Repurposing Language Models for FX Volatility Forecasting: A Data-Efficient and Context-Aware Approach

Quoc Anh Nguyen
Imperial College London
United Kingdom
qn24@imperial.ac.uk

Ce Guo
Imperial College London
United Kingdom
c.guo@imperial.ac.uk

Wayne Luk
Imperial College London
United Kingdom
w.luk@imperial.ac.uk

Abstract

Forecasting foreign exchange (FX) volatility is a critical task in finance. The high data demand of deep learning models creates a practical challenge, because data availability is limited by constantly shifting market dynamics. To address this problem, we introduce Vola-BERT, a novel predictive model that solves this data-efficiency problem by repurposing a pre-trained bidirectional language model. Our core insight is that the model's value lies not in linguistic knowledge, but in its sophisticated capacity to model complex sequential patterns, which we adapt to the financial time-series domain. Vola-BERT is further enhanced by a new semantic conditioning framework that transparently integrates non-time-series data, such as high-impact economic events. This mechanism transforms the model from a black-box forecaster into an interpretable analytical tool capable of directly quantifying the impact of real-world market drivers. Our empirical study against 11 baseline models shows that Vola-BERT achieves state-of-the-art forecasting accuracy, consistently outperforming baselines under both data-rich and data-scarce conditions. This advantage is particularly pronounced in the data-scarce scenario, where it maintains high accuracy using only 10 percent of the training data. This work demonstrates that repurposing language models is a promising, data-efficient direction for context-aware numerical analysis.

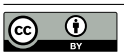
CCS Concepts

• **Applied computing** → **Forecasting**; **Decision analysis**; • **Information systems** → **Business intelligence**; • **Computing methodologies** → *Natural language generation*; **Temporal reasoning**; **Transfer learning**.

Keywords

Volatility Forecasting, Large Language Models, Data-Scarce Learning, Explainable AI, Forex, Time Series Forecasting

ACM Reference Format:
Quoc Anh Nguyen, Ce Guo, and Wayne Luk. 2025. Repurposing Language Models for FX Volatility Forecasting: A Data-Efficient and Context-Aware Approach. In *6th ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3768292.3770386



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICAIF '25, Singapore, Singapore
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2220-2/2025/11
https://doi.org/10.1145/3768292.3770386

1 Introduction

The foreign exchange market is the world's largest financial market, with a daily trading volume exceeding $7.5 trillion [6]. A key characteristic of this market is the rapid fluctuation of currency exchange rates, driven by a complex interplay of monetary policies, macroeconomic indicators, and geopolitical events. Accurately forecasting this volatility is essential for numerous financial applications, including risk management, derivative pricing, and portfolio optimization [1, 10].

Machine learning, particularly deep learning, has been widely applied to financial time series forecasting [17, 31]. Recurrent and Transformer-based models effectively capture long-term temporal patterns [24, 28], but their dependence on large datasets presents a key challenge in finance. Specifically, constantly shifting market dynamics mean that older historical data quickly loses relevance to future conditions [15, 27]. The resulting scarcity of useful training data reduces a model's ability to learn emerging patterns.

Pre-trained language models have demonstrated remarkable transfer learning capabilities in the finance sector [23]. Our work builds on the paradigm of repurposing pre-trained models for time series analysis [32, 35]. Our core insight is that the knowledge being transferred from a text-based LLM is not linguistic, but rather its learned capacity to model complex and long-range sequential patterns. We adapt this general capability for the specific and demanding domain of multivariate FX volatility forecasting. While prior work has focused on patching for univariate series or used causal LLMs [8, 26, 35], our approach is tailored to the multivariate and non-causal nature of the foreign exchange market. Our major contributions include:

(1) A novel predictive model, Vola-BERT, that adapts a bidirectional LLM, which is originally trained from text data, to achieve state-of-the-art accuracy in FX volatility time-series forecasting.
(2) A framework for semantic conditioning that integrates discrete, non-time-series data—such as economic events—into the forecasting process, improving accuracy and transforming the model into an interpretable analytical tool.
(3) An empirical validation against 11 baselines on four currency pairs, demonstrating Vola-BERT's effectiveness and its particular advantage in data-scarce settings, which validates our transfer-learning approach.

Our empirical study shows that Vola-BERT consistently outperforms the baseline models in both data-rich and data-scarce settings. The advantage of our approach is particularly pronounced when models are trained on only 10 percent of the historical data.

Our open-source implementation, including pre-trained model weights, training routines, and scripts to replicate experiments, is publicly available at:

https://github.com/sonnets-project/vola-bert

2 Background and Related Work

Our work builds on three primary areas of research: FX volatility modeling, deep learning for time series, and LLM-based forecasting. A key object of study in finance is realised volatility, which measures historical price variation. While traditional methods often use the variance of returns, range-based estimators like the Parkinson number [30] have been shown to be more statistically efficient and less noisy estimators of volatility [10]. These estimators use the high and low prices within a period to provide a more accurate volatility measure.

The field of time series forecasting has progressed significantly from classical statistical models to advanced deep learning architectures. Early machine learning approaches included models like Support Vector Regressors (SVR) [5, 33] and Gradient Boosted Decision Trees (GDBT) [13], which remain strong baselines. The advent of deep learning popularized Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) [19] and Gated Recurrent Units (GRU) [9], which became the standard for sequence modeling tasks. More recently, Transformer-based models have demonstrated superior performance by using self-attention mechanisms to capture long-range dependencies more effectively. Architectures like PatchTST [28] and iTransformer [24], which treat time series as sequences of patches, have set new state-of-the-art benchmarks. For a broader perspective, we also include Generalized Autoregressive Conditional Heteroskedasticity (GARCH) [7], a popular econometric model for volatility forecasting, as a traditional baseline. These powerful models form the core of our baseline comparisons.

Applying Large Language Models (LLMs) to numerical time series is an emerging and promising research frontier. Three main paradigms have been explored. The first is zero-shot prompting, where a pre-trained LLM is used directly for forecasting without any repurposing by framing the task in natural language [16]. A second approach is embedding alignment, which seeks to align the embedding spaces of textual and time series data to use the LLM's semantic understanding [32]. The third paradigm, which is the focus of our work, is Supervised Fine-Tuning (SFT), a method that adapts a pre-trained LLM's weights for a specific forecasting task. Models like GPT4TS [35] and AutoTimes [26] have pioneered SFT by freezing the core self-attention blocks and repurposing only the normalization layers and embeddings. Our work builds upon the SFT paradigm by tailoring it specifically for the financial domain and introducing a novel mechanism for explainability. This aims to tackle the challenge of constant FX market shifts; we acknowledge that, unlike continual learning [14] that works on the "time and update process" axis, our approach relies on architectural innovations for effective learning from limited recent data. To the best of our knowledge, no prior work has demonstrated the effectiveness of SFT for FX volatility forecasting, particularly in data-scarce and explainable contexts.

3 Proposed Approach

3.1 Motivation

Our main goal is to accurately forecast foreign exchange (FX) volatility. We specifically focus on predicting the Parkinson volatility [30], a statistically efficient measure that uses the high and low prices of an asset to better capture its price range. We compute our target variable, the realised volatility series V_t, using the Parkinson estimator:

$$V_t = \frac{(\ln(H_t) - \ln(L_t))^2}{4\ln(2)}, \tag{1}$$

where H_t and L_t are the high and low prices in interval t. Our multi-horizon forecasting task is to learn a mapping function, $f_{\text{Vola-BERT}}$, that takes a lookback window of historical numerical data along with a set of discrete contextual variables to predict the next H future volatility values. Formally, we define this task as:

$$\{\hat{V}_{t+1}, \ldots, \hat{V}_{t+H}\} = f_{\text{Vola-BERT}}(\{X_{t-L+1}, \ldots, X_t\}, Z_t) \tag{2}$$

where $\{\hat{V}_{t+1}, \ldots, \hat{V}_{t+H}\}$ is the sequence of forecasted volatilities. The inputs consist of the multivariate time series $\{X_{t-L+1}, \ldots, X_t\}$, where each $X_i \in \mathbb{R}^D$, and Z_t, which represents a set of discrete semantic variables providing external context at time t. In this work, Z_t includes information such as the market trading session and scheduled high-impact economic events.

We repurpose a large language model (LLM) that is originally trained on text. Using a text-based model to predict numbers might seem unusual because the domains are so different. Text models learn about language and meaning, while financial forecasting relies on statistics and math. However, we argue that the key knowledge being transferred is not about language, but about how to process complex sequential data. Our main idea is that the self-attention mechanism, having learned from vast amounts of text, becomes a general tool for finding patterns in any sequence. This approach is supported by recent studies showing that text-based LLMs can be successfully adapted for time series forecasting [16, 20, 35].

Our transfer-learning philosophy is very different from training a new model from scratch, like Chronos [4], and offers two key benefits. First, it is much cheaper and faster, as it avoids the huge computational cost of pre-training. Second, we argue that this strategy is particularly effective under data-scarce settings, as repurposing pre-trained models provides access to their representation capacity, making accurate time-series forecasting possible even with limited data. We will further validate and discuss this accuracy advantage in Section 4.

3.2 Model Architecture and Training

The design of Vola-BERT is guided by three principles tailored to the realities of FX market analysis. First, the model must process a rich, multivariate context, recognizing that volatility is driven by a complex interplay of market features, not a single price series. Second, it requires a holistic, non-causal perspective, enabling it to weigh all signals across a time window simultaneously, much like a human analyst. Finally, it must combine capital efficiency with contextual explainability, using pre-trained components while allowing semantic tokens to quantify the impact of external drivers like economic events.

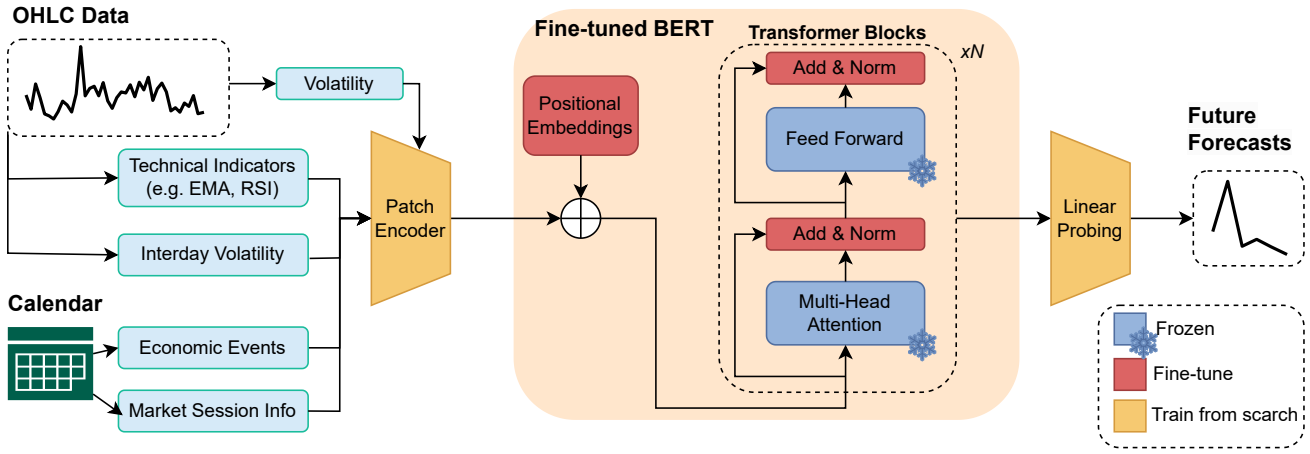


Figure 1: Vola-BERT's architecture. Past realised volatility and other predictors are first transformed into patch embeddings that serve as input tokens for the LLM. These embeddings are processed by a pretrained BERT backbone, where we freeze the core self-attention and feedforward weights while tuning the positional embeddings and layer normalization layers. Finally, lightweight linear probes over the volatility, market session, and event tokens generate future intraday volatility forecasts.

The Vola-BERT architecture, illustrated in Figure 1, follows a three-stage pipeline: input encoding, sequence processing, and forecasting.

First, in the input encoding stage, the model processes two distinct types of data. All numerical time series inputs—including past realized volatility, technical indicators, and interday volatility—are fed into a Patch Encoder. This module transforms the entire time series of each feature into a single, distinct embedding vector, in a one-feature-one-token manner. Separately, the categorical calendar data, such as market session info and economic events, are converted into corresponding learnable embeddings. Crucially, these two sets of embeddings are then combined, and positional embeddings are added to the sequence to retain temporal information.

Second, this combined sequence of tokens is processed by a pretrained BERT backbone [11]. This stage is responsible for understanding the complex relationships between all the different inputs. Through its self-attention mechanism, BERT allows each feature token, corresponding to a key market indicator or signal, to attend to and integrate information from others, yielding contextualized embeddings that capture the holistic market context.

Finally, the forecasting stage uses lightweight linear probes to generate the final prediction. This forecasting head does not simply use the final output of the BERT backbone. Instead, it selectively takes the output hidden states corresponding to the key input tokens—specifically the main volatility token, the market session token, and the event token. These specific representations are concatenated and passed through a final linear layer to produce the future volatility forecasts. This mechanism forces the model to integrate its understanding of the numerical time series with the high-level semantic context to make its prediction.

As shown by the color-coding in the diagram, we employ a parameter-efficient fine-tuning strategy to train our model. The core components of BERT's Transformer Blocks, the Multi-Head Attention and Feed Forward layers, remain frozen to preserve their powerful, pre-learned sequence processing capabilities. Only the Layer Normalisation and positional embedding layers are fine-tuned, allowing the model to adapt to the financial data domain with minimal computational cost. In our open-source implementation, we use the standard Adam optimizer [21] with a small learning rate to avoid catastrophic forgetting. We also apply gradient clipping based on the L2 norm to regulate the magnitude of parameter updates for stable training. This is particularly important in our setup, as we are repurposing an LLM on a non-text modality, where mismatches within the feature space can cause anomalous gradients. Furthermore, FX volatility frequently exhibits abrupt and significant spikes, which can potentially lead to gradient explosion.

Regarding the input sequence, we use 20 supplementary features in addition to the realised volatility itself. These features capture the dynamics of the exchange rate market and fall into three main groups. The first feature group is a set of widely-used technical indicators such as Bollinger Bands, EMA, and RSI, which are standard tools used to model short-term trend and momentum dynamics in financial time series [31]. The second feature group incorporates information about scheduled economic events, which is critical because macroeconomic news announcements are known to be primary drivers of immediate market volatility [3]. The final group adds features for interday volatility by including the volatility from the same time over the previous five days. This allows the model to learn the strong, cyclical intraday patterns common in FX markets, such as increased activity during the London-New York session overlap [2].

3.3 Semantic Conditioning

A key limitation of standard time series models is their inability to incorporate discrete, external information, such as market regimes, human expert analysis, or other categorical signals. To address this

gap, Vola-BERT introduces a flexible and extensible framework for what we term "semantic conditioning". The core idea is to augment the numerical input with special, learnable tokens that represent any external categorical variable. Our proposed framework sets itself apart from the typical explain-then-predict paradigm LLM-based approaches like Chain-of-Thought, as we opt to fix the set of explanatory factors and quantify their impact on the model's forecast with counterfactual examples [34]. This allows the model to be conditioned on specific real-world contexts, a feature that, to our knowledge, is not natively supported by other LLM-based time-series forecasting models.

Before processing by the main LLM backbone, one or more learnable tokens are appended to the sequence of numerical feature embeddings. Each type of semantic information has its own vocabulary and a corresponding embedding layer in the model. This design is highly flexible and extendable. For instance, to incorporate a new signal like analyst sentiment [29], a user would first add this categorical data as a new feature. They would then define its vocabulary, for example 'Positive' and 'Negative', within the dataset class and add a corresponding embedding layer to the model architecture. The resulting embedding would then be integrated into the model's forward pass.

This mechanism transforms the model from a simple predictor into an analytical tool. It enables a form of causal analysis to quantify the impact of specific events. As a practical demonstration of this framework, we implemented two types of semantic tokens highly relevant to the FX market:

(1) Market Session Tokens: A vocabulary representing the market session at the time of the forecast (e.g., `[LONDON_OPEN]`, `[NY_LONDON_OVERLAP]`). This allows the model to learn distinct volatility patterns associated with different trading zones.
(2) Economic Event Tokens: A vocabulary indicating if a high-impact economic news event is scheduled (e.g., `[NO_EVENT]`, `[HIGH_IMPACT_NEWS]`). This enables the model to learn the heightened volatility associated with major announcements.

A detailed guide for semantic conditioning with examples is provided in our open-source repository.

3.4 Comparison with Existing Models

To clarify Vola-BERT's unique design choices, we provide a direct comparison against other prominent LLM-based models in Table 1. Our approach aligns with the capital-efficient, repurposing paradigm of models like GPT4TS and AutoTimes, standing in sharp contrast to the computationally intensive train-from-scratch approach of Chronos. This shared philosophy validates the use of pre-trained text models for numerical tasks. However, the table also highlights the key reasons why we developed Vola-BERT instead of adapting a framework like GPT4TS.

A fundamental difference lies in the LLM architecture. We deliberately chose a bidirectional BERT-style model over the causal, GPT-style models used as a basis for GPT4TS and AutoTimes. This choice is critical for how the model analyzes market data. A causal model processes information sequentially, similar to reading a sentence one word at a time. A bidirectional model, however, considers the entire feature set at once, allowing it to form a complete, holistic picture of the market state.

The use of bidirectional models is crucial in the FX market, where signals are often conflicting and must be interpreted together. For example, a momentum indicator like the RSI at the end of a lookback period might suggest a currency is overbought. At the same time, a trend indicator from the beginning of the period might still show a buy signal. A bidirectional model can see both signals simultaneously and learn to weigh them against each other to assess the true market condition. A causal model would process the early buy signal without the full context of the later overbought signal.

Also, Vola-BERT's bidirectional nature is essential for its explainability framework. When the model receives a semantic token, it does not just affect future predictions. It changes the interpretation of the entire lookback window. The model can learn that the price action leading up to the announcement is not random noise, but pre-event positioning. This ability to re-contextualize the entire input based on a single piece of information is a key advantage of our model for this task. This unique combination of a purpose-chosen architecture, multivariate capability, and built-in explainability defines Vola-BERT's contribution to the field.

Moreover, high-frequency forecasting presents a common challenge, as fixed intervals may not align with market events. Practitioners note that significant news can influence volatility in ways not captured by discrete steps. Our Vola-BERT architecture directly addresses this issue through its Semantic Conditioning framework. The model is conditioned on tokens like `[HIGH_IMPACT_NEWS]` to anticipate atypical volatility patterns.

4 Evaluation

4.1 Experimental Setup

Baseline Models. To evaluate Vola-BERT, we conduct an empirical study against a diverse set of 11 baseline models. This group provides a robust comparison across different modeling paradigms and includes: a traditional econometric model, namely GARCH(1,1) [7]; established machine learning and recurrent models, including GBDT [12], SVR [33], LSTM [19], and GRU [9]; modern Transformer-based architectures, including PatchTST [28], iTransformer [24], and Reformer [22]; and other recent LLM-based forecasters, such as Chronos [4], GPT4TS [35], and AutoTimes [26].

Dataset and Pre-processing. Our study uses 30-minute intraday OHLC data for the GBP/USD, EUR/USD, USD/CHF, USD/JPY currency pairs, retrieved from FirstRate Data. The dataset comprises over 60,000 timepoints spanning from January 2, 2020 to December 31, 2024, which is entirely after the BERT model's pretraining cutoff (October 2018), ensuring no information leakage. To incorporate fundamental context, we use an economic calendar from FXStreet to identify macroeconomic events affecting the considered currencies. These events are classified as low, medium, or high impact based on their expected market influence. We restrict predictions to intraday volatility, that is, within the 240 timepoints spanning a typical trading week (5 business days x 48 half-hour intervals), thereby excluding weekends and holidays when the FX market is closed. The target variable for our forecasting task is the Parkinson volatility.

Table 1: Comparison of Vola-BERT with other LLM-based Time Series Forecasting Models

Model	Vola-BERT (Ours)	Chronos [4]	GPT4TS [35]	AutoTimes [26]
Year	2025	2024	2023	2024
Core Paradigm	Repurpose Text LLM	Train from Scratch	Fine-tune Text LLM	AR forecasting via Text LLM
LLM Architecture	Bidirectional (BERT)	Encoder-Decoder (T5)	Causal (GPT-2)	Causal (LLaMA, GPT-2)
Pre-training Data	Text, then FX Data	Time Series Only	Text, then TS Data	Text, then TS Data
Fine-Tuned Comp.	Norm/Embed/Head	All (From Scratch)	Norm/Embed/Head	Embed/Head
Input Method	Patching (Multivariate)	Quantization (Univariate)	Patching (Univariate)	Patching (Univariate)
Model Scope	Specialist (FX Volatility)	Generalist (Any TS)	Generalist (Fine-tuned)	Generalist (Fine-tuned)
Explainability	Explicit (Semantic Tokens)	None (Implicit)	None (Implicit)	None (Implicit)
Training Cost	Low	Very High	Low	Very Low (Frozen LLM)

Table 2: Forecasting Accuracy on Full-Data Setting (100% of training data). The forecast horizons are denoted by Lookback → Forecast length. Best results are in bold, second-best are underlined.

		GBP/USD			EUR/USD			USD/CHF			USD/JPY		
Method	**Metric**	**40→5**	**50→10**	**60→20**	**40→5**	**50→10**	**60→20**	**40→5**	**50→10**	**60→20**	**40→5**	**50→10**	**60→20**
GARCH(1,1) [7]	rMAE	0.925	0.913	0.905	0.960	0.976	1.000	0.888	0.859	0.869	0.864	0.810	0.829
	rMSE	0.669	0.645	0.640	0.717	0.712	0.736	0.647	0.620	0.650	0.649	0.622	0.659
GBDT [12]	rMAE	0.843	0.678	0.630	0.765	0.652	0.608	0.783	0.670	0.663	0.810	0.713	0.708
	rMSE	0.611	0.480	0.463	0.570	0.461	0.454	0.565	0.476	0.501	0.604	0.529	0.542
SVR [33]	rMAE	0.742	0.639	0.602	0.747	0.617	0.584	0.722	0.647	0.651	0.777	0.673	0.697
	rMSE	0.622	0.511	0.483	0.637	0.488	0.481	0.587	0.501	0.541	0.650	0.551	0.592
LSTM [19]	rMAE	0.691	0.629	0.601	0.666	0.595	0.564	0.682	0.640	0.631	0.700	0.661	0.679
	rMSE	0.540	0.478	0.460	0.559	0.481	0.457	0.534	0.477	0.496	0.589	0.544	0.583
GRU [9]	rMAE	0.686	0.628	0.594	0.667	0.597	0.561	0.685	0.625	0.635	**0.690**	0.645	0.675
	rMSE	0.547	0.476	0.457	0.548	0.475	0.456	0.538	0.470	0.491	0.590	0.540	0.571
PatchTST [28]	rMAE	0.714	0.629	0.598	0.694	0.615	0.585	0.712	0.633	0.637	0.736	0.656	0.673
	rMSE	0.609	0.500	0.477	0.594	0.503	0.487	0.587	0.497	0.519	0.614	0.534	0.559
iTransformer [24]	rMAE	0.686	0.602	0.574	0.680	0.590	0.564	0.695	0.612	0.614	0.709	0.628	0.658
	rMSE	0.577	0.474	0.456	0.577	0.476	0.467	0.561	0.474	0.498	0.587	0.510	0.545
Reformer [22]	rMAE	0.761	0.708	0.630	0.764	0.693	0.619	0.783	0.709	0.668	0.781	0.703	0.722
	rMSE	0.658	0.579	0.505	0.657	0.575	0.509	0.654	0.576	0.541	0.654	0.582	0.607
Chronos [4]	rMAE	0.851	0.753	0.721	0.868	0.760	0.723	0.854	0.775	0.771	0.834	0.757	0.784
	rMSE	0.788	0.671	0.639	0.852	0.684	0.643	0.792	0.701	0.722	0.756	0.672	0.734
GPT4TS [35]	rMAE	0.708	0.621	0.592	0.695	0.613	0.581	0.715	0.634	0.630	0.738	0.653	0.675
	rMSE	0.589	0.492	0.470	0.590	0.495	0.484	0.579	0.495	0.516	0.613	0.532	0.562
AutoTimes [26]	rMAE	0.737	0.642	0.601	0.731	0.617	0.588	0.752	0.650	0.643	0.773	0.686	0.682
	rMSE	0.633	0.511	0.488	0.632	0.507	0.494	0.630	0.516	0.529	0.646	0.561	0.572
Vola-BERT	rMAE	**0.663**	**0.590**	**0.565**	**0.656**	**0.582**	**0.549**	**0.670**	**0.603**	**0.603**	0.693	**0.623**	**0.650**
	rMSE	**0.536**	**0.455**	**0.444**	**0.534**	**0.458**	**0.449**	**0.529**	**0.454**	**0.485**	**0.566**	**0.506**	**0.539**

Evaluation Metric. We evaluate all models across three forecasting horizons: (Lookback = 40, Future = 5), (Lookback = 50, Future = 10), and (Lookback = 60, Future = 20). Unlike one-step-ahead forecasting, predicting volatility over multiple future timepoints aligns better with financial applications, where risk management and portfolio decisions often operate over defined time windows, and thus require forecasts of market dynamics over near-term horizon. For example, during macroeconomic events, market participants are interested not just in the upcoming volatility spikes, but also in how they evolve and return to normal levels, generally forming an inverted V-shape pattern. Our selection of horizons is guided by existing time-series forecasting literature, where lookback-to-future ratios are typically set between 2:1 and 8:1 to assess models' predictive ability in short-term and extended periods [8, 24, 28], and is also within 240 timepoints due to the intraday nature of our volatility forecasting problem. Across all considered horizons, our primary evaluation metrics are the relative Mean Absolute Error (rMAE) and relative Mean Squared Error (rMSE). This is calculated by taking the error metric (MAE/MSE) of a model's forecast and

Table 3: Forecasting Accuracy on Data-Scarce Setting (10% of training data). The forecast horizons are denoted by Lookback → Forecast length. Best results are in bold, second-best are underlined.

		GBP/USD			EUR/USD			USD/CHF			USD/JPY		
Method	**Metric**	**40→5**	**50→10**	**60→20**	**40→5**	**50→10**	**60→20**	**40→5**	**50→10**	**60→20**	**40→5**	**50→10**	**60→20**
GARCH(1,1) [7]	rMAE	0.925	0.913	0.905	0.960	0.976	1.000	0.888	0.859	0.869	0.864	0.810	0.829
	rMSE	0.669	0.645	0.640	0.717	0.712	0.736	0.647	0.620	0.650	0.649	0.622	0.659
GBDT [12]	rMAE	1.173	0.748	0.686	1.043	0.725	0.662	0.864	0.706	0.700	0.848	0.717	0.774
	rMSE	0.879	0.520	0.492	0.739	0.505	0.493	0.610	0.484	0.511	0.629	0.534	0.589
SVR [33]	rMAE	0.872	0.701	0.632	0.790	0.667	0.622	0.785	0.676	0.668	0.848	0.734	0.776
	rMSE	0.625	0.490	0.475	0.609	0.520	0.506	0.597	0.502	0.520	0.717	0.599	0.640
LSTM [19]	rMAE	0.848	0.780	0.823	0.859	0.819	0.726	0.782	0.711	0.696	0.806	0.761	0.774
	rMSE	0.632	0.545	0.616	0.657	0.583	0.530	0.582	0.526	0.533	0.709	0.647	0.660
GRU [9]	rMAE	0.813	0.787	0.749	0.836	0.731	0.627	0.753	0.677	0.693	0.785	0.744	0.772
	rMSE	0.609	0.546	0.536	0.634	0.537	0.486	0.564	0.494	0.544	0.669	0.631	0.645
PatchTST [28]	rMAE	0.725	0.638	0.604	0.713	0.631	0.600	0.730	0.651	0.651	0.754	0.679	0.701
	rMSE	0.605	0.503	0.484	0.612	0.514	0.503	0.597	0.503	0.523	0.624	0.545	0.572
iTransformer [24]	rMAE	0.713	0.620	0.588	**0.703**	0.613	**0.580**	0.725	0.638	0.645	0.743	0.657	**0.682**
	rMSE	0.587	0.484	0.465	0.594	0.490	0.478	0.591	0.492	0.514	0.608	0.522	**0.553**
Reformer [22]	rMAE	0.800	0.755	0.674	0.818	0.776	0.704	0.816	0.747	0.708	0.798	0.724	0.745
	rMSE	0.684	0.642	0.571	0.718	0.667	0.596	0.694	0.619	0.585	0.675	0.600	0.630
Chronos [4]	rMAE	0.851	0.753	0.721	0.868	0.760	0.723	0.854	0.775	0.771	0.834	0.757	0.784
	rMSE	0.788	0.671	0.639	0.852	0.684	0.643	0.792	0.701	0.722	0.756	0.672	0.734
GPT4TS [35]	rMAE	0.725	0.631	0.598	0.711	0.629	0.598	0.731	0.646	0.644	0.769	0.677	0.698
	rMSE	0.597	0.493	0.477	0.602	0.510	0.501	0.592	0.496	0.518	0.636	0.547	0.573
AutoTimes [26]	rMAE	0.785	0.658	0.620	0.792	0.665	0.607	0.790	0.690	0.665	0.780	0.712	0.728
	rMSE	0.677	0.515	0.503	0.687	0.542	0.510	0.652	0.534	0.542	0.645	0.578	0.599
Vola-BERT	rMAE	**0.698**	**0.620**	**0.584**	0.705	**0.611**	0.582	**0.707**	**0.633**	**0.636**	**0.721**	**0.651**	0.689
	rMSE	**0.555**	**0.475**	**0.457**	**0.573**	**0.481**	**0.470**	**0.557**	**0.482**	**0.503**	**0.590**	**0.518**	0.558

normalizing it by the corresponding metric of a naive baseline that simply forecasts the last observed volatility value.

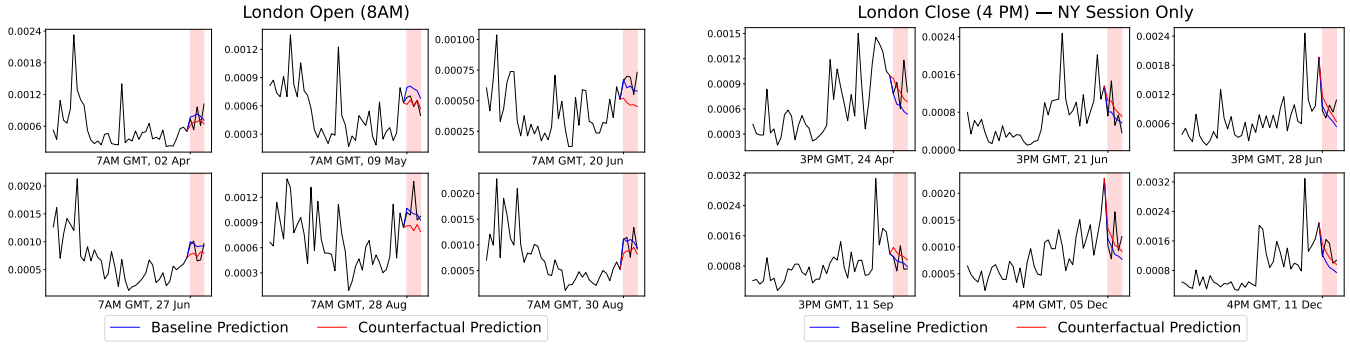
Hyperparameter Settings. To ensure fair comparisons, the architectural hyperparameters for all baseline models were tuned based on their performance on the validation set. For recurrent models like LSTM and GRU, we considered setups with 1 or 2 hidden layers and hidden sizes of 16, 32, or 64. For Transformer models trained from scratch, such as PatchTST and iTransformer, we set the model dimension d_{model} to 16, 32, or 64, the feedforward dimension d_{ff} to twice d_{model}, and varied the number of blocks between 2, 4, and 6. For our proposed model, Vola-BERT, we experimented with using 2, 4, or 6 layers from the pre-trained BERT backbone. All other LLM-based methods are also evaluated using 2, 4, or 6 transformer blocks extracted from the selected LLM.

4.2 Forecasting Accuracy

To test our model under different conditions, we partition the dataset in two ways. The first is a Full-Data Setting, which measures model performance when a large amount of historical data is available. For this, the dataset is divided chronologically into a 70 percent training partition, a 15 percent validation partition, and a final 15 percent held-out test partition. The second is a Data-Scarce Setting, which evaluates performance under data constraints to simulate a scenario where only recent data is trusted. Here, the training set is restricted to only the most recent 10 percent of the original partition.

The results for the full-data and data-scarce settings are summarized in Table 2 and Table 3, respectively. A comparative analysis of these results reveals several key findings:

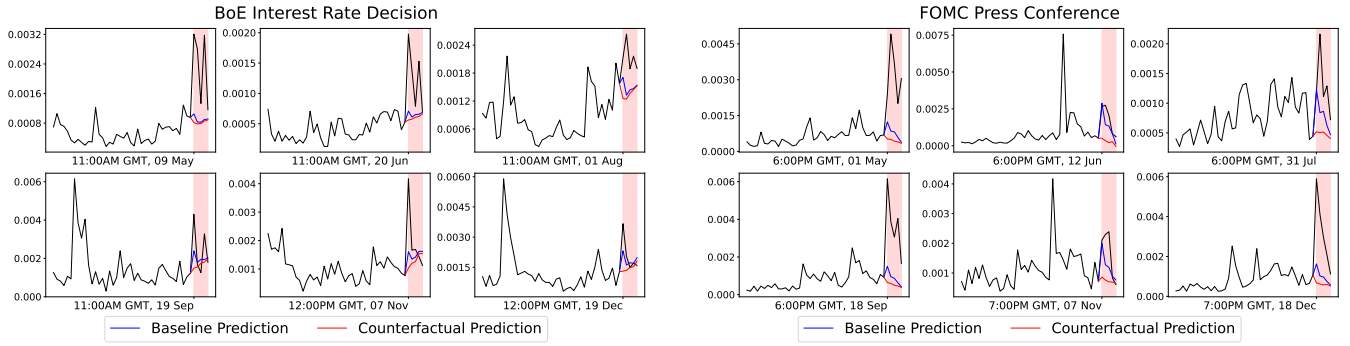
(1) Vola-BERT demonstrates consistent state-of-the-art performance across both data settings. In the full-data scenario, it is the top-performing model in 23 out of 24 metrics. This superiority is maintained in the data-scarce setting, where it still achieves the best result in 20 out of 24 metrics, underscoring its general effectiveness.
(2) Vola-BERT's primary advantage is its data efficiency. This becomes clear when comparing performance degradation between the two settings. While recurrent models like LSTM and GRU degrade sharply, Vola-BERT's error remains stable. For instance, its rMAE advantage over GRU for GBP/USD forecasting widens significantly in the data-scarce setting, with reductions of 14.1 percent, 21.2 percent, and 22.0 percent across the three forecast horizons.



(a) Market Session: London Open

(b) Market Session: London Close & NY Open

Figure 2: Analysis of Vola-BERT's learned understanding of intraday market sessions. Each plot shows the historical volatility (black), the standard forecast (blue), and the counterfactual forecast conditioned on a semantic token (red). Conditioning on the `[LONDON_OPEN]` token (a) consistently increases forecasted volatility, while conditioning on the `[LONDON_CLOSE]` token (b) tends to suppress it, demonstrating that the model learns an interpretable representation of cyclical market behavior.



(a) Bank of England (BoE) Interest Rate Decision

(b) FOMC Press Conference

Figure 3: Quantifying the impact of high-impact news events. For the BoE decision (a) and FOMC press conference (b), we plot the actual volatility (black) against Vola-BERT's forecast conditioned on no event (red) and on a high-impact event (blue). The model learns to anticipate a significant volatility spike when conditioned on the event token, demonstrating its ability to incorporate semantic information about scheduled news.

(3) Vola-BERT's architectural choice of a bidirectional backbone is validated by the results. Vola-BERT consistently outperforms other LLM-based models that use causal architectures, such as GPT4TS and AutoTimes. This suggests that a holistic, non-causal analysis of the entire lookback window is fundamentally better suited for capturing the complex, interacting signals present in financial time series.
(4) Vola-BERT shows a particular robustness against large forecast errors. Its superior performance is not limited to the rMAE metric; it also dominates on rMSE, which penalizes significant miscalculations more severely. This indicates a higher degree of reliability for practical applications like risk management, where avoiding large, unexpected errors is critical.

We hypothesize that Vola-BERT's strong performance in data-scarce settings is due to transfer learning. The LLM's pre-training on a vast text corpus teaches its self-attention layers to become general-purpose sequence processors. Unlike models trained from scratch that must learn these patterns from the limited financial data, Vola-BERT only needs to adapt its existing capabilities, making it far more robust when data is scarce.

4.3 Market Regime Analysis with Semantic Conditioning

We used the semantic conditioning version of Vola-BERT to analyze its learned understanding of market drivers. This transforms the model from a simple predictor into an analytical tool that can quantify the impact of specific, real-world events on volatility.

First, we analyze the model's understanding of intraday market sessions. Figure 2 shows how Vola-BERT's forecasts change when conditioned on tokens representing the London Open and London Close. The model successfully captures the well-known intraday

volatility pattern: it learns that the London open is associated with a spike in volatility, while volatility tends to decline after the London session closes during the New York market hours.

Next, we demonstrate the model's ability to quantify the impact of high-impact news events. Figure 3 shows the forecasted volatility conditioned on the `[HIGH_IMPACT_NEWS]` token versus the `[NO_EVENT]` token around two major events: the Bank of England interest rate decision and the FOMC press conference. While the model cannot predict an event's content, it correctly learns that such events are associated with a significant, sharp increase in volatility followed by a decline.

The analytical capabilities demonstrated here have significant practical implications. The semantic conditioning module moves LLMs from being black-box predictors to analytical tools. A portfolio manager could use this model to ask counterfactual questions, such as "How much will my risk increase if we enter the London session during an FOMC announcement?" In response, the model can provide an expected risk estimate for the given signal, for instance by forecasting that the volatility may increase by a specific percentage. Such information enables market participants to make more informed decisions, including portfolio re-balancing, adjusting hedging strategies, and timing market entry and exit. This bridges the gap between purely quantitative models and qualitative market understanding.

5 Conclusion and Future Work

In this paper, we introduced Vola-BERT, a model that repurposes a pre-trained bidirectional language model for the task of FX volatility forecasting. Our empirical evaluation shows that this approach achieves state-of-the-art accuracy, with a pronounced advantage in data-scarce settings. This result validates our core hypothesis that the general capacity for modeling complex sequences, learned from text, can be effectively transferred to the numerical domain of financial time series. We also proposed an innovative semantic conditioning framework. This method transforms Vola-BERT from a simple forecaster into an interpretable analytical tool by enabling it to learn and quantify the impact of real-world market drivers, such as scheduled economic events.

The proposed framework is ready for practical use, as it can be directly applied by providing signals that drive FX volatility, such as economic events and analyst sentiment. Once trained, our model can effectively assess and quantify the impact of these features, enabling flexibility for various forecasting scenarios and signal types. A promising direction for future work is a multi-modal architecture that expands our semantic conditioning framework. Integrating real-time news text would allow the model to move beyond simply recognizing scheduled events and learn to distinguish between expected outcomes and market-moving surprises, further refining its forecasts and explainability. We consider this as a crucial improvement, as news sources arguably offer the most direct reflection of the current market sentiment, but are also among the hardest data types to utilize as model inputs. Lastly, exploring alternative bidirectional LLMs with improved architectures or training procedures, such as RoBERTa [25] or DeBERTa [18], could enhance our proposed model's forecast accuracy and is a valuable direction for future research.

Acknowledgments

We thank the anonymous reviewers for their thorough evaluation and valuable suggestions. We are also grateful to Dr Livia Paranhos (Bank of England) and Dr Marios Kogias (Imperial College London) for feedback following the submission. This work is supported by the United Kingdom EPSRC (grant numbers UKRI256, EP/V028251/1, EP/N031768/1, EP/S030069/1, and EP/X036006/1), Altera, Intel, AMD, SiliconFlow, and Google Cloud.

References

[1] Alia Afzal and Philipp Sibbertsen. 2023. Long memory, spurious memory: Persistence in range-based volatility of exchange rates. *Open Economies Review* 34, 4 (2023), 789–811.

[2] Torben G Andersen and Tim Bollerslev. 1998. Deutsche Mark-Dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *The Journal of Finance* 53, 1 (1998), 219–265.

[3] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Clara Vega. 2003. Micro effects of macro announcements: Real-time price discovery in foreign exchange. *American economic review* 93, 1 (2003), 38–62.

[4] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the Language of Time Series. arXiv:2403.07815 [cs.LG] https://arxiv.org/abs/2403.07815

[5] Farhad Bahramy and Sven F. Crone. 2013. Forecasting foreign exchange rates using Support Vector Regression. In *2013 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. 34–41. doi:10.1109/CIFEr.2013.6611694

[6] Bank for International Settlements. 2022. Triennial Central Bank Survey of Foreign Exchange and Over-the-counter (OTC) Derivatives Markets in 2022. https://www.bis.org/statistics/rpfx22.htm.

[7] Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 3 (1986), 307–327. doi:10.1016/0304-4076(86)90063-1

[8] Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. 2024. LLM4TS: Aligning Pre-Trained LLMs as Data-Efficient Time-Series Forecasters. arXiv:2308.08469 [cs.LG] https://arxiv.org/abs/2308.08469

[9] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[10] Ray Yeutien Chou and Nathan Liu. 2010. The economic value of volatility timing using a range-based volatility model. *Journal of Economic Dynamics and Control* 34, 11 (2010), 2288–2301.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).

[13] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[14] FundYourFX. 2025. The Importance of Continuous Learning in Forex Trading. https://fundyourfx.com/importance-of-continuous-learning-in-forex-trading/

[15] Paul Geertsema and Helen Lu. 2022. Measuring information decay in financial markets. *Available at SSRN 3965171* (2022).

[16] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2024. Large Language Models Are Zero-Shot Time Series Forecasters. arXiv:2310.07820 [cs.LG] https://arxiv.org/abs/2310.07820

[17] Himanshu Gupta and Aditya Jaiswal. 2024. A Study on Stock Forecasting Using Deep Learning and Statistical Models. arXiv:2402.06689 [q-fin.ST] https://arxiv.org/abs/2402.06689

[18] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654 [cs.CL] https://arxiv.org/abs/2006.03654

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. arXiv:2310.01728 [cs.LG] https://arxiv.org/abs/2310.01728

[21] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] https://arxiv.org/abs/1412.6980

[22] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. arXiv:2001.04451 [cs.LG] https://arxiv.org/abs/2001.04451

[23] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2024. Large Language Models in Finance: A Survey. arXiv:2311.10723 [q-fin.GN] https://arxiv.org/abs/2311.10723

[24] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. arXiv:2310.06625 [cs.LG] https://arxiv.org/abs/2310.06625

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[26] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024. AutoTimes: Autoregressive Time Series Forecasters via Large Language Models. arXiv:2402.02370 [cs.LG] https://arxiv.org/abs/2402.02370

[27] Michael C. Münnix, Takashi Shimada, Rudi Schäfer, Francois Leyvraz, Thomas H. Seligman, Thomas Guhr, and H. E. Stanley. 2012. Identifying States of a Financial Market. arXiv:1202.1623 [q-fin.ST] https://arxiv.org/abs/1202.1623

[28] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. arXiv:2211.14730 [cs.LG] https://arxiv.org/abs/2211.14730

[29] O. F. Olaiyapo. 2024. Applying News and Media Sentiment Analysis for Generating Forex Trading Signals. *Review of Business and Economics Studies* 11, 4 (Feb. 2024), 84–94. doi:10.26794/2308-944x-2023-11-4-84-94

[30] Michael Parkinson. 1980. The extreme value method for estimating the variance of the rate of return. *Journal of business* (1980), 61–65.

[31] Sahabeh Saadati and Mohammad Manthouri. 2024. Forecasting Foreign Exchange Market Prices Using Technical Indicators with Deep Learning and Attention Mechanism. arXiv:2411.19763 [cs.LG] https://arxiv.org/abs/2411.19763

[32] Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. 2024. TEST: Text Prototype Aligned Embedding to Activate LLM's Ability for Time Series. arXiv:2308.08241 [cs.CL] https://arxiv.org/abs/2308.08241

[33] Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.

[34] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for Large Language Models: A Survey. arXiv:2309.01029 [cs.CL] https://arxiv.org/abs/2309.01029

[35] Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All:Power General Time Series Analysis by Pretrained LM. arXiv:2302.11939 [cs.LG] https://arxiv.org/abs/2302.11939