**Working Paper** (WP/25-10)

# Forecasting Federal Fund Rates with AI: LSTM, GRU, and Large Language Model Approaches

Jorge A. Chan-Lau, Toan Long Quach, and
Aruhan Rui Shi

September 2025

[This page is intentionally left blank]

# Forecasting Federal Funds Rates with AI: LSTM, GRU, and Large Language Model Approaches

Prepared by Jorge A. Chan-Lau, Toan Long Quach, and Aruhan Rui Shi [*]

Approved by Abdurohman (Deputy Director)

September 2025

**Abstract**

Forecasting the U.S. federal funds rate is important since it impacts global financial conditions, and in turn the real economy. This study examines whether AI models could forecast policy rates accurately by inferring the implicit decision rules the U.S. Federal Open Markets Committee (FOMC) follows when setting the policy rate. We examine two recurrent neural networks (RNNs), LSTM and GRU, and a large zero-shot language model (LLM) capable of combining numerical economic data and interpreting qualitative information. The forecasting performance of the three models is good, especially during the zero lower bound and early COVID 19 periods. The policy rules implied from the RNN models suggest that the FOMC might have been overly accommodative during the COVID-19 pandemic. Since 2020, the models suggest a deviation from prior monetary policy patterns, with the FOMC adjusting rates less frequently than predicted by the model-implied policy rules. These findings suggest that incorporating AI insights could enhance our understanding of and ability to predict future Fed rate decisions.

| | |
| --- | --- |
| JEL classification: | C45, C54, E47, E52 |
| Keywords: | Federal funds rate, forecasting, deep learning, generative AI, large language model, long-short term memory model, gated recurrent unit |

---

[*] We benefitted from comments and suggestions by Yoki Ogawa, Andrew Tsang, and internal seminar participants at AMRO. We remain responsible for any errors or omissions. Please address correspondence to all authors. Email addresses: Chan-Lau: jchanlau@gmail.com (permanent) and jorge.chan-lau@amro-asia.org; Quach: quachtoanlong@gmail.com (permanent) and quach.toanlong@amro-asia.org; Shi: aruhanruishi@outlook.com (permanent) and aruhan.rui.shi@amro-asia.org

## Contents



## List of Figures



## List of Tables

## 1 Introduction

Given the central role of the U.S. dollar in international finance and global trade, understanding and predicting changes in the Fed's policy rate is crucial as its impact on global financial conditions affect exchange rates, investment decisions, and borrowing costs in other countries—particularly in emerging markets—through interest rate differentials, risk sentiment, and portfolio rebalancing channels.

The forecasting task is inherently difficult since it involves inferring the implicit decision rules the U.S. Federal Open Markets Committee (FOMC) follows when setting the policy rate. The rules are not necessarily fixed and could change over time due to changes in the Committee's composition, evolving quantitative and qualitative policy frameworks, and the overall discretionary nature of monetary policy. In addition, rate decisions are influenced by a wide range of factors, including real-time economic data, financial market conditions, geopolitical developments, and communication strategies such as forward guidance, all of which add complexity and uncertainty to the forecasting task.

This paper argues that artificial intelligence (AI) models may better capture the implicit rate decision rules the FOMC follows. Specifically, it examines the forecasting performance of three AI-based models. The first two models are modern recurrent neural networks (RNNs) developed for modeling long sequences of observations. The first RNN is a long short-term memory (LSTM) model—a type of RNN with a gating mechanism that allows information to be filtered and passed from one time step to the next. This structure makes LSTMs well-suited for time series forecasting when the model inputs are numerical data. The second RNN is a gated recurrent unit (GRU) network, another RNN which was developed to reduce the computational time required to estimate the network parameters. It shared the same principles as the LSTM, but it has a simpler gate structure, which enables faster training (estimation) times.

The LSTM and GRU networks only process numerical inputs and cannot incorporate qualitative factors that may also influence FOMC decisions. To address this, we also explore the performance of a third model, a zero-shot application of a state-of-the-art large language model (LLM), Anthropic's Claude 3.5 Sonnet.[1] The LLM is capable of processing and interpreting contextual and qualitative information—such as economic narratives, forward guidance, and geopolitical developments. We input into the LLM the same quantitative information input into the RNNs together with a textual prompt asking the model to review the data in its capacity as a central bank advisor.

The results show that the LSTM, GRU, and zero-shot LLM models perform well in forecasting the federal funds rate, with the LLM outperforming the LSTM and GRU by a small margin. Notably, the RNN models underperform the LLM during the zero-lower bound period and the early phase of the COVID-19 pandemic as they systematically predicted rates higher than those observed. This

1. The calculations reported here are based on the Claude 3.5 Sonnet model Anthropic released in June 20 2024 and accessed in April 2025.

underscores the added value of incorporating qualitative information alongside quantitative data. Furthermore, we find that, relative to the policy rate adjustments suggested by the models, the actual policy rate path has been more persistent, with fewer rate changes over time, regardless of whether the economy was in a low-rate environment, such as during the pandemic years, or a high-rate environment, as during the post-pandemic high inflation period. Arguably, this finding might suggest a recent policy change the models have not been able to capture fully. Finally, we identify a potential structural shift in monetary policy conduct after 2020: had the FOMC followed the implicit rules derived from our RNN models, it likely would have implemented smaller rate cuts during the pandemic and consequently required less aggressive tightening when inflationary pressures reemerged.

In the remainder of this document Section 2 reviews some selected applications of LSTM and GRU networks, and LLMs in economics and finance. Section 3 describes the data and methods used to implement the models, and Section 4 presents the results. Finally, Section 5 concludes by discussing the results and their implications.

## 2 LSTM, GRU, and LLM applications in economics and finance

Modern recurrent neural networks, such as LSTMs (Hochreiter and Schmidhuber 1997) and GRUs (Cho et al. 2014), were developed to address the main weakness of standard recurrent neural networks (RNNs), namely their difficulty for capturing long-term dependencies as exploding and vanishing gradients when using back-propagation during the training phase hampered their estimation (Goodfellow, Bengio, and Courville 2016). Ease of calculation have partly contributed to the increased prominence of LSTM and GRU networks in time series forecasting, as they can model nonlinear dynamics and capture long-range dependencies. The LSTM and GRU-based forecasting models rely on numerical data inputs, such as historical time series values, to identify temporal patterns and make predictions.

An alternative approach to the use of modern RNNs involves leveraging the multimodality of LLMs, which have proved very effective to understand and process various data types, including numerical and textual data. When fed numerical data, LLMs can exploit their extensive training corpora, which include qualitative information like textual descriptions, economic trends, or domain-specific knowledge, to contextualize and enhance the forecasting task. LLM's multimodal capability allows them to potentially capture nuanced relationships that purely numerical models like LSTMs and GRUs might miss, offering a complementary method for time series prediction.

### 2.1 LSTM applications

Initially applied to univariate time series, LSTMs have demonstrated consistent predictive gains over classical linear models such as ARIMA and exponential smoothing, particularly when applied to macroeconomic indicators or financial series with nonlinear or regime-dependent behavior. In the macroeconomic domain, LSTM models have been employed effectively in forecasting aggregate output (Hamiane et al. 2023; Zhang, Wen, and Yang 2022; Hamiane et al. 2024; Xie et al. 2024; Zhao 2024), anticipating turning points associated with economic and financial crises (Park and Yang 2022), and producing inflation forecasts across medium- and long-term horizons (Lakshmi Narayanaa et al. 2023; Zhao 2024; Liu and Lan 2025; Paranhos 2025). These models' capacity to internalize complex functional relationships without requiring explicit structural specification has proven particularly advantageous in high-dimensional or noisy environments.

Applications in financial economics have similarly confirmed LSTMs usefulness. Their ability to replicate intricate market dynamics has led to improved predictive performance across a range of applications, including stock return forecasting (Serin and Kemalbay 2024; Furizal et al. 2024; Pilla and Mekonen 2025), particularly at short horizons (Kobiela et al. 2022). LSTMs have also outperformed traditional univariate models in interest rate forecasting (Salem, Jummah, and Albourawi 2024) and financial risk modeling (Xu et al. 2024). Notably, when compared against the benchmark Generalized Autoregressive Conditional Heteroskedasticity (GARCH) framework of Bollerslev (1986), LSTMs have yielded more accurate estimates of Value-at-Risk (VaR) in several empirical settings (Ormaniec et al. 2022).[2]

LSTM have also been applied to analyze multivariate models. Cao, Li, and Li (2019) find that LSTMs outperform vector autoregressions (VARs), particularly during high-volatility periods and Hopp (2022) shows LSTMs are more accurate than dynamic factor models for nowcasting trade aggregates. Using U.S. data, Chan-Lau and Quach (2025) report that quarterly LSTM models clearly outperform VARs in one-step-ahead forecasts, with path dependence affecting the shock propagation dynamics. Recent developments have introduced hybrid LSTM architectures to improve forecasting accuracy. They incorporate attention mechanisms (Hollis, Viscardi, and Yi 2018; X. Zhang et al. 2019; Ju and Liu 2021), combine them with other machine learning methods such as SVR, CatBoost, or DFMs (Lashina and Grishunin 2023), hidden markov models (Sivakumar 2025), and GRUs (Lawi, Mesra, and Amir 2021).

### 2.2 GRU applications

GRUs, with only two gates (functions) to process past and current information, are faster to train than LSTMs, which have three gates, and tend to deliver better results in language modeling (Jozefowicz, Zaremba, and Sutskever 2015) and show a comparable performance in financial time series forecasting (Shiri et al. 2024). In particular, their faster training speed made GRUs especially suitable for high frequency forecasting in finance. Examples include Dai (2025), who recommends gold investment strategies partly based on a GRU's long-term predictions; and Umezuruike et al. (2024),

2. However, in high-frequency contexts characterized by volatility clustering and microstructure noise, GARCH-type models may continue to offer comparative advantages (Sezer, Gudelek, and Ozbayoglu 2020).

which show that a GRU predicts stock prices better than LSTM and transformer models. Pirani et al. (2022) find that a GRU model performs better than ARIMA and multiple LSTM variants when forecasting financial time series when data was processed in reverse order, that is, when the time steps were traversed backwards.

As for economic applications, the results in Patel, Sanghavi, and Singh (2023) demonstrate that the GRU and LSTM models perform equally well in forecasting U.S. GDP, outperforming traditional univariate forecasting models such as ARIMA, exponential smoothing models, and their many variants. Guo and Wang (2024) show that a GRU model, enhanced with feature engineering and exogenous variables, significantly improves economic forecasting accuracy for ice and snow tourism. Zhu, Zhang, and Tan (2025) obtain reasonable predictions GDP predictions using a GRU model that include among its explanatory variables power factors such as electricity consumption. However, it is not necessarily the case that the most sophisticated models perform better. Naas and Zouaoui (2024) identify linear regression as the top-performing model for forecasting exchange rate volatility, with the GRU model ranking a close second.

As in the case of LSTMs, hybrid models are being developed rapidly with many of them including both LSTM and GRU layers. Zhu et al. (2025), by replacing one of the gates in an otherwise standard GRU model with an attention mechanism (Vaswani et al. 2017), obtain more accurate forecasts of stock market movements in China and the U.S. Liu and Lai (2025) using a hybrid PCA-GRU-LSTM, demonstrate that environmental factors could improve stock market forecasts. Gu et al. (2024) build an ensemble exchange rate forecasting model, where the weak learners combines LSTM and GRU layers, and AdaBoost serves to construct the final strong learner. They apply the model to different currency pairs, including the U.S. dollar, the Chinese yuan, and the British pound.

### 2.3 LLM applications

The emergence of large language models, such as OpenAI's GPT, Google's PaLM, and Anthropic's Claude, has significantly advanced the integration of artificial intelligence into economic and financial forecasting. Initially developed for understanding natural language, LLMs have shown great potential in reasoning, simulation, and prediction tasks, suggesting promising and novel ways to conduct macroeconomic modeling, market forecasting, and decision analysis.

Recent studies have evaluated LLMs' capabilities in economic prediction tasks. Faria-e-Castro and Leibovici (2024) examine the performance of a foundation model (Google's PaLM) in generating conditional forecasts of U.S. inflation. They find that the model's forecasts are competitive with, and often outperform, those from professional forecasters, especially over longer horizons. Similarly, Carriero, Pettenuzzo, and Shekhar (2025) explore LLM-based forecasting for macroeconomic time series. Their study shows that foundation models can match or exceed traditional models in out-of-sample forecasting, suggesting potential for reducing model specification risk in economics.

In finance, LLMs are being applied for sentiment analysis, risk modeling, and earnings prediction. Wu et al. (2023) introduce BloombergGPT, a domain-specific LLM trained on a massive financial text corpus. It outperforms general-purpose LLMs on various financial NLP tasks, such as question answering and document classification. Kim, Muhn, and Nikolaev (2025) analyze the effectiveness of prompting LLMs like GPT to perform financial statement analysis. They find that the LLM's assessments of earnings quality correlate strongly with future returns and outstrip the performance of human analysts and traditional text-based models.

Another frontier application involves using LLMs to simulate economic decision-making. Horton (2023) demonstrates that LLMs can replicate behavioral patterns uncovered by behavioral economic experiments, including framing effects and loss aversion. Hao and Xie (2025) propose a multi-agent framework using several LLMs as artificial economic agents for policy simulations. These agents display heterogeneity and responsiveness to policy scenarios, illustrating how LLMs can be used for dynamic agent-based modeling in economics.

While promising, the use of modern RNNs and LLMs raises critical challenges common to all AI models. Model interpretability and reliability remain central concerns in economic applications. Molnar (2022) emphasizes the importance of explainability in machine learning, highlighting techniques to open the black box of LLM decision processes. Additionally, Fuster et al. (2022) show that machine learning models, if improperly deployed, may exacerbate existing inequalities, particularly in credit and lending markets. This underscores the need for ethical and transparent model deployment in finance.

## 3 Empirical implementation

### 3.1 Data

Analysts often rely on a core set of macroeconomic and financial indicators to anticipate the Federal Open Market Committee (FOMC) policy decisions. These typically include inflation rates, unemployment figures, GDP growth, labor market conditions, and asset price movements. In line with this practice, our analysis incorporates a similar set of variables to inform both the recurrent neural network RNN models, namely LSTM and GRU, and the LLM. While acknowledging that additional explanatory variables could enhance the analysis, we deliberately focus on this core set to examine whether the fundamental information they contain is sufficient for accurate Federal Funds rate prediction. The explanatory variables, sourced from Haver Analytics and the Federal Reserve Economic Data (FRED), are listed in Table 1. The data sample period is January 1998 to April 2025, and the observations reflect revised figures available as of August 2024 rather than real-time data. The models' results, hence, reflect the underlying structure of the economy rather than being nowcasts of the Federal Funds rate.

In the RNN-based models, some variables are transformed from raw data into lags, growth rates or changes to improve stationarity and enhance model learning. In contrast, the LLM was prompted directly using the level and growth information in tabular format, preserving the semantic of the raw indicators. The input set for the LLM included monetary aggregates (M1 and M2), labor market indicators (e.g., U-3 and U-6 unemployment rates, changes in nonfarm payrolls, average wages, labor force participation), inflation measures (CPI, PPI, and PCE—both headline and core), inflation expectations (1-year and 5-year), real GDP, and asset prices (housing and commercial real estate), alongside their respective growth rates.

Table 1: RNN-based models: variables, description and data transformation

| Haver code | Description | Data transformation |
|---|---|---|
| FF | Federal Funds rate | 1 month lag |
| FM2 | Money stock (M2) | Monthly change, YoY % change |
| GSACPPIC | Green Street Advisors commercial property price index | YoY % change |
| JCBM | Personal consumption expenditure (PCE), chain price index | YoY % change |
| JCXFEBM | PCE less food & energy (core PCE), chain price index | YoY % change |
| LANAGRD | Change in total nonfarm employment | Level (no transformation) |
| LEPRIVA | Avg hourly earnings, total private industries | Dropped from analysis |
| LKPRIVA | Avg weekly earnings, production and nonsupervisory | YoY % change |
| MGDPN | Nominal GDP | YoY % change |
| MLU6 | U-6 unemployment rate | MoM and YoY difference |
| PZRAW | CRB spot commodity price index | YoY % change |
| PCU | CPI-U (all urban consumers) | YoY % change |
| PCUSLFE | CPI-U less food and energy | YoY % change |
| USPHPIM | FHFA house price index (purchase-only) | YoY % change |
| YC_2y_10y | Yield curve (10Y minus 2Y) | Spread |
| YC_6m_1y | Yield curve (1Y minus 6m) | Spread |

Sources: Haver Analytics and Federal Reserve Economic Data (FRED).

### 3.2 LSTM and GRU networks

The Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) and Gated Recurrent Unit (GRU) (Cho et al. 2014) networks are specialised types of recurrent neural network (RNN) designed to effectively learn and remember patterns in sequential data over long periods.[3] LSTMs and GRUs address the difficulties traditional RNNs encounter due to the potential presence of vanishing and exploding gradients, which could cause relevant past information to be lost during the training process. These modern RNNs avoid problems associated with badly-behaved gradients using a memory and gating mechanism. Processing each input sequence element at a time, the networks update their internal state with each new input, storing the information in a memory cell. The gates control what information is added, retained, or discarded at each step, enabling learning new data patterns while preserving valuable past knowledge.

3. Please see the appendix for a formal description of the networks. For more details, see the original references, and the textbook treatments in Goodfellow, Bengio, and Courville (2016) and A. Zhang et al. (2023).

The basic LSTM block include three gates—input, forget, and output— that regulate the flow of information, allowing the network to selectively remember or forget data as needed. This structure enables LSTMs to capture long-range dependencies and context, making them particularly well-suited for tasks involving time series forecasting. GRUs were introduced as a simpler alternative to LSTM networks, aiming to address the same challenge of capturing long-term dependencies in sequences while reducing computational complexity. The GRU achieves this by combining the memory cell and gating mechanisms into a more streamlined structure, using only two gates: the reset gate and the update gate. These gates control how much of the previous information is retained or updated with new input, allowing the network to adaptively remember or forget information as needed.

Unlike LSTMs, which use separate input, output, and forget gates, a GRU block merges some of these functionalities, resulting in fewer parameters and faster training times. The update gate determines the extent to which the unit updates its activation or content, while the reset gate controls how much of the past information to forget. This design enables GRUs to capture complex temporal patterns and dependencies in data, making them well-suited for tasks like time series forecasting, speech recognition, and natural language processing. Their simplicity and efficiency often allow GRUs to perform comparably to LSTMs, making them preferable when computational resources or training time are limited. Since our forecasting project is relatively small and resource constraints are minimal, we use both models to compare their performance.

### 3.2.1 LSTM and GRU architecture

The LSTM and GRU architecture used is similar to the multi-step architecture in Chan-Lau and Quach (2025) but specialized for forecasting a single variable. Let $\boldsymbol{X}_t$ be the vector collecting the observations of the explanatory variables at time $t$, and $n$ the number of lagged observations used to predict the federal funds rate, $y_t$. We set up the architecture by chaining $n$ LSTM (or GRU) blocks (or units) sequentially. In this setup, unit $-i$ processes the information corresponding to the lagged observation vector $\boldsymbol{X}_{t-i}$, and transmits its output, summarized by the updated memory cell and hidden state, to the next block, unit $-i+1$. At the end of the chain, the hidden state of the last block, unit $-1$, is input into a fully connected output projection head, consisting of a dropout layer followed by two linear transformations with a rectified linear unit (ReLU) activation in between, to predict $y_t$. This setup allows the LSTM and GRU to learn a mapping function $F_{\boldsymbol{\Theta}}\left(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots, \boldsymbol{X}_{t-n}\right) = y_t$, where $\boldsymbol{\Theta}$ represents the model parameters, that is, the weights and biases of the different components of the neural network blocks.[4]

4. See Chan-Lau and Quach (2025) for details on how LSTM blocks are chained together. GRU blocks are similarly chained together.

### 3.2.2 Hyperparameter tuning, model training, and out-of-sample backtests

Hyperparameters are the settings or configuration choices that define the structure and learning process of the RNN models before any training begins. Hyperparameter tuning is the process of finding the best hyperparameter settings for the LSTM and GRU models. The hyperparameters of interest are the number of hidden units, the learning rate, the dropout rate, the batch size, the number of time steps (number of lagged observations and units), and the number of training epochs (or number of complete passes through the entire training set).[5] Once the hyperparameters' values are set, the model training process begins, during which the model parameters—such as weights and biases—are estimated from the data. These model parameters are adjusted iteratively by the learning algorithm to minimize the loss function and improve predictive accuracy.

To select the hyperparameters and estimate the model parameters, the dataset was divided into a training set, covering all observations up to December 2015, and a test set, spanning January 2016 to April 2025. Within the training set, a five-fold time series cross-validation was implemented. In this approach, each fold used a progressively larger subset of the training data for training, while the subsequent time window served as the validation set. This setup ensured that the temporal structure of the data was preserved, and that future data was never used to predict the past.

Hyperparameter tuning was performed using Bayesian optimization, which systematically searches over a predefined range of hyperparameter values to identify the combination that minimizes the loss function on the validation folds (Garnett 2023). Once the optimal hyperparameters were selected, the model was retrained on the full training set before conducting a final evaluation on the test set. For the LSTM, the selected hyperparameter values were 12 time steps, 109 hidden units, two LSTM layers with no dropout, a learning rate of 0.0001, and 644 training epochs. For the GRU, the hyperparameter values were 12 time steps, 188 hidden units, a single LSTM layer with no dropout, a learning rate of 0.0002, and 449 training epochs.

In addition, recursive out-of-sample backtests were conducted to assess the predictive power of the model in a real-time, nowcasting setting. This involved using an expanding window and re-estimating the model at each point in time. Starting in January 2016, at any given period $t$, the model was trained using the data sample ending in period $t-2$, and afterwards, the observations $\boldsymbol{X}_{t-1}$ were used to predict the Federal Funds rate $y_t$.

## 3.3 LLM model

The LLM employed in this analysis is a zero-shot implementation of Claude 3.5 Sonnet (developed by Anthropic, released on June 20, 2024, and accessed in April 2025). This state-of-the-art model is capable of processing not only structured economic data but also unstructured, qualitative in-

5. See Goodfellow, Bengio, and Courville (2016) and A. Zhang et al. (2023) for a comprehensive discussion of these hyperparameters.

formation—such as economic narratives, forward guidance, and geopolitical developments. It can synthesize these diverse inputs to generate coherent, context-aware predictions and explanations.

A key advantage of using Claude 3.5 in this context lies in its ability to incorporate real-world context into economic forecasting. This includes the model's capacity to reason over policy signals, macroeconomic trends, and textual cues that may not be easily encoded in traditional statistical models. However, like many LLMs, it is susceptible to certain limitations. One such issue is "knowledge leakage," whereby the model may produce responses that suggest undue certainty about past events, despite having no direct access to historical data. While this phenomenon complicates the retrospective validation of its predictions, its impact may be less pronounced when the model is used for forward-looking analysis, such as forecasting future decisions by the Federal Reserve.

Our methodological approach consists of two key components. First, we explore the model's ability to generate real-time forecasts by prompting it to predict the Federal Reserve's most recent policy decision. In this step, the prompt was modified to explicitly request the model's view on the likely direction of policy—whether the Fed would raise, lower, or maintain the target rate—based on the most current economic data and contextual indicators.

Second, we assess the model's predictive accuracy by simulating historical forecasts of the Federal Funds Target Rate and comparing them with actual outcomes. In this validation phase, the model was provided with economic indicators and relevant contextual information available at the time of each decision. It was then prompted to generate a forecast of the target rate, along with a confidence estimate and a concise explanation of its reasoning. Together, these exercises allow us to evaluate not only the model's point forecasts, but also its interpretive and explanatory capabilities when reasoning over contemporary economic scenarios.

When conducting retrospective simulations to assess the LLM's underlying knowledge and predictive capabilities, the prompt presented to the model was crafted to emulate a realistic forecasting scenario and was structured as follows:

> *As an AI central bank advisor, review the past n months of U.S. economic data and predict the Fed Funds Target Rate the FOMC is likely to set on target date. In addition, please only use your knowledge up to the date of prediction.*

The model was instructed to return its forecast in a consistent, structured format:

> *Predicted rate; confidence level; and explanation limited to 100 words.*

This standardized prompt–response structure was implemented to ensure consistency across multiple simulations, enabling systematic comparison and reproducibility of the results. We also constrained its access to information that would have been available only up to the date of each forecast—effectively emulating a real-time decision-making environment. We conducted a simulation comprising 100 historical replications in which the model generated forecasts for previous FOMC meetings using only contemporaneous data. The average value of the replications was then compared to the actual policy rate, which enabled us to assess the LLM's aggregate forecasting behavior and its ability to approximate the FOMC's decisions over time.

# 4 Results

## 4.1 LSTM and GRU

Figure 1 displays the recursive out-of-sample predictions of the LSTM and GRU models. The forecasts closely track the FMOC rate decisions during the 2016-2019 rate normalization period and the rapid tightening cycle which started in mid-2022. Table 2 shows several performance metrics, including the root mean squared error (RMSE), the mean absolute error (MAE), and the mean bias error (MBE). Most metrics indicate the models perform very well. For instance, the MAEs are only 0.12 and 0.15 percentage points for the LSTM and the GRU models respectively. The MBEs show the forecasts are mostly unbiased.

Table 2: LSTM and GRU models: forecasting performance

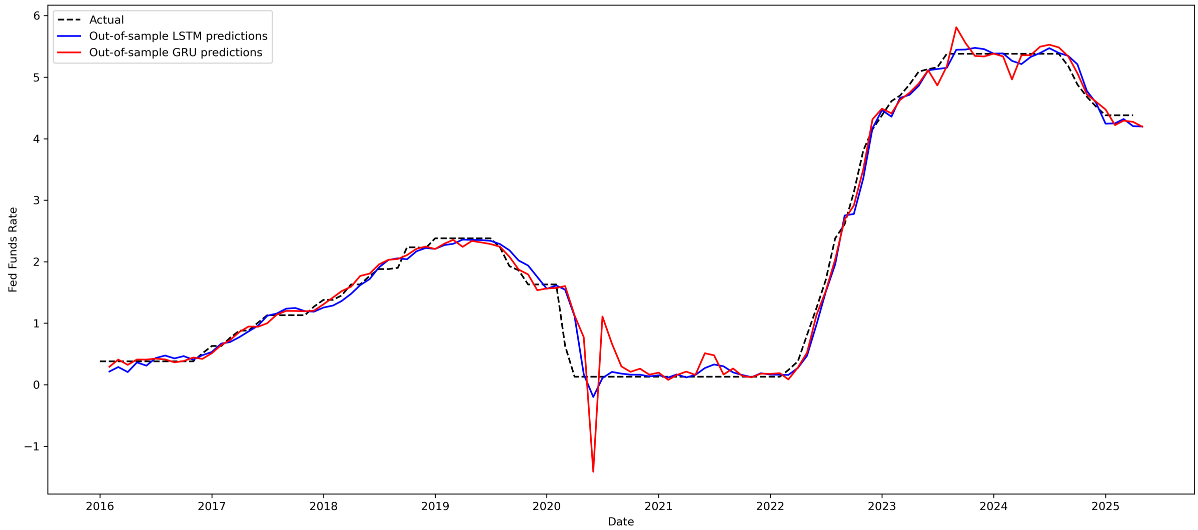
| **Model** | **RMSE** | **MAE** | **MBE** |
|---|---|---|---|
| LSTM | 0.1868 | 0.1164 | -0.0051 |
| GRU | 0.2672 | 0.1476 | 0.0186 |

Note: RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) measure the average size of forecast errors, with RMSE placing greater weight on larger errors. MBE (Mean Bias Error) indicates the average directional bias: positive values correspond to overprediction; negative values correspond to underprediction. All errors are in percentage points unless otherwise indicated. Lower values mean more accurate estimates.
Sources: Haver Analytics and authors' calculations.

While the forecasting performance is generally good, two periods stand out where the LSTM and GRU models appear to have missed key rate dynamics: first, at the onset of the COVID-19 pandemic, and second, during 2021–22. These divergences between model forecasts and observed federal funds rates can be interpreted if the models are viewed as providing guidance for policy rate decisions. At the onset of the COVID-19 pandemic, both the LSTM-based and GRU-based models missed the emergency rate cuts by the Fed in March 2020. As both models were trained to be consistent with past FOMC behavior, they could not capture the preemptive rate cuts given the favorable economic data recorded in March 2020.[6] Nonetheless, the rapid deterioration of economic conditions in April—driven by severe disruptions to global supply chains and the collapse of the domestic economy due to lockdowns—highlighted the need for aggressive easing of monetary

6. The FOMC statements issued in March 03 and March 15, 2020, show that the FOMC, despite strong economic data, decided to cut rates as it expected the pandemic to drive a rapid deterioration of the economy.

Figure 1: Recursive Out-of-Sample LSTM and GRU predictions vs. actual Fed Funds rate




Sources: Haver Analytics and authors' calculations.

conditions. The model aligns with this view, predicting that the Fed Funds rate should be negative. However, this prediction conflicts with the lower bound policy rate, which prevents the Fed Funds rate from going below zero—a constraint not explicitly incorporated into the models.[7]

Inflation accelerated rapidly in 2021-2022, but the FMOC kept the Fed Funds rate low, considering the inflationary pressures to be transitory. In contrast, the LSTM and GRU models initially recommended starting to raise rates early and to ease the monetary policy stance later on, which brought its one-step rate forecasts closer to the observed policy rate path by mid-2022. Since the model is estimated recursively, we attribute this convergence to the models updating their parameters by learning from the most recent policy actions. From that point onward, both models' predictions mostly follow the FMOC decisions. Nonetheless, both models suggested another potential rate cut in early 2024 and 2025.

From both episodes, during which the predictions diverge from the policy rate decision but only briefly, we infer that the models' parameters are being updated frequently on an online basis as new data become available. In this regard, the model learns and captures in real-time how the FMOC policy decision framework evolves over time, reflecting the rotation of committee members, potential political pressures, and changing views and reassessments of the inflation outlook and its potential drivers. However, it is instructive to consider what the model predictions would have been if the policy framework had remained unchanged - that is, if we assume the decision process was stable over time. This involves training the models on an earlier subsample to fix the model-implied

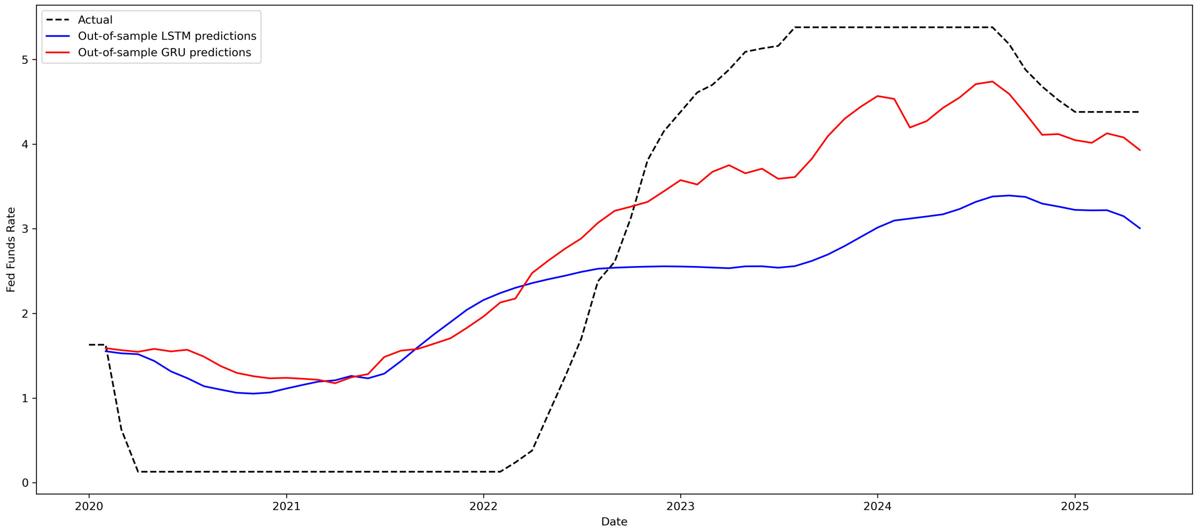
7. The bound could have been accommodated by including the following constraint on the output layer of the LSTM and GRU: $\max(25\text{bps}, \text{LSTM prediction})$.

rate decision rules, and then applying the rules to a more recent subsample without further models' updates. Comparing the models' predictions in this second subsample to the observed Fed Funds rate could provide insights into whether the policy decision framework changed over time.

We assess whether a policy shift occurred from 2020 onward using a simple counterfactual analysis. When trained only on data through December 2019, the models predict that the FOMC would have lowered rates in 2020 to counter the pandemic-induced economic contraction—driven in part by supply-side shocks from lockdown measures—but not by as much as it ultimately did. This smaller-than-observed rate cut can be attributed to the models' reliance solely on hard numerical data. The FOMC, however, had to account for heightened uncertainty surrounding the pandemic's economic impact, prompting a more aggressive easing stance. As the economic shock faded by mid-2021, the models suggested that rates should begin rising. The FOMC delayed the start of the tightening cycle, partly due to its view that inflation was transitory—a belief that likely contributed to the delay. Arguably, policy during this period was also shaped by the adoption of the Flexible Average Inflation Targeting (FAIT) framework in 2020, which allows inflation to run moderately above target following periods of below-target inflation.

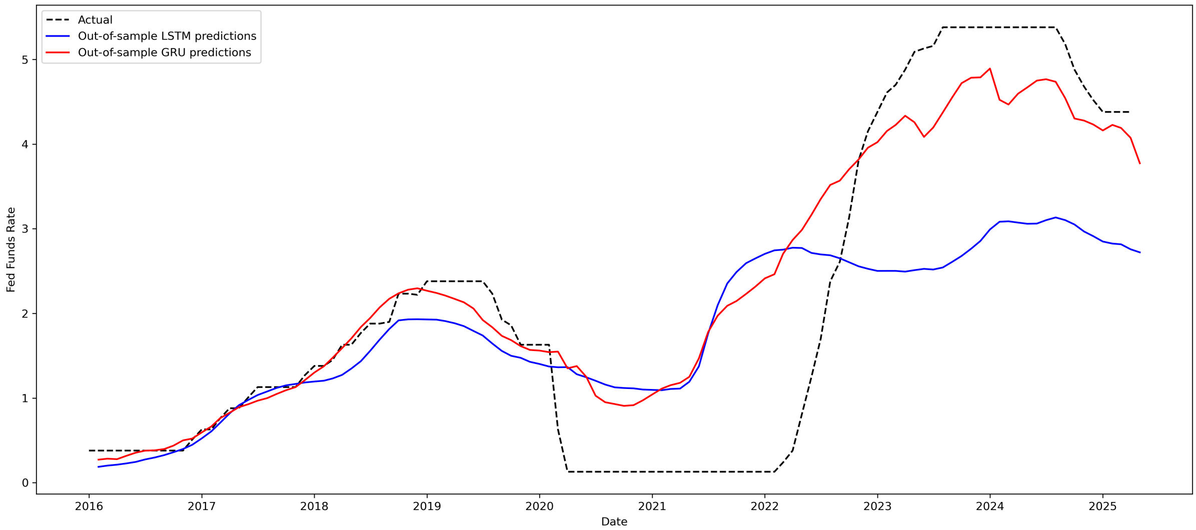
From 2022 onward, the policy rate began converging toward the model-implied rate and has remained elevated since. However, the LSTM model currently suggests that rates should fall to about 3 percent—near the upper end of the 2–3 percent range viewed as neutral by market analysts. The GRU model points to a relatively higher rate of 4 percent, though still below the actual policy rate. Similar results are obtained when the training sample ends in 2015. If evaluated against pre-COVID model-implied policy rules, one could argue that changes in the policy framework led the FOMC to ease too aggressively and tighten too slowly. Equally important, the RNN models' implied decision rules suggest that, despite the models' propensity for more frequent rate adjustments, the FOMC tends to keep rates unchanged for longer periods at both low and high interest rate levels—maintaining either an accommodative or tight policy stance for extended durations relative to pre-pandemic behavior. This points to a possible shift toward a more gradualist approach in monetary policy implementation.

Figure 2: LSTM and GRU: counterfactual Fed Funds rate analysis, training period Jan. 1998 - Dec. 2019




Sources: Haver Analytics and authors' calculations.

Figure 3: LSTM and GRU: counterfactual Fed Funds rate analysis, training period Jan. 1998 - Dec. 2015




Sources: Haver Analytics and authors' calculations.

### 4.2 LLM

#### 4.2.1 Real time prediction

A key strength of LLM in economic forecasting lies in their ability to incorporate and contextualize real-time developments directly within the prompt. A prompt is the textual input provided to a LLM to guide its response. It defines the task; provides context such as background facts, a scenario or economic indicators; and may include input data, instructions, examples, or questions. The prompt serves as the starting point from which the model generates its output, drawing on patterns learned during training. Hence, the LLM is able to interpret not only historical data but also contemporary macroeconomic conditions and policy signals, including qualitative data and unstructured data.

As an example, we prompt the LLM to obtain the real-time forecast of the Federal Reserve's most recent policy decision in March 2025. In this example, the prompt included a reference to ongoing trade policy uncertainty—such as the imposition of U.S. tariffs—to assess how such qualitative context might influence the model's reasoning. When prompted with this additional information, the model returned the following forecast:

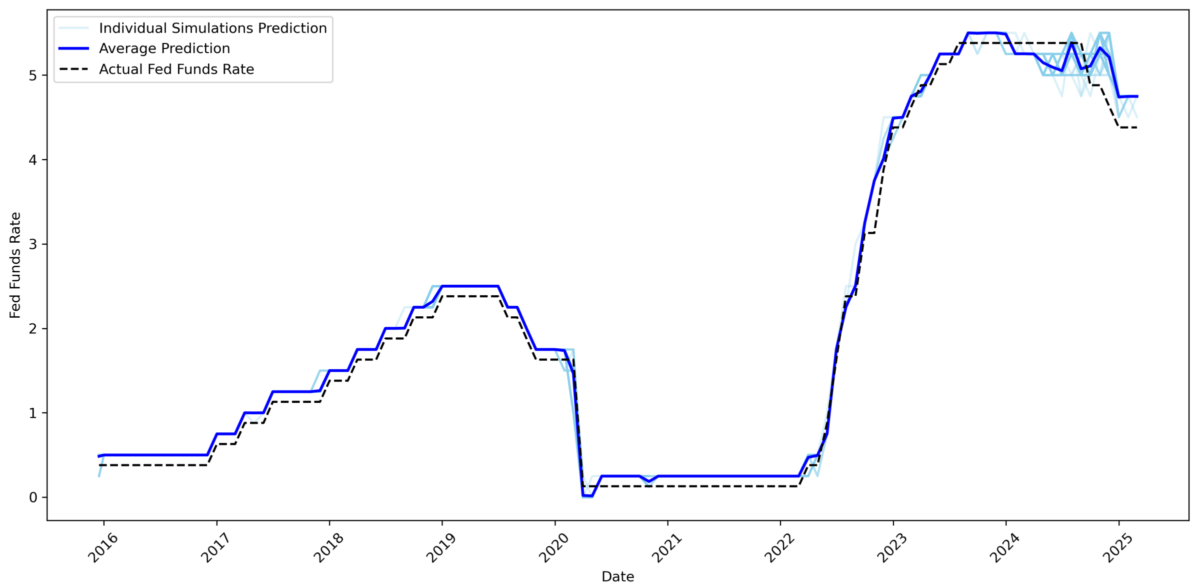
> *Direction: unchanged*
>
> *Explanation: Based on the data, inflation appears to be moderating with CPI and PCE growth rates declining. Unemployment remains low at 4.1 percent. However, GDP growth is slowing and trade tensions create uncertainty. The Fed may hold rates to balance inflation control with economic support, given mixed signals and external risks.*

This simple example demonstrates the model's capacity to synthesize up-to-date economic narratives, including inflation trends, labor market conditions, and geopolitical risks, into its policy assessment. By incorporating structured data alongside recent qualitative developments, the model is able to produce forecasts that are also grounded in a coherent rationale. This interpretive capability makes it a valuable complement to traditional quantitative forecasting methods, such as the RNN-based models described earlier, as these methods may overlook contextual or geopolitical factors. Nonetheless, LLMs remain prone to surface-level reasoning and may generate confident-sounding outputs without a grounded causal framework. As such, their predictions are best used in conjunction with other forecasting approaches to provide a more comprehensive and balanced view.

#### 4.2.2 Prediction validation

Figure 4 displays the individual predictions from each of the 100 LLM's replications in the simulation described in Section 3.3 as well as the average prediction across all replications. Overall, the predicted policy rate generated by the LLM closely tracks the actual federal funds rate and exhibits a better alignment than the RNN-based models. This improved performance likely stems from the LLM's ability to draw on a broader and more diverse information set, including not only quantitative economic indicators, but also text-based inputs such as forward guidance and macroeconomic narratives that shape policy expectations. Moreover, as the LLM incorporates qualitative information, it respects the zero lower bound constraining the policy rate not to fall below 25 bps, a constraint that the RNNs violate as it was not explicitly included in the models' setup. The ability to process such qualitative context allows the LLM to complement traditional inputs and produce forecasts that better reflect the Fed's decision-making environment. Note, however, that this is offset by the finding that prior to the COVID-19 pandemic, the LLM predictions exhibit an upward bias vis-a-vis the observed policy rates. Taken at face value, the LLM suggests the policy rate could have been slightly higher during the pandemic period, in agreement with the RNNs' predictions.

Figure 4: Average Predictions of LLM vs Actual Fed Fund Rate




Sources: FRED, Haver Analytics, and authors' calculations.

The interpretive capacity of the LLM is illustrated in the forecast plot, where the average prediction from 100 replications follows the actual policy rate closely, while the spread of individual simulations highlights the range of plausible paths the model considered. In itself, this is an interesting finding. While large language models (LLMs) are often associated with variability and hallucinations—particularly in natural language tasks such as open-ended question answering or creative writing—their behavior in structured numerical forecasting tasks appears markedly different. In our setting, where the LLM is tasked with forecasting the federal funds rate based on a fixed set of numerical and textual economic indicators, we observe remarkably low variance across 100 replications using the same input. This consistency contrasts sharply with LLM performance in more generative domains, where outputs can vary significantly between runs, even with identical prompts.

LLM hallucinations in certain domains typically occur because the model must infer or "fill in" missing or underspecified information, drawing on broad patterns learned during pretraining. These inferences may result in plausible but factually inaccurate responses, especially when the model is asked to generate open-ended text without grounding in verifiable data. In contrast, when used for numerical forecasting with well-specified inputs and a constrained output space, the model's generative behavior is more tightly anchored to the data. As a result, the scope for hallucination is limited, and the model tends to produce consistent and stable predictions.

To evaluate the LLM's predictive performance and the absence of hallucinations more systematically we compute three standard error metrics, MSE, RMSE, and MAE, for each of the 100 replications. On average, the model achieves an MSE of approximately 0.0311, an RMSE of 0.1762, and an MAE of 0.1483. The distribution of errors is relatively narrow, with RMSE values ranging from 0.158 to 0.201 and MAE values between 0.140 and 0.161, indicating a high degree of stability in the model's outputs across repeated runs. These results indicate the LLM provides consistently accurate forecasts of the federal funds rate, particularly when considering both squared and absolute deviations. Table 3 summarizes the distribution of the three error metrics.

Table 3: LLM Forecast Errors: simulation summary statistics

| **Statistic** | **MSE** | **RMSE** | **MAE** |
|---|---|---|---|
| Number of replications | 100 | 100 | 100 |
| Mean | 0.0311 | 0.1762 | 0.1483 |
| Std. deviation | 0.0034 | 0.0097 | 0.0045 |
| Minimum | 0.0250 | 0.1581 | 0.1398 |
| 10th Percentile | 0.0274 | 0.1656 | 0.1425 |
| 25th Percentile | 0.0286 | 0.1692 | 0.1452 |
| Median | 0.0308 | 0.1754 | 0.1474 |
| 75th Percentile | 0.0334 | 0.1829 | 0.1516 |
| 90th Percentile | 0.0355 | 0.1885 | 0.1542 |
| Maximum | 0.0405 | 0.2013 | 0.1606 |

Note: MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) measure the average size of forecast errors. All errors are in percentage points unless otherwise indicated. Lower values mean more accurate estimates.
Sources: Haver Analytics and authors' calculations.

We now examine two episodes where the model's forecast diverges from the actual policy decision and prompt the LLM to explain each discrepancy. The first occurs in early 2020, when the LLM predicts a rate cut to the zero lower bound. The second is in early 2024, where the model suggests that the policy rate should be lower than the rate actually observed. Since there were 100 replications, we present below a representative sample of the LLM's explanations for each deviation.

For the March 2020 deviation, the LLM suggests that a rate cut is imminent in order to respond rapidly to the severe economic disruption. One explanation is as follows:

> *Given severe economic disruption from COVID-19 pandemic, including sharp rise in unemployment and negative GDP growth, the Fed is likely to cut rates to near-zero to provide maximum monetary stimulus. Recent data shows rapid economic deterioration requiring urgent policy response.*

The LLM's explanations suggest that a cautious rate cut could be considered as inflation moderates, while the labor market remains strong and GDP growth stays positive. A few such explanations are as follows:

> *Based on recent data, inflation is moderating but remains above target. Unemployment is low and GDP growth is steady. The Fed may maintain a slightly restrictive stance to ensure inflation continues downward, but could consider a small rate cut if economic conditions warrant.*

> *Based on recent economic data, inflation appears to be moderating but remains above the Fed's 2% target. Labor market remains strong with low unemployment. GDP growth is positive but slowing. The Fed may continue its cautious approach, potentially implementing a small rate cut to support economic growth while monitoring inflation closely.*

> *Based on recent economic data, inflation appears to be moderating but remains above the Fed's 2% target. Labor market remains tight with low unemployment. GDP growth is steady. The Fed may maintain current rates or make a small cut to support economic growth while ensuring inflation continues to decline.*

While these findings are encouraging, it is important to acknowledge a potential limitation. The model, as any other LLM, may benefit from hindsight or lookahead bias (Sarkar and Vafa 2024), as its underlying training data could include documents or narratives released after the forecast dates. This access to future knowledge might partially explain the model's apparent ability to anticipate policy decisions and possibly its relative performance deterioration at the end of the sample.[8] Caution is therefore warranted when interpreting these results. That said, this concern is less relevant for forward-looking forecasts, where the model necessarily lacks access to future outcomes and must rely solely on contemporaneous information.

8. One possible safeguard is to mask or remove explicit date information from the input dataset to prevent the model from anchoring predictions to known historical events. However, this strategy may only offer limited protection, as many macroeconomic indicators contain temporal signals, such as seasonality or structural shifts, that can implicitly reveal the time period. Thus, even without date labels, the model may still infer approximate timing. Ludwig, Mullainathan, and Rambachan (2025) suggests some potential solutions, such as using only open-source LLMs specifying knowledge cutoff dates. Even in this case, however, the effective cutoff date might differ from the official date (Cheng et al. 2024).

## 5 Conclusions

Forecasting the U.S. federal funds rate remains an analytically complex yet policy-relevant task, given its central role in transmitting monetary policy to financial conditions, capital flows, and macroeconomic outcomes. The inherent difficulty lies in the discretionary and evolving nature of monetary policy, which must weigh a broad set of economic, financial, and geopolitical considerations—many of which are difficult to formalize or quantify within standard econometric frameworks.

This paper investigates the potential of artificial intelligence (AI)–based approaches to augment monetary policy analysis, focusing on the predictive capabilities of recurrent neural networks (LSTM and GRU) and a large language model (LLM) in forecasting the federal funds rate. While LSTM and GRU architectures are well-suited for modeling temporal dependencies in structured numerical data, the LLM introduces a novel dimension by leveraging unstructured textual inputs, including central bank communication and economic narratives, which may carry implicit forward guidance or sentiment cues not captured by quantitative indicators.

Empirical results indicate that all three models generate forecasts that are broadly consistent with observed policy decisions. The LLM demonstrates a modest performance edge, particularly in episodes where qualitative signals, such as shifts in communication strategy or geopolitical shocks, appear to exert a stronger influence on policy outcomes. The RNN models accuracy declines during the zero lower bound period and the early phase of the COVID-19 pandemic, likely reflecting high uncertainty and deviations from historical policy rules. Forecast performance improves post-pandemic, as macroeconomic conditions and policy responses begin to normalize.

The decision rules inferred from the RNN-based models suggest a systematic preference for more frequent rate adjustments than observed in actual FOMC behavior. Instead, the empirical record points to a pattern of policy inertia, with the FOMC maintaining rates at the lower or upper bound of the range for prolonged periods—potentially reflecting a shift toward a more gradualist policy stance, consistent with the adoption of the Flexible Average Inflation Targeting (FAIT) framework.

Overall, the findings underscore the potential of AI-based forecasting tools to complement traditional policy analysis. While such models are not a substitute for structural economic reasoning or institutional insight, they can serve as valuable inputs in environments characterized by uncertainty, narrative-driven dynamics, and evolving policy frameworks. As the role of qualitative information in monetary policy continues to expand, the integration of language-based models into forecasting toolkits may offer a promising avenue for future research and operational use.

## References

Bollerslev, Tim. 1986. “Generalized Autoregressive Conditional Heteroskedasticity.” *Journal of Econometrics* 31 (3): 307–327. https://doi.org/10.1016/0304-4076(86)90063-1.

Cao, Jianlong, Zhi Li, and Jingwei Li. 2019. “Financial time series forecasting model based on CEEMDAN and LSTM.” *Physica A: Statistical Mechanics and its Applications* 519:127–139.

Carriero, Andrea, Davide Pettenuzzo, and Shubhranshu Shekhar. 2025. *Macroeconomic Forecasting with Large Language Models.* arXiv: 2407.00890 [econ.EM]. https://arxiv.org/abs/2407.00890.

Chan-Lau, Jorge A., and Toan Long Quach. 2025. Shock propagation in LSTM multivariate time series systems. Working paper. ASEAN+3 Macroeconomic Research Office.

Cheng, Jeffrey, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. *Dated Data: Tracing Knowledge Cutoffs in Large Language Models.* https://doi.org/10.48550/arXiv.2403.12958. arXiv: 2403.12958.

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches.” In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation,* edited by Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi, 103–111. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4012.

Dai, Huning. 2025. “Quantitative Investment Analysis Based on GRU Networks.” In *Frontier Computing: Vol 3,* edited by Jason C. Hung, Neil Yen, and Jia-Wei Chang, 108–117. https://doi.org/10.1007/978-981-96-2798-1_12.

Faria-e-Castro, Miguel, and Fernando Leibovici. 2024. “Artificial Intelligence and Inflation Forecasts.” *Federal Reserve Bank of St. Louis Review* 106 (12): 1–14. https://www.stlouisfed.org/publications/review/2024/nov/artificial-intelligence-and-inflation-forecasts.

Furizal, Furizal, Asdelina Ritonga, Alfian Ma’arif, and Iswanto Suwarno. 2024. “Stock price forecasting with multivariate time series long-short term memory: a deep learning approach.” *Journal of Robotics and Control* 5 (5). https://doi.org/10.18196/jrc.v5i5.22460.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. “Predictably Unequal? The Effects of Machine Learning on Credit Markets.” *Journal of Finance* 77 (1): 5–47. https://doi.org/10.1111/jofi.13084. https://doi.org/10.1111/jofi.13084.

Garnett, Roman. 2023. *Bayesian Optimization.* Cambridge University Press.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* Cambridge, MA: MIT Press. ISBN: 9780262035613. https://www.deeplearningbook.org/.

Gu, Jincheng, Shiqi Zhang, Yanling Yu, and Feng Liu. 2024. “AB-LSTM-GRU: a novel ensemble composite deep neural network model for exchange rate forecasting.” *Computational Economics,* https://doi.org/10.1007/s10614-024-10754-7.

Guo, Song, and Chengmin Wang. 2024. "Research on ice and snow toursim economic forecasting method based on improved gated recurrent unit model." In *ICCSIE '24: Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering,* 388–394. Association for Computing Machinery.

Hamiane, Sana, Yousseff Ghanou, Hamid Khalifi, and Meryam Telmem. 2024. "Comparative analysis of LSTM, ARIMA, and hybrid models for forecasting future GDP." *Ingénierie des Systèmes d'Information* 29 (3): 853–861. https://doi.org/10.18280/isi.290306.

Hamiane, Sana, Hamid Khalifi, Youssef Ghanou, and Gabriella Casalino. 2023. "Forecasting the Gross Domestic Product using LSTM and ARIMA." In *2023 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD),* 1–6. https://doi.org/10.1109/ICTMOD59086.2023.10438159.

Hao, Yuzhi, and Danyang Xie. 2025. *A Multi-LLM-Agent-Based Framework for Economic and Public Policy Analysis.* arXiv: 2502.16879 [cs.AI]. https://arxiv.org/abs/2502.16879.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hollis, Thomas, Antoine Viscardi, and Seung Eun Yi. 2018. A Comparison of LSTMs and Attention Mechanisms for Forecasting Financial Time Series. Working paper. http://arxiv.org/abs/1812.07699.

Hopp, Daniel. 2022. "Economic nowcasting with long short-term memory artificial neural networks (LSTM)." *Journal of Official Statistics* 38 (3): 847–873. https://doi.org/https://doi.org/10.2478/jos-2022-0.

Horton, John J. 2023. "Large Language Models as simulated economic agents." *arXiv preprint arXiv:2301.07543,* https://arxiv.org/abs/2301.07543.

Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever. 2015. "An empirical exploration of recurrent network architectures." In *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37,* edited by Francis Bach and DAvid Blei, 2342–2350. https://doi.org/10.5555/3045118.3045367.

Ju, Jie, and Fang-Ai Liu. 2021. "Multivariate time series data prediction based on ATT-LSTM network." *Applied Sciences* 11 (9373).

Kim, Alex, Maximilian Muhn, and Valeri Nikolaev. 2025. *Financial Statement Analysis with Large Language Models.* arXiv: 2407.17866 [q-fin.ST]. https://arxiv.org/abs/2407.17866.

Kobiela, Dariusz, Dawid Krefta, Weronika Krol, and Pawel Weichbroth. 2022. "ARIMA vs LSTM on NASDAQ stock exchange data." *Procedia Computer Science* 207:3836–3845. https://doi.org/10.1016/j.procs.2022.09.445.

Lakshmi Narayanaa, T, R R Skandarsini, S. Jhansi Ida, S. Rathana Sabapathy, and P Nanthitha. 2023. "Inflation Prediction: A Comparative Study of ARIMA and LSTM Models Across Different Temporal Resolutions." In *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA),* 1390–1395. https://doi.org/10.1109/ICIMIA60377.2023.10425970.

Lashina, Maria, and Sergey Grishunin. 2023. "Comparison of forecasting power of statistical models for GDP growth under conditions of permanent crises for application in strategic risk controlling." *Procedia Computer Science* 221:442–449. https://doi.org/10.1016/j.procs.2023.07.059.

Lawi, Armin, Hendra Mesra, and Supri Amir. 2021. "Implementation of long short-term memory and gated recurrent units on grouped time-series data to predict stock prices accurately." *Journal of Big Data* 9:89. https://doi.org/doi.org/10.1186/s40537-022-00597-0.

Liu, Bingchun, and Mingzhao Lai. 2025. "Advanced machine learning for financial markets: a PCA-GRU-LSTM approach." *Journal of the Knowledge Economy* 16:3140–3174. https://doi.org/10.1007/s13132-024-02108-3.

Liu, Sixing, and Quanxiang Lan. 2025. "LSTM economic forecasting model based on Adam optimization." *ICCSMT '24: Proceeding of the 2024 5th International Conference on Computer Science and Management Technology,* 1408–1413.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2025. Large language models: an applied econometric framework. NBER Working Paper 33344. National Bureau of Economic Research.

Molnar, Christoph. 2022. *Interpretable Machine Learning.* 2nd ed. Leanpub. https://christophm.github.io/interpretable-ml-book/.

Naas, Meryem-Nadjat, and Habib Zouaoui. 2024. "Forecasting foreign exchange rate volatility using deep learning: case of US dollar/Algerian dinar during the COVID-19 pandemic." *Research Papers in Economics and Finance* 8 (1). https://doi.org/10.18559/ref.2024.1.1172.

Ormaniec, Weronika, Marcin Pitera, Sajad Safarveisi, and Thorsten Schmidt. 2022. Estimating value at risk: LSTM vs. GARCH. Working paper. https://doi.org/https://arxiv.org/abs/2207.10539.

Paranhos, Livia. 2025. "Predicting inflation with recurrent neural networks." *International Journal of Forecasting,* https://doi.org/10.1016/j.ijforecast.2024.07.010.

Park, Sanjing, and Jae-Suk Yang. 2022. "Interpretable deep learning LSTM model for intelligent economic decision-making." *Knowledge-Based Systems* 248 (108907). https://doi.org/10.1016/j.knosys.2022.108907.

Patel, Soham, Dhyey Sanghavi, and Artika Singh. 2023. "Forecasting Economic Trends: Analyzing FRED Data with Time Series Methods and RNNs." In *2023 Global Conference on Information Technologies and Communications (GCITC),* 1–7. https://doi.org/10.1109/GCITC60406.2023.10426363.

Pilla, Prashant Reddy, and Raji Mekonen. 2025. Forecasting S&P 500 Using LSTM Models. Working paper. https://doi.org/10.5281/ZENODO.14759118.

Pirani, Muskaan, Paurav Thakkar, Pranay Jivrani, Mohammed Husain Bohara, and Dweepna Garg. 2022. "A Comparative Analysis of ARIMA, GRU, LSTM and BiLSTM on Financial Time Series Forecasting." In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE),* 1–6. https://doi.org/10.1109/ICDCECE53908.2022.9793213.

Salem, Abdorfwf A Mohamed, Amaal Jummah, and Abdullah Albourawi. 2024. “Predictive models for interest rate forecasting using machine learning: a comparative analysis and practical applications.” *Brilliance: Research of Artificial Intelligence* 4 (2). https://doi.org/10.47709/brilliance.v4i2.4983.

Sarkar, Suproteem K., and Keyon Vafa. 2024. Lookahead bias in pretrained language models. Technical report. Harvard University. http://dx.doi.org/10.2139/ssrn.4754678.

Serin, Selim, and Gülder Kemalbay. 2024. “LSTM-based multivariate deep neural networks for stock price forecasting.” In *Güncel Ekonometrik ve İstatistiksel Uygulamalar ile Akademik Çalışmalar.* Özgür Yayınları.

Sezer, Omer Berat, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2020. “Financial time series forecasting with deep learning: A systematic literature review: 2005–2019.” *Applied Soft Computing* 90:106181. https://doi.org/10.1016/j.asoc.2020.106181. https://www.sciencedirect.com/science/article/pii/S1568494620300960.

Shiri, Farhad Mortezapour, Thinagaran Perumal, Norwati Mustapha, and Raihani Mohamed. 2024. “A comprehensive overview and comparative analysis on deep learning models.” *Journal on Artificial Intelligence* 6:301–360. https://doi.org//doi.org/10.32604/jai.2024.054314.

Sivakumar, Guhan. 2025. HMM-LSTM Fusion Model for Economic Forecasting. Working paper. https://arxiv.org/abs/2501.02002.

Umezuruike, Chinecherem, Deborah Olaniyan, Julius Olaniyan, Abidemi Emmanuel Adeniyi, Adedoyin Oyebade, and David Abaneme. 2024. “Comparative Analysis of Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and Transformer Models in Predicting Stock Prices.” In *2024 IEEE 5th International Conference on Electro-Computing Technologies for Humanity (NIGERCON),* 1–6. https://doi.org/10.1109/NIGERCON62786.2024.10927198.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is All you Need.” In *Advances in Neural Information Processing Systems,* edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol. 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. *BloombergGPT: A Large Language Model for Finance.* arXiv: 2303.17564 [cs.LG]. https://arxiv.org/abs/2303.17564.

Xie, Huaqing, Xingcheng Xu, Fangjia Yan, Xun Qian, and Yanqing Yang. 2024. Deep Learning for Multi-Country GDP Prediction: A Study of Model Performance and Data Impact. Working paper. https://arxiv.org/abs/2409.02551.

Xu, Ke, Yu Cheng, Shiqing Long, Junjie Guo, Jue Xiao, and Mengfang Sun. 2024. Advancing Financial Risk Prediction Through Optimized LSTM Model Performance and Comparative Analysis. Working paper. https://arxiv.org/abs/2405.20603.

Zhang, Aston, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 2023. *Dive into Deep Learning.* Https://D2L.ai. Cambridge University Press.

Zhang, Junhuan, Jiaqi Wen, and Zhen Yang. 2022. "China's GDP forecasting using long short term memory recurrent neural network and hidden markov model." *PLOS One* 17 (6): e0269529. https://doi.org/10.1371/journal.pone.0269529.

Zhang, Xuan, Xun Liang, Aakas Zhiyuli, Shusen Zhang, Rui Xu, and Bo Wu. 2019. "AT-LSTM: and attention-based LSTM model for financial time series prediction." *IOP Conference Series: Materials Science and Engineering* 569 (5). https://doi.org/10.1088/1757-899X/569/5/05237.

Zhao, Jiayao. 2024. "The Macroeconomic Prediction Model based on LSTM Improved Algorithm." In *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS),* 1–7. https://doi.org/10.1109/ICICACS60521.2024.10498821.

Zhu, Dongge, Jiahui Zhang, and Xinxin" Tan. 2025. "Considering Power Factors in Macroeconomic Forecasting Based on the Gated Recurrent Unit (GRU) Model." In *Proceedings of the 5th International Conference on Big Data Analytics for Cyber-Physical System in Smart City—Volume 1,* edited by Mohammed Atiquzzaman, Neil Yen, and Zheng Xu, 279–290. https://doi.org/10.1007/978-981-96-0208-7_26,.

Zhu, Peng, Yuante Li, Yifan Hu, Sheng Xiang, Qinyuan Liu, Dawei Cheng, and Yuqi Liang. 2025. "MCI-GRU: stock prediction model based on mlti-head corss-attention and imprvoed GRU." *Neurocomputing* 638 (130168). https://doi.org/10.1016/j.neucom.2025.130168.

# A LSTM and GRU models: mathematical formulation

This appendix draws heavily on A. Zhang et al. (2023).

## A.1 Long Short-Term Memory (LSTM) Networks

LSTMs manage information through a sophisticated memory cell mechanism that can selectively retain, update, or forget information across long sequences. The key innovation lies in memory cells equipped with multiplicative gates that regulate information flow, enabling the network to capture long-term dependencies in sequential data. Figure A1 shows the internal structure of the LSTM memory cell and the computation flow, which are explained next.

Each memory cell is equipped with an internal state and a number of multiplicative gates that determine whether (i) a given input should impact the internal state (the input gate), (ii) the internal state should be flushed to 0 (the forget gate), and (iii) the internal state of a given neuron should be allowed to impact the cell's output (the output gate).

#### Input Gate, Forget Gate, and Output Gate

The data feeding into the LSTM gates are the input at the current time step and the hidden state of the previous time step. Three fully connected layers compute the values of the input, forget, and output gates. Formally, let $\mathbf{X}_t \in \mathbb{R}^{n\times d}$ be the minibatch input and $\mathbf{H}_{t-1} \in \mathbb{R}^{n\times h}$ the hidden state in period $t$, where $n$ is the batch size and $d$ the size of an input observation. Correspondingly, the gates at time step $t$ are defined as follows: the input gate, $\mathbf{I}_t \in \mathbb{R}^{n\times h}$; the forget gate $\mathbf{F}_t \in \mathbb{R}^{n\times h}$; and the output gate $\mathbf{O}_t \in \mathbb{R}^{n\times h}$. They are defined as:

$$\begin{aligned}
\mathbf{I}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{\mathrm{xi}} + \mathbf{H}_{t-1}\mathbf{W}_{\mathrm{hi}} + \mathbf{b}_{\mathrm{i}}),\\
\mathbf{F}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{\mathrm{xf}} + \mathbf{H}_{t-1}\mathbf{W}_{\mathrm{hf}} + \mathbf{b}_{\mathrm{f}}),\\
\mathbf{O}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{\mathrm{xo}} + \mathbf{H}_{t-1}\mathbf{W}_{\mathrm{ho}} + \mathbf{b}_{\mathrm{o}}),
\end{aligned}$$

where $\sigma$ is the sigmoid activation function, $\mathbf{W}_{\mathrm{xi}}, \mathbf{W}_{\mathrm{xf}}, \mathbf{W}_{\mathrm{xo}} \in \mathbb{R}^{d\times h}$ and $\mathbf{W}_{\mathrm{hi}}, \mathbf{W}_{\mathrm{hf}}, \mathbf{W}_{\mathrm{ho}} \in \mathbb{R}^{h\times h}$ are weight parameters and $\mathbf{b}_{\mathrm{i}}, \mathbf{b}_{\mathrm{f}}, \mathbf{b}_{\mathrm{o}} \in \mathbb{R}^{1\times h}$ are bias parameters.

#### Input Node

The input node $\tilde{\mathbf{C}}_t \in \mathbb{R}^{n\times h}$, an auxiliary node, plays an important role in the acquisition of new information. Its computation at time step $t$ is given by:

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{X}_t\mathbf{W}_{\mathrm{xc}} + \mathbf{H}_{t-1}\mathbf{W}_{\mathrm{hc}} + \mathbf{b}_{\mathrm{c}}),$$

where $\mathbf{W}_{\mathrm{xc}} \in \mathbb{R}^{d\times h}$ and $\mathbf{W}_{\mathrm{hc}} \in \mathbb{R}^{h\times h}$ are weight parameters and $\mathbf{b}_{\mathrm{c}} \in \mathbb{R}^{1\times h}$ is a bias parameter.

#### Memory Cell Internal State

In LSTMs, the input gate $\mathbf{I}_t$ governs how much of the new data is acounted for via $\tilde{\mathbf{C}}_t$ and the forget gate $\mathbf{F}_t$ addresses how much of the old cell internal state $\mathbf{C}_{t-1} \in \mathbb{R}^{n\times h}$ is retained. Together, they update the internal state according to:

$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t.$$

where $\odot$ is the Hadamard (elementwise) product operator.

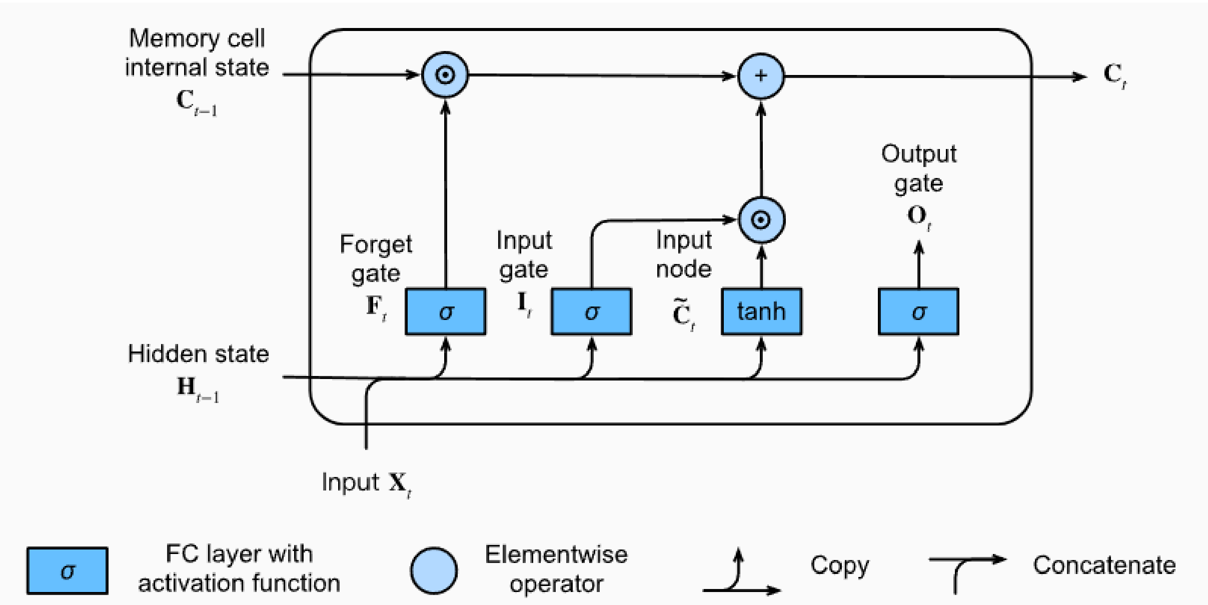
#### Hidden State

Finally, the hidden state $\mathbf{H}_t$ is computed using the output gate and the memory cell internal state:

$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t),$$

which together with the cell internal state, is passed to the next LSTM memory cell. Once the final memory cell is reached, the final hidden state is input into a fully connected output projection head consisting of a dropout layer followed by two linear transformations with a ReLU activation to predict the policy rate.

Figure A1: LSTM memory cell




Source: A. Zhang et al. (2023).

### A.2 Gated Recurrent Unit (GRU) Networks

The GRU manages information through a streamlined gating mechanism that controls how much past information to retain and how much new information to incorporate at each time step. The key difference with the LSTM lies in a simpler two-gate structure while maintaining the ability to capture long-term dependencies in sequential data with reduced computational complexity. Figure A2 shows the GrU memory cell structure and computation flow, which are described next.

#### Reset Gate and Update Gate

There are two gates, the reset gate and the update gate, with sigmoid activation functions. The reset gate controls how much of the previous state information is preserved and the update gate controls how much of the information in the new state is the same as in the old state.

Formally, for a given time step $t$ let $\mathbf{X}_t \in \mathbb{R}^{n\times d}$ be the minibatch input and $\mathbf{H}_{t-1} \in \mathbb{R}^{n\times h}$ the state in the previous time step. The reset gate $\mathbf{R}_t \in \mathbb{R}^{n\times h}$ and update gate $\mathbf{Z}_t \in \mathbb{R}^{n\times h}$ are:

$$\mathbf{R}_t = \sigma(\mathbf{X}_t\mathbf{W}_{\mathrm{xr}} + \mathbf{H}_{t-1}\mathbf{W}_{\mathrm{hr}} + \mathbf{b}_{\mathrm{r}}),$$
$$\mathbf{Z}_t = \sigma(\mathbf{X}_t\mathbf{W}_{\mathrm{xz}} + \mathbf{H}_{t-1}\mathbf{W}_{\mathrm{hz}} + \mathbf{b}_{\mathrm{z}}),$$

where $\mathbf{W}_{\mathrm{xr}}, \mathbf{W}_{\mathrm{xz}} \in \mathbb{R}^{d \times h}$ and $\mathbf{W}_{\mathrm{hr}}, \mathbf{W}_{\mathrm{hz}} \in \mathbb{R}^{h \times h}$ are weight parameters and $\mathbf{b}_{\mathrm{r}}, \mathbf{b}_{\mathrm{z}} \in \mathbb{R}^{1 \times h}$ are bias parameters.

**Candidate Hidden State**

The candidate hidden state $\tilde{\mathbf{H}}_t \in \mathbb{R}^{n \times h}$ in the GRU plays the same roll as the memory cell internal state in LSTM. It is computed using the reset gate:

$$\tilde{\mathbf{H}}_t = \tanh(\mathbf{X}_t \mathbf{W}_{\mathrm{xh}} + (\mathbf{R}_t \odot \mathbf{H}_{t-1}) \mathbf{W}_{\mathrm{hh}} + \mathbf{b}_{\mathrm{h}}),$$

where $\mathbf{W}_{\mathrm{xh}} \in \mathbb{R}^{d \times h}$ and $\mathbf{W}_{\mathrm{hh}} \in \mathbb{R}^{h \times h}$ are weight parameters and $\mathbf{b}_{\mathrm{h}} \in \mathbb{R}^{1 \times h}$ is a bias parameter.

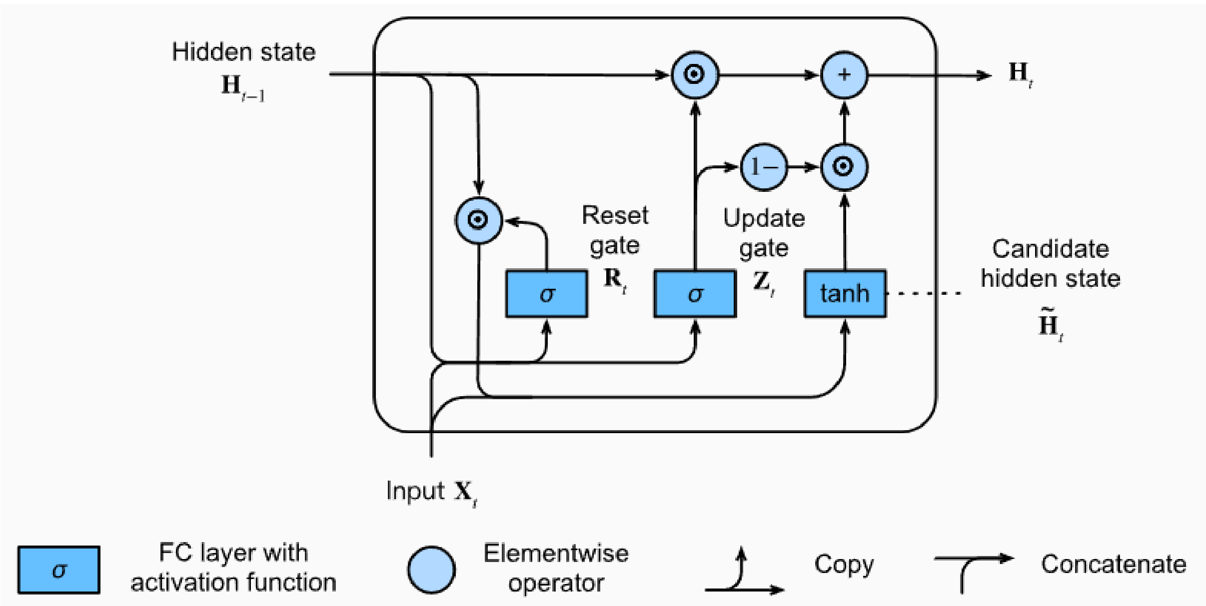
**Hidden State Update**

The final hidden state $\mathbf{H}_t$ is computed as a convex combination of the previous hidden state and the candidate hidden state:

$$\mathbf{H}_t = \mathbf{Z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{H}}_t.$$

Similarly to the LSTM, the cell internal state and hidden states are passed to the next GRU memory cell sequentailly until reaching the final memory cell. Then the hidden state serves as an input to a projection head sharing the same architecture as that in the LSTM model to predict the policy rate.

Figure A2: GRU memory cell




Source: A. Zhang et al. (2023).

[This page is intentionally left blank]

Address: 10 Shenton Way, #15-08
MAS Building, Singapore 079117
Website: www.amro-asia.org
Tel: +65 6323 9844
Email: enquiry@amro-asia.org