like the price direction of the FTSE 100 [13].

Methods

Data Pipeline

Our dataset spans January 1991 to October 2025, comprising daily S&P 500 returns and mixed-frequency predictor variables. Features include daily market indicators (the CBOE Volatility Index (VIX), 10-year Treasury yield, 10Y-2Y yield spread) and monthly macroeconomic releases (CPI-derived inflation with a roughly 14-day publication lag). VIX measures the expected 30-day S&P 500 volatility and is commonly used as a proxy to characterize market regimes.

Macroeconomic data presents a look-ahead bias risk. Indicators may be revised post-release, and naive implementations are at risk of using values unavailable at prediction time. We address this through vintage date alignment using the ALFRED database [14], retrieving only data vintages available on each historical date.

Data is partitioned chronologically: training (1991-2015), validation (2015-2020), and test (2020-2025). This ensures evaluation spans distinct market regimes, with validation including the COVID-19 disruption and test covering the post-pandemic period and 2022-2023 Fed tightening cycle (Figure 1).

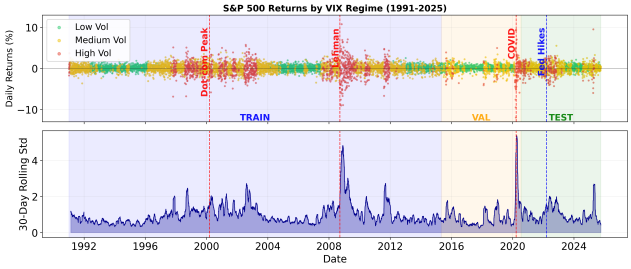

Figure 1. S&P 500 daily returns colored by VIX regime (green: low volatility, yellow: medium, red: high) with 30-day rolling volatility below. Background shading indicates train/validation/test splits. Major market events (dot-com peak, Lehman collapse, COVID-19, Fed rate hikes) cluster with high-volatility regimes, motivating regime-aware modeling approaches.

Model Architecture

We employ the Temporal Fusion Transformer (TFT) [4], which combines LSTM encoders for local temporal processing with interpretable multi-head attention for long-range dependencies. The model outputs predictions for seven quantiles (0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98), trained with quantile loss, which asymmetrically penalizes over- and under-prediction at each quantile level. We use 2 attention heads, a hidden dimension of 16, a hidden continuous size of 16, and a dropout of 0.1. The encoder observes 20 historical time steps. Training uses the Ranger optimizer (RAdam + Lookahead) with a learning rate of 5e-4, ReduceLROnPlateau scheduler (patience=4), gradient clipping of 0.1, and early stopping with patience of 10 epochs on validation loss.

For baseline comparison, we implement ARIMAX [15] and standard LSTM models on the same prediction task. The LSTM baseline uses sequences from the past 15 trading days. It features a three-layer LSTM with 128 hidden units, a 0.2 dropout rate, and two fully connected layers that map to the next-day return. The model is trained with MSE loss using Adam, which has a learning rate of 10^{-3} and a batch size of 64. The ARIMAX baseline determines its (p, d, q) order through grid search over $p \in [0, 3], d \in [0, 2], q \in [0, 3]$, applying AIC to the univariate return series. After that, it fits an $\mathrm{ARIMAX}(p, d, q)$ model with the same exogenous covariates using an 80/20 chronological train-test

split. Preliminary exploration found that TFT hidden dimensions above 18 consistently produced degenerate predictions biased toward exclusively positive returns regardless of market conditions. We term this failure mode "prediction collapse", characterized by low output variance and persistent unidirectional bias, and constrained subsequent experiments to smaller architectures.

Architectural Modifications

We extend the base TFT architecture with several modifications targeting the failure modes identified in our experiments. Figure 2 illustrates the overall system, showing the dual-head output structure and regime-aware attention mechanism.

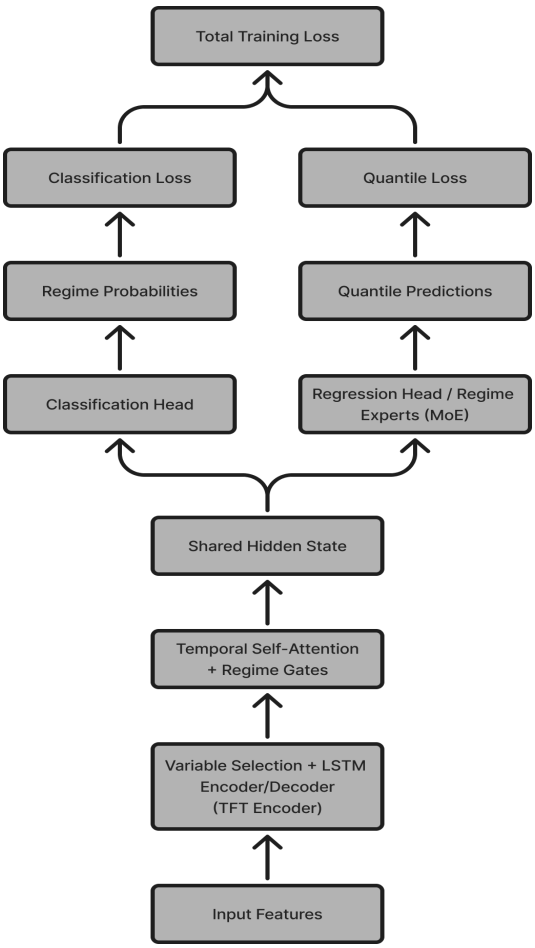

Figure 2. Modified TFT architecture. The base TFT encoder (Variable Selection + LSTM) feeds into temporal self-attention augmented with regime-conditional gates. The shared hidden state branches to a classification head (regime prediction) and regression head with optional mixture-of-experts. When enabled, both losses contribute to the total training objective.

Loss Function Penalties. We extend the standard quantile loss with regularization terms to prevent common predictive failure modes. A directional diversity penalty discourages unidirectional predictions by means of penalizing batches where more than 90% of predictions share the same sign:

$$\mathcal{L}_{div} = \lambda_{Div} \cdot \max(0, \hat{b} - \tau)^2 \tag{1}$$

where $\hat{b}$ is the observed directional bias and $tau = 0.9$ is the threshold. This regularization is based on empirical analysis showing real market returns never exceed this threshold over 30-day windows. Additional penalties target minimum prediction variance (anti-collapse) and temporal consistency between sequential predictions. These modifications are implemented by subclassing pytorch-forecasting's QuantileLoss [16].

Regime-Conditional Output Layer. Inspired by mixture-of-experts (MoE) architectures [17], we replace the TFT output layer with an MoE architecture, where parallel expert heads specialize in different market regimes. A router network assigns samples to experts based on either learned hidden state features or deterministic VIX thresholds. This is intended to test whether regime-specific prediction heads can improve performance when the model detects regime shifts in attention patterns but fails to adapt the output behavior accordingly.

Classification Head. Following multi-task approaches [18], we add a parallel classification head to the TFT encoder output, trained jointly with quantile regression via a combined loss:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{quantile} + \beta \mathcal{L}_{CE} \tag{2}$$

This serves as a diagnostic tool to assess if the classification head learns while regression fails, indicating that the encoder representation is informative, but the regression output layer cannot exploit it. We evaluate both direction prediction (binary up/down) and regime classification (VIX-based volatility states).

Regime-Aware Attention. We modify the interpretable multi-head attention mechanism to explicitly condition on the market regime. Each attention head is scaled by a regime-specific learned gate:

$$\tilde{a}_h = \sigma(g_{r,h}) \cdot a_h \tag{3}$$

where $g_{r,h}$ is the learned gate regime r and head h. This allows the heads to specialize, for example, by one head being amplified in high-volatility periods, and another during times of low-volatility. Regime assignment uses VIX thresholds for interpretability. This adds minimal parameters (4 total for 2 regimes with 2 attention heads) while providing a direct mechanism for regime-dependent attention behavior.

Evaluation Framework

We evaluate models using quantile loss, directional accuracy, and Sharpe ratio of a long-only strategy, alongside

standard regression metrics (RMSE, MAE) and financial performance measures (hit rate, maximum drawdown).

Beyond aggregate metrics, we monitor training dynamics by tracking gradient flow through each layer, prediction variance, and directional bias over epochs. This revealed that output layer gradients collapse early in training while encoder layers continue learning (Figure 3), a pattern consistent across experiments that motivated our architectural modifications. We classify evaluation windows by prediction quality (healthy vs. degraded/collapsed) based on variance, directional bias, and correlation with actual returns.

To assess robustness across market regimes, we employ rolling evaluation with 10-year training, 1-year validation, and 1-year test windows, stepping forward annually. This produces performance distributions across different market conditions (e.g., 2016 bull market vs. 2022 bear market) rather than single-point estimates from one fixed split.

We found that validation loss converged tightly (0.39-0.40) across experiments regardless of downstream performance. We instead save and evaluate checkpoints based on multiple criteria, including directional accuracy, prediction diversity, and composite metrics combining accuracy with distribution matching.

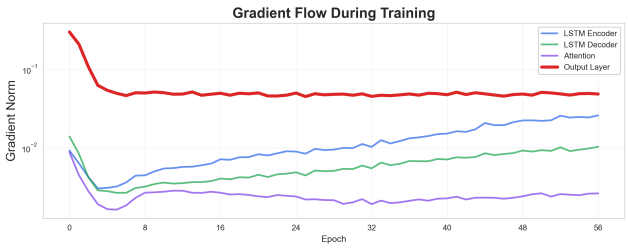

Figure 3. Gradient norm by layer during training. The output layer (red) collapses within the first 10 epochs and remains flat, while LSTM encoder and decoder gradients (blue, green) continue increasing throughout training. This pattern, consistent across experiments, indicates the encoder learns meaningful representations that the output layer fails to translate into predictions.

Current Status and Next Steps

We are currently evaluating regime-aware attention modifications and weekly frequency models. Preliminary classification experiments show that the encoder learns regime structure, achieving 100% accuracy on VIX-based regime classification while direction prediction remains at base rate. This suggests daily returns lack a predictable signal despite meaningful feature representations. Remaining work includes multi-horizon forecasting, alternative feature configurations, multi-target prediction across constituents, and comparison with ARIMAX and LSTM baselines.

References

[1] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970. 1

[2] P. Chen and C.-H. Shen, "Regime-switching models: Capturing structural changes in time series," in *Proc. SAS Global Forum*, 2018. 1

[3] A. Ang and A. Timmermann, "Regime changes and financial markets," *Annual Review of Financial Economics*, vol. 4, no. 1, pp. 313–337, 2012. 1

[4] B. Lim, S. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021. 1, 2, 3

[5] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 22419–22430, Curran Associates, Inc., 2021. 2

[6] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *CoRR*, vol. abs/2012.07436, 2020. 2

[7] P. Liu, B. Wu, Y. Hu, N. Li, T. Dai, J. Bao, and S. Xia, "Timebridge: Non-stationarity matters for long-term time series forecasting," in *Proc. International Conference on Machine Learning (ICML)*, 2025. 2

[8] E. Ghysels, P. Santa-Clara, and R. Valkanov, "The midas touch: Mixed data sampling regression models," *Journal of Financial Economics*, vol. 78, no. 2, pp. 213–244, 2005. 2, 3

[9] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018. 2

[10] International Finance Corporation, "Leveraging machine learning to enhance credit data quality," technical report, World Bank Group, 2025. 3

[11] "Temporal fusion transformers for enhanced multivariate time series forecasting in stock market prediction," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 1–8, 2024. TFT obtained a remarkable SMAPE of 0.0022. 3

[12] Y. Chen *et al.*, "Economic system forecasting based on temporal fusion transformers: Multi-dimensional evaluation and cross-model comparative analysis," *Neurocomputing*, vol. 549, p. 126433, 2023. 3

[13] Y. Pei, J. Cartlidge, A. Mandal, D. Gold, E. Marcilio, and R. Mazzon, "Cross-modal temporal fusion for financial market forecasting," in *Proc. European Conference on Artificial Intelligence (ECAI)*, (Bologna, Italy), 2025. Manuscript accepted to PAIS at ECAI-2025. 3

[14] F. R. B. of St. Louis, "ALFRED: Archival Federal Reserve Economic Data." https://alfred.stlouisfed.org/, 2025. 3

[15] G. E. P. Box and G. C. Tiao, "Intervention analysis with applications to economic and environmental problems," *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 70–79, 1975. 3

[16] J. Beitner and contributors, "PyTorch Forecasting: QuantileLoss_pkg Documentation." https://pytorch-forecasting.readthedocs.io/en/stable/api/pytorch_forecasting.metrics._quantile_pkg._quantile_loss_pkg.QuantileLoss_pkg.html, 2025. 4

[17] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *CoRR*, vol. abs/1701.06538, 2017. 4

[18] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997. 4

AI Detection Report

All sections of this report were checked using the Grammarly AI detector.

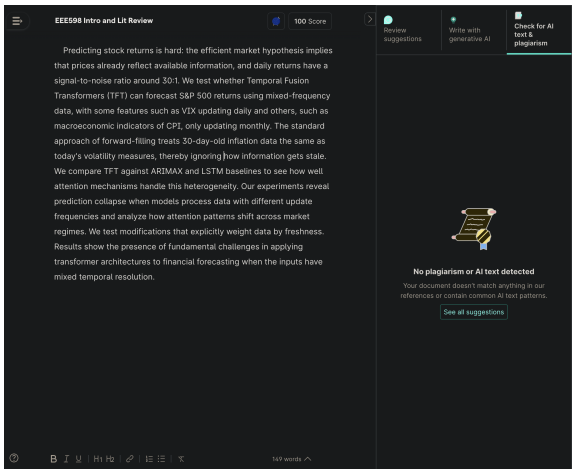

Figure 4. Abstract: 0% AI-generated

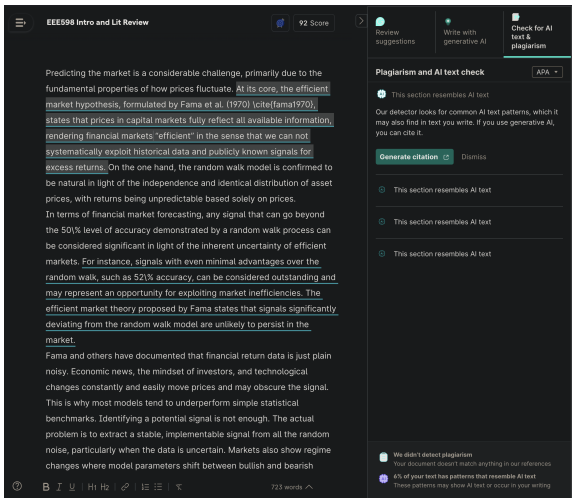

Figure 5. Introduction: 6% AI-generated

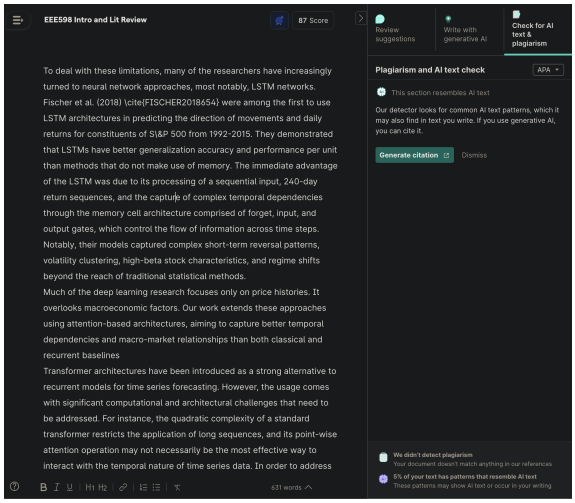

Figure 6. Related work: 5% AI-generated

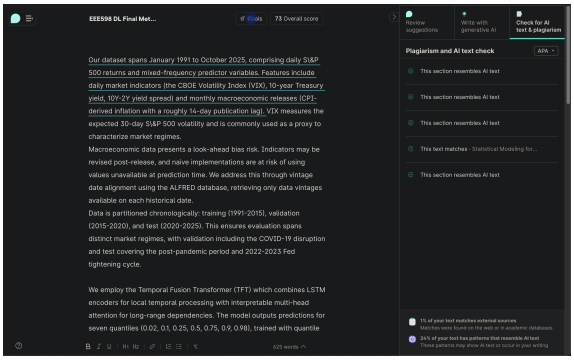

Figure 7. Methods: 24% AI-generated