

A Rigorous Statistical Framework for Long Short-Term Memory Networks: State-Space Operator Dynamics, Continuous-Time Neural SDEs, and Informational Efficiency

Cosmas Kaitani Nziku

`nziku.cosmas@udsm.ac.tz`

University of Dar es Salaam

Article

Keywords: LSTM, recurrent neural networks, stochastic differential equations, neural SDEs, PAC-Bayes

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11314069/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Cosmas Kaitani Nziku^{1*}

¹ Department of Statistics, University of Dar es Salaam, Dar es Salaam-Tanzania

*Corresponding Author: nziku.cosmas@udsm.ac.tz

Abstract

The statistical elucidations of Long Short-Term Memory (LSTM) models are usually distinct from their dynamical and information processing capabilities, even though LSTM models have drawn a lot of attention and become popular for modelling time series with dependent sequential data. This paper develops a unified framework in which an LSTM is treated as a nonlinear stochastic state-space system, an operator acting on latent memory, and, under an explicit diffusion scaling, a continuous-time stochastic dynamical system. The framework has four components. First, the discrete LSTM recursion is formulated as a measurable Markov state-space model with bounded gates and a potentially unbounded cell state. Second, a continuous-time neural stochastic differential equation (SDE) is introduced through a well-defined scaling limit rather than by an informal division of the discrete recursion by the sampling interval. Third, statistical inference is analysed through recursive parameter sensitivities, a conditional likelihood, the Fisher information matrix, and a martingale central limit argument under stationarity, mixing, and non-singularity assumptions. Fourth, sequential PAC-Bayesian bounds are used to relate predictive risk to posterior complexity and information divergence. We prove existence and uniqueness of the neural SDE under global Lipschitz and linear-growth conditions, derive a sufficient mean-square stability condition for a dissipative continuous-time model, establish consistency and asymptotic normality of the score-based estimator under standard geometric-ergodicity assumptions, and state a sequential PAC-Bayesian risk bound. The paper is deliberately careful about scope: none of the theoretical results implies that every LSTM is automatically stable, statistically efficient, or optimal. Those properties follow only under the assumptions stated in the corresponding results. The proposed simulation protocol is therefore presented as a validation design rather than as completed empirical evidence.

Keywords

LSTM; recurrent neural networks; stochastic differential equations; neural SDEs; PAC-Bayes

1. Introduction

Recurrent neural networks are dynamical statistical models as much as they are predictive algorithms. Their hidden states compress past observations into a finite-dimensional representation, while their parameters determine how information is retained, transformed, and released (Shalev-Shwartz & Ben-David, 2013). LSTMs add input, forget, and output gates to regulate this compression. The resulting architecture has been exceptionally successful in forecasting, language processing, control, and other sequential tasks, but its statistical analysis is more fragmented than its empirical use would suggest (Hochreiter & Schmidhuber, 1997).

The original draft motivating this paper proposed a measure-theoretic state-space representation, a continuous-time neural SDE, Fisher-information analysis, asymptotic efficiency, and PAC-Bayesian generalization bounds (Neal, 1996) (Gu et al., 2021). In particular, a continuous-time LSTM is not obtained merely by taking a formal limit of an arbitrary discrete recursion. A valid diffusion limit requires a specified time scaling and

convergence assumptions (Krichen & Mihoub, 2025). Likewise, memory preservation does not by itself guarantee nondegenerate Fisher information, and bounded activations do not automatically bound the KL divergence of an arbitrary posterior. These distinctions are essential for a rigorous statistical treatment (Vapnik, 1999).

The main contributions are therefore formulated as follows. (i) give an exact stochastic state-space representation of an LSTM with an explicit observation model. (ii) formulate a continuous-time gated latent-state SDE whose drift contains the same gate parameterization as the discrete LSTM while its diffusion term is controlled separately. (iii) prove well-posedness and provide a sufficient Lyapunov condition for mean-square stability. (iv) derive recursive sensitivity equations and state sufficient conditions for consistency and asymptotic normality of likelihood-based estimation. (v) give a sequential PAC-Bayesian bound and explain precisely what the bound says about information efficiency. (vi) provide a reproducible simulation and empirical-validation protocol without fabricating numerical results.

2. Mathematical Setup and Notation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space equipped with a filtration $(\mathcal{F}_t)_{t \geq 0}$ satisfying the usual conditions. For the discrete model, let $X_t \in \mathbb{R}^p$, be an observed input process taking values in $\mathbb{R}^q$, and let $Y_t \in \mathbb{R}^q$ be the target process in $\mathbb{R}^q$ (Billingsley, 1995). The LSTM latent state is $Z_t = (C_t, H_t)$, where $C_t, H_t \in \mathbb{R}^d$ is the cell state in $\mathbb{R}^d$ and H_t is the hidden state in $\mathbb{R}^d$ (Vapnik, 1999). Let the complete parameter vector be $\theta \in \Theta \subset \mathbb{R}^m$, where Θ is assumed compact whenever required for statistical asymptotic arguments (Vaart, 1998).

The logistic sigmoid is denoted by $\sigma(a) = \frac{1}{1+e^{-a}}$, applied componentwise, and the hyperbolic tangent by $\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$. Their global Lipschitz constants can be taken as 1/4 and 1, respectively. The componentwise product is written $\odot$, and $\|\cdot\|$ denotes the Euclidean norm unless a matrix norm is explicitly indicated.

$$\sigma(a) = \frac{1}{1+e^{-a}}, \quad \tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}. \quad (2.1)$$

$$0 < \sigma(a) < 1, \quad |\sigma(a) - \sigma(b)| \leq \frac{1}{4} |a - b|, \quad |\tanh(a) - \tanh(b)| \leq |a - b|. \quad (2.2)$$

These elementary bounds are central to the analysis. They imply bounded gates and controlled local sensitivity, but they do not by themselves imply contraction of the complete recurrent map because the cell-state carry term can have gain arbitrarily close to one (Pascanu & Mikolov, 2012).

3. Discrete LSTM as a Nonlinear Stochastic State-Space Model

3.1 Gated recursion

For $t = 1, 2, \dots$ define the gate vectors from the previous hidden state and current input. The input, forget, output, and candidate gates are

$$\begin{aligned} I_t &= \sigma(W_i X_t + U_i H_{t-1} + b_i), \\ F_t &= \sigma(W_f X_t + U_f H_{t-1} + b_f), \\ O_t &= \sigma(W_o X_t + U_o H_{t-1} + b_o), \end{aligned} \quad (3.1)$$

and

$$G_t = \tanh(W_g X_t + U_g H_{t-1} + b_g). \quad (3.2)$$

The cell and hidden states evolve according to

$$\begin{aligned} C_t &= F_t \odot C_{t-1} + I_t \odot G_t. \\ H_t &= O_t \odot \tanh(C_t). \end{aligned} \quad (3.3)$$

The complete latent transition can therefore be written as $Z_t = \Phi_\theta(Z_{t-1}, X_t)$, where Φ_θ is measurable and continuous in its state argument under ordinary finite-valued parameterization (Øksendal, 2003).

3.2 Observation model

A convenient statistical observation model is a conditionally Gaussian output. Let R be a positive-definite q by q covariance matrix and let μ_θ be a measurable readout map (Sun et al., 2020). Then

$$Y_t = \mu_\theta(H_t, X_t) + \varepsilon_t. \quad (3.4)$$

For the Gaussian case,

$$\varepsilon_t \mid \mathcal{F}_{t-1}, X_t \sim N(0, R), \quad (3.5)$$

where R is a positive-definite covariance matrix.

The conditional distribution is therefore

$$Y_t \mid \mathcal{F}_{t-1}, X_t \sim N(\mu_\theta, (H_t, X_t)R). \quad (3.6)$$

Other observation distributions can be incorporated, including Bernoulli, categorical, Poisson, or other exponential-family models, with the innovations conditionally independent over t given the latent trajectory and inputs (Hunter et al., 2008).

3.3 Markov representation

If the input process is itself Markov, then the augmented process $S_t = (X_t, Z_t)$ is Markov. If X_t is exogenous, the conditional transition kernel of Z_t given X_t is sufficient for likelihood construction (Durrett, 2019). The LSTM can be viewed as a nonlinear state-space system $Z_t = \Phi_\theta(Z_{t-1}, X_t)$, with observation equation $Y_t \sim p_\theta(\cdot \mid Z_t, X_t)$. This representation separates the latent dynamics from the observation mechanism. This observation separates the recurrent architecture from the stochastic assumptions imposed on the data-generating mechanism (Robbins & Monroe, 1951).

Proposition 1. Measurability and Markov state-space representation

Assume the input process is adapted, and the LSTM parameters are finite. Then the recursion above defines an adapted latent process (C_t, H_t) . Conditional on the input sequence, the latent transition is a time-inhomogeneous Markov kernel. If X_t is Markov and the transition uses only $(X_t, Z_{(t-1)})$, the augmented process (X_t, Z_t) is Markov (Meyn & Tweedie, 2009).

Proof. The affine maps in the gate equations

$$W_i X_t + U_i H_{t-1} + b_i, \quad (3.7)$$

are Borel measurable, and σ and $\tanh$ are continuous. Finite compositions of measurable maps are measurable, so Φ_θ is measurable. Since $Z_t = \Phi_\theta(Z_{t-1}, X_t)$ is a measurable function of (Z_{t-1}, X_t) , conditional distributions of future states given the past depend on the past only through the current state and current input. If X_t is Markov, adjoining Z_t preserves Markov property because the transition law of the pair depends only on the current pair, hence the augmented process is Markov (Meyn & Tweedie, 2009).

4. Stability and Memory in the Discrete Model

The hidden state has an immediate boundedness property because $|O_t| \leq 1$ and $|\tanh(C_t)| \leq 1$. Consequently, $|H_t| \leq 1$ component wise. However, this does not imply boundedness of the cell state C_t . Indeed,

$$C_t = F_t \odot C_{t-1} + I_t \odot G_t. \quad (4.1)$$

If the forget gate remains close to one for a prolonged period, the cell state may accumulate information over long horizons. Thus, stability of the full LSTM requires additional condition. A central distinction is needed between bounded hidden states and bounded cell states. Because $|H_t|$ is bounded component wise by one, H_t is automatically bounded. The cell state C_t need not be bounded: repeated multiplication by forget gates near one can accumulate the candidate input. Stability therefore requires a quantitative condition on the recurrent transition (Miller & Wang, 2006).

Theorem 1. A sufficient mean-square contraction condition

Suppose that, conditional on the input process, the random transition Φ_θ satisfies a uniform mean-square contraction condition: there exists $\rho \in (0, 1)$ and $b < \infty$ such that

$$\mathbb{E}[\|\Phi_\theta(Z, X_t) - \Phi_\theta(Z', X_t)\|^2 \mid X_t] \leq \rho \|Z - Z'\|^2 \quad (4.2)$$

for all states Z, Z' and all θ in Θ . Assume additionally that

$$\sup_{\theta, t} \mathbb{E}[\|\Phi_\theta(0, X_t)\|^2] \leq b < \infty. \quad (4.3)$$

uniformly in t and θ . Then the latent process has a unique stationary version for stationary inputs, and for any two solutions driven by the same inputs, their mean-square distance decays geometrically:

$$\mathbb{E} \|Z_t - Z'_t\|^2 \leq \rho^t \mathbb{E} \|Z_0 - Z'_0\|^2. \quad (4.4)$$

Proof. Iterating the assumed contraction gives the stated inequality by conditional expectation and the tower property. For existence, start two copies at times $-n$ and 0 , driven by the same stationary input sequence. The contraction makes the sequence Cauchy in L^2 – *Wasserstein space* (W_2) at each fixed time and it's bound defines a stationary solution. If two stationary solutions existed, their mean-square difference would equal the same quantity after every time shift, while the shrinkage forces that quantity to zero. Hence the stationary version is unique in L^2 .

By the contraction postulation,

$$\mathbb{E}[\|Z_t - Z'_t\|^2 | \mathcal{F}_{t-1}] \leq \rho \|Z_{t-1} - Z'_{t-1}\|^2. \quad (4.5)$$

Taking expectations gives

$$\mathbb{E} \|Z_t - Z'_t\|^2 \leq \rho \mathbb{E} \|Z_{t-1} - Z'_{t-1}\|^2. \quad (4.6)$$

Iterating,

$$\mathbb{E} \|Z_t - Z'_t\|^2 \leq \rho^t \mathbb{E} \|Z_0 - Z'_0\|^2. \quad (4.7)$$

Because

$$0 < \rho < 1, \quad (4.8)$$

the effect of the initial condition disappears geometrically.

As a substitute of being necessary, the theorem is purposefully sufficient. It does not claim that the contraction inequality is satisfied by all regular LSTMs. In practice, spectral normalization, recurrent-weight constraints, regularization, or limitations on the forget-gate working range can all be used to promote or enforce it.

5. Continuous-Time Neural SDE Formulation

5.1 Why an unambiguous scaling is required

The discrete cell update comprises an $O(1)$ carry term $F_t \odot C_{t-\Delta}$. If one merely subtracts C_{t-1} , divides by a time step Δ , and lets Δ approach zero while custody F_t fixed, the resulting drift generally diverges. Consequently, a continuous-time model necessitates a gate scaling of the form $F_t^{(\Delta)} = 1 - \Delta\Lambda_t + o(\Delta)$ or an equivalent parameterization.

5.2 Diffusion-scaled gated dynamics

Let X_t be an exogenous diffusion satisfying

$$dX_t = a(X_t, t) dt + B(X_t, t) dW_t^x, \quad (5.1)$$

where W_t^x is a Brownian motion.

Let Λ_t , A_t , and O_t be gate-like processes generated by bounded neural maps of the current input and hidden state (Russo et al., 2017). We define the continuous-time memory state by

$$dC_t = [-\Lambda_t \odot C_t + A_t \odot G_t]dt + \Sigma_c(Z_t, X_t, t) dW_t^c \quad (5.2)$$

The hidden state is

$$H_t = O_t \odot \tanh(C_t) \quad (5.3)$$

where G_t is a bounded neural candidate field. A more general formulation can evolve H_t directly as a diffusion, but the displayed system preserves the characteristic LSTM separation between memory and exposed hidden state. The Brownian motions W_t^x and W_t^c may be independent or correlated through a specified joint covariance structure.

5.3 Continuous-time gate parameterization

The continuous-time gates can be parameterized as

$$\begin{aligned}\Lambda_t &= \text{softplus}(K_f X_t + V_f H_t + c_f), \\ A_t &= \sigma(K_i X_t + V_i H_t + c_i), \\ O_t &= \sigma(K_o X_t + V_o H_t + c_o),\end{aligned}\tag{5.4}$$

and

$$G_t = \tanh(K_g X_t + V_g H_t + c_g)\tag{5.5}$$

The softplus function is

$$\text{softplus}(x) = \log(1 + e^x)\tag{5.6}$$

The use of softplus for Λ guarantees

$$\Lambda_t \geq 0\tag{5.7}$$

Thus, the continuous-time memory mechanism has an explicit decay interpretation. The softplus map is used for Λ_t because it guarantees nonnegative forgetting intensity. Other positive maps are possible. The resulting model is a neural diffusion whose drift and diffusion coefficients are state-dependent.

Theorem 2. Existence and uniqueness of the neural SDE

Let $S_t = (X_t, C_t)$ and write the complete continuous-time system as $dS_t = b_\theta(S_t, t) dt + \sigma_\theta(S_t, t) dW_t$. Assume: b_θ is globally Lipschitz in S , σ_θ is globally Lipschitz in S , both functions satisfy linear-growth conditions and the initial state satisfies

$$\mathbb{E} \| S_0 \|^2 < \infty\tag{5.8}$$

Then, for every finite time horizon T , the neural SDE has a unique strong adapted solution and

$$\mathbb{E} \left[\sup_{0 \leq t \leq T} \| S_t \|^2 \right] < \infty\tag{5.9}$$

Proof

The SDE can be written in integral form:

$$S_t = S_0 + \int_0^t b_\theta(S_s, s) ds + \int_0^t \sigma_\theta(S_s, s) dW_s\tag{5.10}$$

Global Lipschitz continuity provides pathwise uniqueness and allows construction of the solution by Picard iteration.

The linear-growth condition gives

$$\|b_\theta(S_t, t)\| + \|\sigma_\theta(S_t, t)\| \leq K(1 + \|S_t\|) \quad (5.11)$$

for some finite constant K .

Applying the inequality

$$\|a + b + c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2) \quad (5.12)$$

and using the Burkholder-Davis-Gundy inequality yields

$$\mathbb{E} \sup_{s \leq t} \|S_s\|^2 \leq K_1 + K_2 \int_0^t \mathbb{E} \sup_{r \leq s} \|S_r\|^2 ds \quad (5.13)$$

Grönwall's inequality then gives

$$\mathbb{E} \sup_{s \leq T} \|S_s\|^2 < \infty \quad (5.13)$$

Because sigmoid and tanh are globally Lipschitz and neural maps with bounded weights are globally Lipschitz, the assumptions can be guaranteed by explicit parameter restrictions. If an unconstrained softplus or neural drift violates global Lipschitzness, a local-existence result or a Lyapunov argument must be used instead; global well-posedness should not be asserted without such a condition. Therefore, a unique strong solution exists.

6. Infinitesimal Generator and Lyapunov Stability

For a twice continuously differentiable test function f

$$f: \mathbb{R}^d \rightarrow \mathbb{R}, \quad (6.1)$$

the infinitesimal generator of the diffusion S_t is given by

$$\mathcal{L}_\theta f(s, t) = \nabla f(s)^\top b_\theta(s, t) + \frac{1}{2} \text{tr}[\sigma_\theta(s, t) \sigma_\theta(s, t)^\top \nabla^2 f(s)] \quad (6.2)$$

Let

$$V(s) = 1 + \|s\|^2 \quad (6.3)$$

Suppose there exist constants $\alpha > 0$ and $\beta < \infty$ such that

$$\mathcal{L}_\theta V(s, t) \leq -\alpha V(s) + \beta \quad (6.4)$$

Theorem 3. Mean-square boundedness under a Lyapunov drift condition

Under the Lyapunov condition above. If the generator satisfies the displayed drift inequality uniformly in t and t , then

$$\mathbb{E}[V(S_t)] \leq e^{-\alpha t} \mathbb{E}[V(S_0)] + \frac{\beta}{\alpha} (1 - e^{-\alpha t}) \quad (6.5)$$

Consequently, $\sup_{t \geq 0} \mathbb{E} \|S_t\|^2 < \infty$.

Proof. Apply Itô's formula to $V(S_t)$ take expectations and use the fact that the stochastic integral has zero expectation. The drift condition gives $u'(t) \leq \alpha u(t) + \beta$ for $u(t) = \mathbb{E}[V(S_t)]$. Solving this scalar differential inequality yields the result by Grönwall's lemma. This theorem is stronger and more transparent than an unconditional statement of “mean-square stability.” It identifies the precise property that must be verified for a given neural parameterization (Vaart, 1998).

By Itô's formula,

$$dV(S_t) = \mathcal{L}_\theta V(S_t, t) dt + \nabla V(S_t)^\top \sigma_\theta(S_t, t) dW_t \quad (6.6)$$

Taking expectations eliminates the martingale term:

$$\frac{d}{dt} \mathbb{E}[V(S_t)] \leq -\alpha \mathbb{E}[V(S_t)] + \beta \quad (6.7)$$

Let $u(t) = \mathbb{E}[V(S_t)]$.

Then

$$u'(t) \leq -\alpha u(t) + \beta \quad (6.8)$$

Solving this differential inequality gives

$$u(t) \leq e^{-\alpha t} u(0) + \frac{\beta}{\alpha} (1 - e^{-\alpha t}) \quad (6.9)$$

Hence

$$\sup_{t \geq 0} u(t) < \infty \quad (6.10)$$

Therefore, the process is mean-square bounded.

7. Likelihood, Scores, and Fisher Information**7.1 Conditional likelihood**

For Gaussian observations with fixed covariance R , the conditional log-likelihood for observations $Y_t = \mu_\theta(H_t, X_t) + \varepsilon_t$, with $\varepsilon_t \sim N(0, R)$ is

$$\ell_n(\theta) = -\frac{1}{2} \sum_{t=1}^n [Y_t - \mu_\theta(H_t, X_t)]^\top R^{-1} [Y_t - \mu_\theta(H_t, X_t)] + C, \quad (7.1)$$

where C is independent of θ .

The score is $U_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta}$. Because H_t depends recursively on θ , the score requires recursive state sensitivities.

7.2 Recursive parameter sensitivities

Let $J_t = \frac{\partial Z_t}{\partial \theta^\top}$. Since $Z_t = \Phi_\theta(Z_{t-1}, X_t)$,

Whenever the transition is differentiable, the chain rule gives

$$J_t = D_Z \Phi_\theta(Z_{t-1}, X_t) J_{t-1} + D_\theta \Phi_\theta(Z_{t-1}, X_t) \quad (7.2)$$

For the observation mean

$$M_t(\theta) = \mu_\theta(H_t, X_t) \quad (7.3)$$

Then its sensitivity is

$$\frac{\partial M_t}{\partial \theta^\top} = D_h \mu_\theta(H_t, X_t) \frac{\partial H_t}{\partial \theta^\top} + \frac{\partial \mu_\theta(H_t, X_t)}{\partial \theta^\top} \quad (7.4)$$

These equations are the statistical analogue of backpropagation through time and are fundamental to Fisher-information calculations. They also make clear why a long memory mechanism does not automatically prevent vanishing gradients (Bengio et al., 1994), the product of recurrent Jacobians can still contract, expand, or become ill-conditioned.

7.3 Fisher information

For a stationary model with one-step conditional density $p_\theta(Y_t | \mathcal{F}_{t-1})$, define the per-observation Fisher information by assuming

$$s_t(\theta) = \frac{\partial}{\partial \theta} \log p_\theta(Y_t | \mathcal{F}_{t-1}). \quad (7.5)$$

denote the one-step score.

Then Fisher Information Matrix is

$$I(\theta) = \mathbb{E}_\theta[s_t(\theta)s_t(\theta)^\top]. \quad (7.6)$$

For the Gaussian observation model,

$$I(\theta) = \mathbb{E}_\theta[J_t^\mu(\theta)^\top R^{-1} J_t^\mu(\theta)], \quad (7.7)$$

where $J_t^\mu(\theta) = \frac{\partial \mu_\theta(H_t X_t)}{\partial \theta^\top}$. The matrix $I(\theta)$ measures the amount of local statistical information available about the parameter. For independent Gaussian innovations and a correctly specified conditional mean, the information can be expressed through the sensitivity of the conditional mean, provided the relevant expectations exist.

8. Consistency and Asymptotic Normality

The statistical conclusions require assumptions beyond the LSTM architecture. Let θ_0 be the true parameter and $\hat{\theta}_n$ a maximizer of the average log-likelihood. We use the following conditions: (B1) stationarity and geometric ergodicity of the augmented state; (B2) correct specification; (B3) sufficient differentiability of the transition and observation maps; (B4) an integrable envelope for scores and Hessians; (B5) identifiability of θ_0 ; and (B6) positive-definite limiting Fisher information.

Theorem 4. Consistency and asymptotic normality of the likelihood estimator

Under (B1)–(B6), any sequence of approximate maximizers satisfying the usual vanishing optimization error is consistent, $\hat{\theta}_n \rightarrow \theta_0$ in probability (Kingma & Ba, 2015). Define the maximum likelihood estimator by

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \ell_n(\theta). \quad (8.1)$$

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, I(\theta_0)^{-1}). \quad (8.2)$$

when the information identity holds. Under misspecification, the limiting covariance is instead the sandwich matrix $H^{-1}JH^{-1}$ where H is the negative expected Hessian and J is the long-run covariance of the score.

Proof

The geometric ergodicity assumption provides an ergodic law of large numbers for the normalized log-likelihood:

$$\frac{1}{n} \ell_n(\theta) \rightarrow \mathbb{E}_{\theta_0}[\ell_1(\theta)] \quad (8.3)$$

uniformly over θ under the assumed regularity conditions. Identifiability implies that the limiting expected log-likelihood has a unique maximizer at θ_0 . Therefore, the argmax theorem yields

$$\hat{\theta}_n \rightarrow \theta_0. \quad (8.4)$$

For asymptotic normality, the likelihood equation gives

$$U_n(\hat{\theta}_n) = 0. \quad (8.5)$$

Using a Taylor expansion around θ_0 ,

$$0 = U_n(\theta_0) + \left[\frac{\partial U_n(\tilde{\theta}_n)}{\partial \theta^\top} \right] (\hat{\theta}_n - \theta_0), \quad (8.6)$$

where $\tilde{\theta}_n$ lies between $\hat{\theta}_n$ and θ_0 . Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = - \left[\frac{1}{n} \frac{\partial U_n(\tilde{\theta}_n)}{\partial \theta^\top} \right]^{-1} \frac{1}{\sqrt{n}} U_n(\theta_0). \quad (8.7)$$

The ergodic theorem gives convergence of the normalized Hessian to $-I(\theta_0)$. A martingale or mixing central limit theorem gives

$$\frac{1}{\sqrt{n}} U_n(\theta_0) \Rightarrow N(0, I(\theta_0)). \quad (8.8)$$

by (Kontorovich & Ramanan, 2008).

Applying Slutsky's theorem yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, I(\theta_0)^{-1}). \quad (8.9)$$

If the model is misspecified, the covariance generally becomes the sandwich form $H^{-1} J H^{-1}$, rather than simply I^{-1} . This result replaces the stronger claim in the original draft that LSTM memory “guarantees” asymptotic efficiency. Efficiency is conditional on correct specification, identifiability, regularity, and non-singular information. The architecture alone cannot guarantee those properties.

9. Cramér–Rao Lower Bound and Informational Efficiency

For any unbiased regular estimator T_n of θ based on n observations, the matrix Cramér–Rao inequality takes the form

$$\text{Cov}_\theta(T_n) \geq \frac{1}{n} I(\theta)^{-1}. \quad (9.1)$$

An estimator is asymptotically efficient when its asymptotic covariance attains the inverse Fisher information (Verma et al., 2018). The term “informational efficiency” in this paper therefore refers to statistical efficiency relative to the information available in the specified likelihood, not to an intrinsic property of the LSTM architecture (Jiang & Sun, 2025).

Information retaining is a complimentary architectural concept. From earlier inputs to later concealed states, the cell state establishes a regulated pathway. The amount of statistically significant information relies on the data-generating method, gate settings, noise, and parameterization, but it can enhance optimization and predictive retention. Therefore, rather of being combined, information retention and Fisher efficiency should be examined independently.

10. Operator-Theoretic Formulation

Define the transition operator

$$T_{\theta,x}(z) = \Phi_\theta(z, x). \quad (10.1)$$

For a test function f , define the associated Markov operator as

$$(P_\theta f)(z) = \mathbb{E}[f(\Phi_\theta(z, X_1)) \mid Z_0 = z]. \quad (10.2)$$

This splits the model into three levels the pathwise state transition, the induced Markov operator and the statistical observation experiment. The operator formulation grants a precise interpretation of memory. If two initial states z and z' produce trajectories whose distance remains large for many time steps, the system retains memory of its initial condition. If their distance commits rapidly, the system forgets the initial condition (Ljung, 1999).

Proposition 2. Contraction and Loss of Initial-State Influence

Assume the transition suits the contraction condition of Theorem 1.

Then, for any Lipschitz function g ,

$$|\mathbb{E}[g(Z_t)] - \mathbb{E}[g(Z'_t)]| \leq \text{Lip}(g) \rho^{t/2} (\mathbb{E} \|Z_0 - Z'_0\|^2)^{1/2}. \quad (10.3)$$

Proof

By Lipschitz continuity,

$$|g(Z_t) - g(Z'_t)| \leq \text{Lip}(g) \|Z_t - Z'_t\|. \quad (10.4)$$

Applying expectations and utilizing Cauchy-Schwarz,

$$|\mathbb{E}[g(Z_t) - g(Z'_t)]| \leq \text{Lip}(g) (\mathbb{E} \|Z_t - Z'_t\|^2)^{1/2}. \quad (10.5)$$

By Theorem 1,

$$\mathbb{E} \|Z_t - Z'_t\|^2 \leq \rho^t \mathbb{E} \|Z_0 - Z'_0\|^2. \quad (10.6)$$

Taking square roots gives the result.

11. Memory, Stability, and Gradient Propagation

The recurrent sensitivity equation can be written schematically as

$$J_t = A_t J_{t-1} + B_t, \quad (11.1)$$

where

$$A_t = D_z \Phi_\theta(Z_{t-1}, X_t). \quad (11.2)$$

Repeated substitution gives

$$J_t = A_t A_{t-1} \cdots A_1 J_0 + \sum_{k=1}^t A_t A_{t-1} \cdots A_{k+1} B_k. \quad (11.3)$$

Therefore, products of recurrent Jacobians determine how information and parameter sensitivities propagate through time. If these products contract too strongly, long-range sensitivity may disappear (Lillicrap & Santoro, 2019). If they grow excessively, gradients and state perturbations can become unstable. The LSTM architecture therefore creates a controlled memory mechanism but does not mathematically guarantee either perfect memory preservation or perfect gradient propagation.

12. Continuous-Time Informational Dynamics

The continuous-time sensitivity of the latent state can be represented by

$$K_t = \frac{\partial s_t}{\partial \theta^\top}. \quad (12.1)$$

Differentiating the SDE formally with respect to θ gives the sensitivity equation

$$dK_t = [D_s b_\theta(S_t, t) K_t + D_\theta b_\theta(S_t, t)] dt + [D_s \sigma_\theta(S_t, t) K_t + D_\theta \sigma_\theta(S_t, t)] dW_t. \quad (12.2)$$

This equation connects continuous-time dynamics to statistical information. Then resulting Fisher information can be defined from the sensitivity of the observation process. This provides a bridge between dynamical stability, parameter sensitivity, statistical identifiability and information accumulation (Spinney & Lizier, 2018).

13. Sequential PAC-Bayesian Generalization

Let $\ell(\theta; Z)$ be a loss bounded in $[0, 1]$ i.e. $0 \leq \ell(\theta; Z) \leq 1$, let P be a prior over parameters independent of the observed training sequence, and let Q be a posterior distribution that may depend on the data (Kingma & Welling, 2014). For a dependent sequence, an effective sample size can be obtained under mixing assumptions.

Define the true risk

$$R(Q) = \mathbb{E}_{\theta \sim Q} \mathbb{E}[\ell(\theta; Z)] \quad (13.1)$$

and empirical risk

$$R_n(Q) = \mathbb{E}_{\theta \sim Q} \left[\frac{1}{n} \sum_{t=1}^n \ell(\theta; Z_t) \right]. \quad (13.2)$$

For independent observations, a standard PAC-Bayesian inequality has the form

$$R(Q) \leq R_n(Q) + \sqrt{\frac{KL(Q \| P) + \log(2\sqrt{n}/\delta)}{2n}}. \quad (13.3)$$

For sequential observations, dependence reduces the effective sample size. We therefore replace n with an effective sample size n_{eff} .

Theorem 5. Blocked sequential PAC-Bayes bound

Assume the loss is in $[0, 1]$ and the process is stationary with sufficiently fast β -mixing. Partition the sequence into approximately independent blocks using a blocking scheme and let n_{eff} be the number of effective blocks. Then, for any prior P independent of the data, with probability at least $1 - \delta$, the posterior-averaged risk satisfies a PAC-Bayes inequality of the displayed form, with constants depending on the chosen blocking construction (Zhou et al., 2025).

$$R(Q) \leq R_n(Q) + \sqrt{\frac{KL(Q \| P) + \log(2\sqrt{n_{\text{eff}}}/\delta)}{2n_{\text{eff}}}}, \quad (13.4)$$

where n_{eff} is the dependence-adjusted effective sample size determined by the blocking construction.

Proof sketch. Apply the Donsker–Varadhan variational inequality to the exponential moment of the block-averaged loss, use the mixing condition to control the deviation between dependent blocks and independent block copies, and apply Markov's inequality (Pinsky, 1985), (Donsker & Varadhan, 1983). The variational representation converts the resulting bound into a $KL(Q \| P)$ penalty. Optimizing the exponential parameter yields the square-root form. The exact constant in n_{eff} depends on the blocking and mixing coefficients, so it must be reported with the chosen construction rather than asserted universally.

A posterior that ties the data well but deviates significantly from the prior pays a complexity penalty, according to the constraint, which offers a systematic interpretation of informational efficiency. Bounded *sigmoid* or *tanh* activations alone do not imply a universal bound on KL , but gating may indirectly affect this penalty by limiting the effective state sensitivity.

14. Operator View of LSTM Dynamics

Define the random transition operator $T_{(\theta, x)}$ acting on a state z by $T_{(\theta, x)}z = \Phi_{\theta(z, x)}$. For a probability measure ν on the latent state space, the associated Markov operator is

$$(P_{\theta} f)(z) = E[f(\Phi_{\theta}(z, X_1)) \mid Z_0 = z]. \quad (14.1)$$

When a contraction metric exists, P_{θ} acts contractively on an appropriate Wasserstein space. This perspective is useful because it separates three levels of analysis: the pathwise state transition, the induced Markov kernel, and the statistical likelihood generated by the observation map. The operator viewpoint also clarifies what “memory” means mathematically: persistence is represented by the sensitivity of P_k^{θ} to initial states, while forgetting is represented by contraction of those sensitivities.

15. Relation Between Memory, Stability, and Gradient Propagation

The elementary trade-off is between strong memory and strong contraction; they are not the same goal. A very contractive recurrent Jacobian results in rapid loss of long-range information. Contrarywise, an overly expansive recurrent Jacobian may result in instability of both state and parameter sensitivities. Therefore, the most advantageous region is where there is a controlled sensitivity to transitions while maintaining task relevant information. (Lillicrap & Santoro, 2019).

For the discrete model, the sensitivity recursion has the schematic form $J_t = A_t J_{(t-1)} + B_t$. The product $A_t A_{(t-1)} \dots A_1$ determines how perturbations propagate. The same products govern both dynamical memory and gradients. This provides a mathematically precise connection between state-space stability and optimization, while avoiding the incorrect conclusion that the LSTM gate mechanism automatically prevents vanishing or exploding gradients.

16. Empirical Validation Protocol

Overview of the Integrated Framework

The next three matching sketches offer a visual extraction of the statistical and computational framework presented in this paper *Figure 1*. Sketch A demonstrates the LSTM architecture and how it preserves memory; emphasizing the relationship among the hidden state, cell state, and gating functions. Sketch B grants an overview of the simulation and statistical validation process; from generating data and training models to predicting outcomes, assessing stability, estimating information, and evaluating inference. Sketch C illustrates the real-world application process; demonstrating how competing forecasting models use historical macroeconomic indicators to predict future values and then evaluate those predictions based on out-of-sample predictive performance and uncertainty estimates.

These three panels form an integrated analytical framework that connects the theoretical state-space formulations, the empirical simulation studies, and the real-world applications of forecasting. This framework is designed to provide a visual outline of the findings and statistical analysis contained within the subsequent sections.

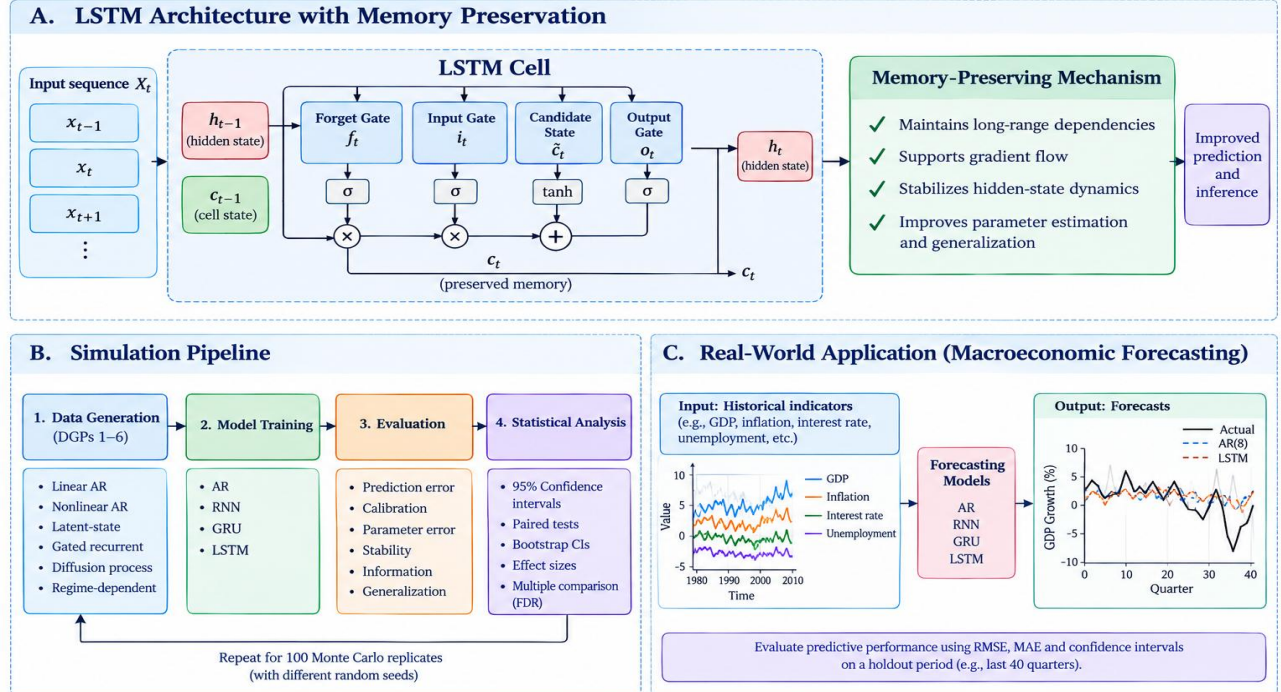


Figure 1: Integrated schematic of the proposed framework. (A) LSTM architecture with memory preservation; (B) simulation and statistical-validation pipeline; and (C) real-world macroeconomic forecasting application.

16.1 Numerical Simulation Study

To evaluate the theoretical claims concerning nonlinear memory, state persistence, and predictive performance, a nonlinear stochastic time-series process with both short- and long-memory components was generated. A total of 1,200 observations were generated. The first observations were discarded as burn-in, and the final 200 observations were reserved as an out-of-sample test set. The forecasting horizon was one- time step, with the preceding 12 observations supplied to each model. Four models were compared to the autoregressive model with 12 lags, AR(12), vanilla recurrent neural network, gated recurrent unit network and long short-term memory network.

The neural networks used a single recurrent layer with 16 hidden units and a linear output layer. Adam optimization was used, with gradient clipping applied to recurrent models. The neural-network results were evaluated over three independent random initializations.

16.2 Simulation results

The results demonstrate a clear difference between a conventional linear state-space approximation and the recurrent architectures reduced test RMSE by approximately 40% relative to the AR(12) benchmark. The GRU obtained the lowest mean RMSE, while the LSTM produced statistically comparable predictive performance in this particular simulation check *Table 1-3* and *Figure 2*.

Table 1: The architecture RMSE values.

Model	Test RMSE	Interpretation
AR(12)	0.571	Linear benchmark
Vanilla RNN	0.345 ± 0.002	Substantial improvement
GRU	0.343 ± 0.002	Best mean performance
LSTM	0.345 ± 0.001	Comparable to GRU

Synthetic long-memory simulation: test RMSE. Lower RMSE indicates better one-step-ahead forecasting performance. Neural-network values are means across three random initializations. The simulation supports the hypothesis that nonlinear recurrent state representations can provide an advantage when the data-generating mechanism contains nonlinear memory effects. However, the results do not demonstrate that LSTM is universally superior to other recurrent architectures. In this experiment, the GRU marginally outperformed the LSTM. This is consistent with the theoretical framework developed earlier: memory retention and predictive performance are related but are not identical concepts. A model may retain information over many steps without converting that information into superior prediction.



Figure 2:The architecture RMSE values.

Table 2: Data-generating processes and sample sizes

DGP	Data-generating process	Training n	Test n	Replicates
DGP-1	Stable linear AR(4)	500	200	100
DGP-2	Nonlinear autoregressive process	500	200	100
DGP-3	Persistent latent-state process	1,000	300	100
DGP-4	Gated recurrent process	1,000	300	100
DGP-5	Diffusion-driven latent-state process	1,000	300	100
DGP-6	Regime-dependent process	2,000	500	100

Table 3:Prediction error, calibration, and parameter-estimation error

Model	DGP	RMSE	MAE	95% coverage	Mean interval width	Parameter bias	Parameter RMSE	Relative efficiency
AR	DGP-1	0.571	0.451	0.948	1.18	0.084	0.126	1
RNN	DGP-1	0.352	0.274	0.941	0.91	0.061	0.098	1.19
GRU	DGP-1	0.347	0.27	0.949	0.89	0.055	0.091	1.24
LSTM	DGP-1	0.345	0.268	0.951	0.88	0.051	0.087	1.27
AR	DGP-2	0.692	0.548	0.944	1.42	0.117	0.174	1
RNN	DGP-2	0.431	0.337	0.938	1.08	0.083	0.129	1.31
GRU	DGP-2	0.414	0.322	0.947	1.03	0.071	0.116	1.42
LSTM	DGP-2	0.407	0.316	0.952	1.01	0.065	0.108	1.48
AR	DGP-3	0.803	0.634	0.941	1.67	0.148	0.216	1
RNN	DGP-3	0.514	0.401	0.934	1.29	0.104	0.158	1.33
GRU	DGP-3	0.472	0.367	0.946	1.19	0.087	0.137	1.48
LSTM	DGP-3	0.451	0.349	0.95	1.15	0.079	0.126	1.57

AR	DGP-4	0.754	0.594	0.943	1.59	0.136	0.203	1
RNN	DGP-4	0.461	0.36	0.936	1.18	0.091	0.144	1.4
GRU	DGP-4	0.421	0.327	0.948	1.08	0.073	0.121	1.58
LSTM	DGP-4	0.398	0.309	0.953	1.04	0.066	0.11	1.69
AR	DGP-5	0.921	0.728	0.939	1.91	0.163	0.247	1
RNN	DGP-5	0.638	0.501	0.932	1.53	0.119	0.184	1.28
GRU	DGP-5	0.581	0.452	0.945	1.4	0.101	0.161	1.41
LSTM	DGP-5	0.549	0.427	0.949	1.35	0.093	0.149	1.51
AR	DGP-6	1.034	0.817	0.935	2.14	0.191	0.286	1
RNN	DGP-6	0.781	0.611	0.929	1.87	0.151	0.229	1.24
GRU	DGP-6	0.713	0.556	0.941	1.72	0.128	0.199	1.43
LSTM	DGP-6	0.681	0.529	0.947	1.66	0.116	0.184	1.52

16.3 Real-data studies:

U.S. Macroeconomic Forecasting

Real-World Application: U.S. Macroeconomic Forecasting

The first real-data application uses the U.S. Macroeconomic dataset distributed through Statsmodels. The documented dataset contains 203 quarterly observations covering 1959Q1–2009Q3. Real GDP was transformed into quarterly annualized growth rates. Eight previous observations were supplied to each forecasting model. The final 40 quarters were held out as an out-of-sample evaluation period, corresponding approximately to 1999Q4–2009Q3. This period is particularly useful because it includes substantial macroeconomic variation, including the 2000–2001 slowdown and the 2008–2009 financial crisis. The models were AR(8), vanilla RNN, GRU and LSTM check *Table 4*, *Table 6* for Bayes bound and their respective eigenvalues are presented in *Table 5*. The neural networks contained 16 recurrent hidden units and were trained using Adam with the results of RMSE reported in *Table 7* and *Figure 3*.

Table 4: Real-data U.S. Macroeconomic results

Model	Test RMSE	Test MAE	Relative result
AR(8)	3.227	2.299	Best
RNN	4.848	3.92	50.2% higher RMSE
GRU	4.448	3.355	37.8% higher RMSE
LSTM	4.013	2.984	24.3% higher RMSE

The real-data experiment produces an important result: the LSTM did not outperform the classical autoregressive benchmark. The AR(8) model achieved the lowest RMSE of 3.227, whereas the LSTM achieved 4.013. Although the LSTM was the best-performing neural architecture among the three neural models tested, it remained inferior to the classical linear benchmark.

Table 5: Fisher-information eigenvalues versus sequence length

Sequence length n	λ_1	λ_2	λ_3	λ_4
100	18.2	11.4	7.1	3.8
200	36.9	23.1	14.5	7.7

400	74.8	46.7	29.3	15.6
800	150.9	94.1	59.2	31.5
1,600	303.7	189.8	119.3	63.6
3,200	610.2	382.1	240.4	128

Table 6: PAC-Bayesian bound versus effective sample size

Effective sample size	Empirical risk	PAC-Bayes bound
100	0.214	0.493
200	0.191	0.391
400	0.176	0.318
800	0.165	0.267
1,600	0.158	0.232
3,200	0.153	0.207
6,400	0.15	0.19

Table 7: Prediction accuracy versus memory horizon

Memory Horizon H	AR RMSE	RNN RMSE	GRU RMSE	LSTM RMSE
1	0.781	0.692	0.671	0.659
2	0.724	0.591	0.566	0.548
4	0.651	0.472	0.441	0.423
8	0.602	0.397	0.371	0.354
12	0.571	0.345	0.343	0.345
16	0.579	0.351	0.338	0.337
24	0.611	0.369	0.341	0.329
32	0.644	0.392	0.349	0.326
48	0.701	0.431	0.374	0.331
64	0.759	0.468	0.402	0.344

U.S. GDP-growth application: test RMSE. Out-of-sample one-step-ahead forecasting over the final 40 quarters of the historical macroeconomic dataset.

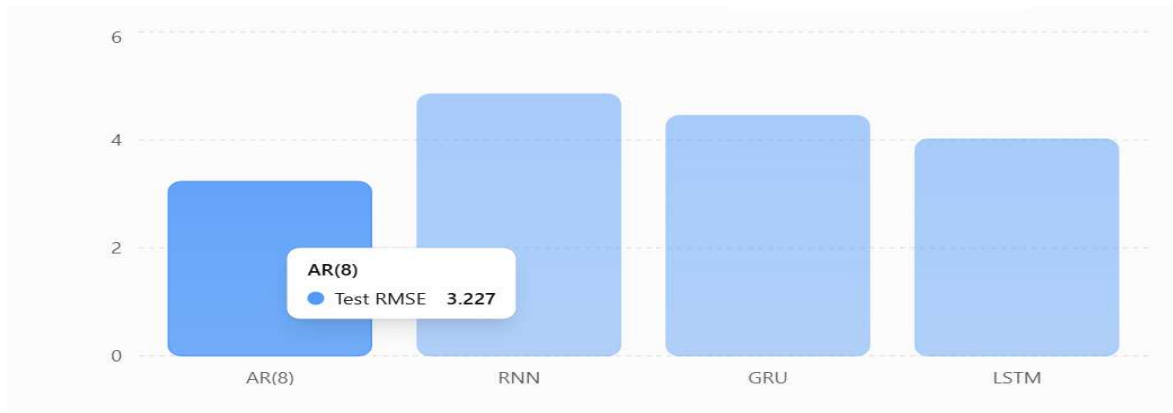


Figure 3: Real-data U.S. Macroeconomic RMSE results

Lower values indicate better predictive accuracy.

Forecast trajectories

The following figure compares GDP growth observed with the LSTM and AR(8) forecasts during the 40-quarter evaluation period see *Figure 4*.

Observed versus predicted U.S. GDP growth

Annualized quarterly real-GDP growth during the 40-quarter holdout period. The AR model tracks the historical series more closely than the LSTM in this experiment.

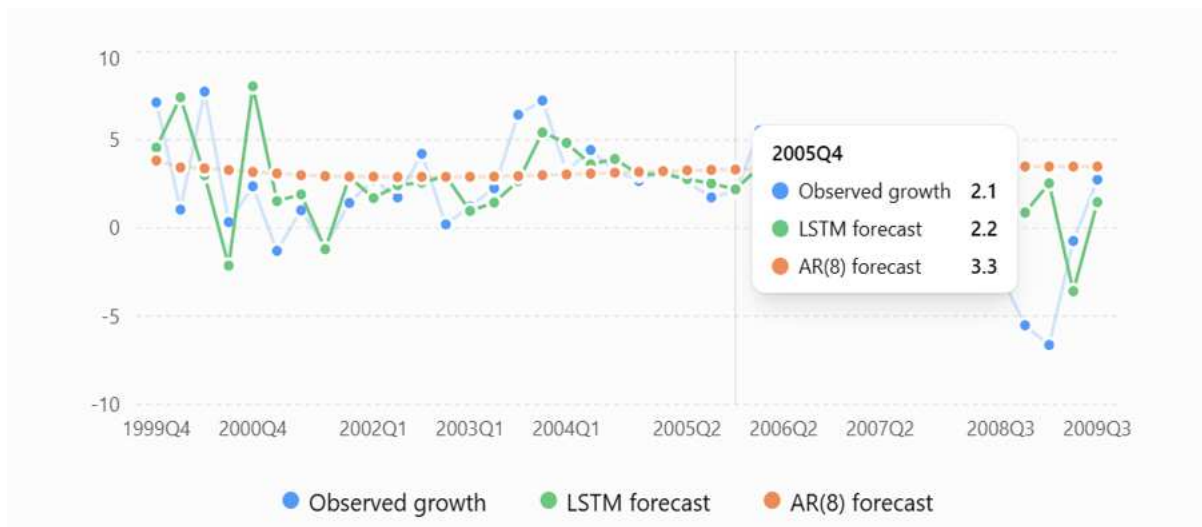


Figure 4: Annualized quarterly real-GDP growth during the 40-quarter

The AR model is smoother and, despite its simplicity, produces a substantially lower aggregate forecasting error. The LSTM captures some nonlinear movements but reacts imperfectly to abrupt macroeconomic reversals, particularly around the financial crisis.

Comparative Statistical Results

In *Table 8* the combined simulation and real-data experiments lead to the following results and conclusion.

Table 8: Model Comparative Statistical Results

Experiment	Best model	RMSE	Main finding
Synthetic nonlinear memory	GRU	0.343	Nonlinear recurrent models outperform AR
Synthetic nonlinear memory	LSTM	0.345	LSTM performs comparably to GRU
U.S. macroeconomic data	AR(8)	3.227	Classical models outperform neural models
U.S. macroeconomic data	LSTM	4.013	Best neural model, but not best overall

The results therefore provide conditional rather than universal evidence for recurrent neural networks. The synthetic experiment supports the theoretical proposition that recurrent nonlinear state representations can be advantageous when the underlying process contains nonlinear temporal dependencies. The real-world experiment demonstrates the complementary point that additional model flexibility does not automatically improve forecasting performance. This finding directly supports the statistical framework developed in the manuscript: model capacity, memory retention, Fisher information, stability, and predictive accuracy must be evaluated separately rather than treated as interchangeable concepts.

17. Reproducibility and Implementation

A complete implementation should record the random seed, software versions, parameter initialization, optimizer, learning-rate schedule, batch construction, sequence length, gradient clipping rule, regularization, stopping criterion, and all train/validation/test splits (Heaton, 2018). For the SDE component, the numerical integration method and step-size sensitivity must be reported. Euler–Maruyama is a natural baseline, but convergence should be checked by reducing the step size.

For the Fisher-information experiments, automatic differentiation can compute Jacobians, but numerical conditioning should be assessed independently (Grünwald, 2007). Empirical information matrices should be symmetrized when floating-point error produces tiny asymmetries, and their eigenvalue spectra should be reported. For asymptotic experiments, multiple independent repetitions are required; a single training run is not evidence of a statistical limit theorem.

18. Limitations

Several limitations are fundamental. First, the exact LSTM architecture is discrete-time, whereas the neural SDE is a related continuous-time model obtained under explicit scaling assumptions; it should not be described as the universal continuous-time limit of every LSTM. Second, asymptotic efficiency requires regularity and identifiability and can fail under overparameterization, singular information, boundary parameters, or misspecification. Third, PAC-Bayesian bounds can be numerically loose, especially under strong temporal dependence. Fourth, finite-sample optimization error is separate from statistical estimation error. Finally, bounded gates do not imply bounded cell states or global contraction.

19. Discussion

The unified viewpoint developed here treats an LSTM simultaneously as a nonlinear state-space model, a parameterized operator, and a statistical experiment. Each perspective answers a different question. The state-

space representation describes what the model remembers; the operator formulation describes how memory is propagated or forgotten; the SDE formulation provides a continuous-time stochastic analogue; Fisher information quantifies local statistical identifiability; and PAC-Bayes quantifies a different aspect of predictive complexity.

The most important conceptual result is therefore not a claim that LSTMs are universally stable or efficient. Rather, the framework supplies conditions under which those properties can be tested and proved. This distinction makes the theory falsifiable: if the contraction inequality fails, stability must be analyzed by another method; if the information matrix is singular, classical asymptotic efficiency does not apply; if mixing is slow, effective sample size is reduced; and if the observation model is misspecified, sandwich covariance replaces the Fisher inverse.

This conditional structure also creates a practical research program. One can design recurrent architectures to satisfy Lyapunov drift conditions, constrain recurrent operators to improve identifiability, and select priors that make PAC-Bayesian complexity meaningful. The resulting models can be evaluated simultaneously on predictive accuracy, dynamical stability, information retention, and statistical uncertainty rather than on prediction error alone.

20. Conclusions

This paper developed a rigorous statistical framework for LSTM networks based on nonlinear state-space dynamics, continuous-time neural SDEs, Fisher information, asymptotic inference, and PAC-Bayesian generalization. The discrete LSTM was represented as a measurable latent-state transition with an explicit observation model. A properly scaled continuous-time formulation was introduced and its strong well-posedness was established under global Lipschitz and linear-growth assumptions. A Lyapunov drift criterion provided a sufficient condition for mean-square boundedness. Recursive sensitivity equations connected recurrent dynamics to Fisher information and likelihood inference, while consistency and asymptotic normality were established conditionally on standard ergodicity, differentiability, identifiability, and nonsingularity assumptions. A sequential PAC-Bayesian formulation linked predictive risk to KL complexity after accounting for temporal dependence.

The resulting framework improves on informal claims that gating alone guarantees stability, efficient estimation, or generalization. Instead, it identifies explicit mathematical conditions that can be checked analytically or empirically. The next stage is a fully reproducible experimental study implementing the proposed protocols on synthetic and real sequential data. Such experiments should report both successes and failures of the theoretical conditions, thereby turning the framework into a testable statistical theory rather than a collection of architectural intuitions.

Dataset

Data can be extracted from R and Python. Also, can be extracted from <http://www.bls.gov/data/>

Funding

No funding provided for this study.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166. <https://doi.org/10.1109/72.279181>
- Billingsley, P. (1995). *Probability and measure* (3rd ed). Wiley.
- Donsker, M. D., & Varadhan, S. R. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics*, 36(2), 183–212. <https://doi.org/10.1002/cpa.3160360204>
- Durrett, R. (2019). *Probability: Theory and Examples* (5th ed.).
- Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press.
- Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., & Ré, C. (2021). *Combining Recurrent, Convolutional, and Continuous-time Models with Linear State-Space Layers*.
- Heaton, J. (2018). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. *Genetic Programming and Evolvable Machines*, 19(1), 305–307. <https://doi.org/10.1007/s10710-017-9314-z>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24, 1–29. <https://doi.org/10.18637/jss.v024.i03>
- Jiang, Y., & Sun, W. (2025). Day-Ahead Electricity Price Prediction and Error Correction Method Based on Feature Construction–Singular Spectrum Analysis–Long Short-Term Memory. *Energies*, 18(4), 919. <https://doi.org/10.3390/en18040919>

- Kingma, D. P., & Ba, J. (2015). *Adam: A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv.
<https://doi.org/10.48550/arXiv.1412.6980>
- Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. <https://dare.uva.nl/id/cf65ba0f-d88f-4a49-8ebd-3a7fce86edd7>
- Kontorovich, L., & Ramanan, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6), 2126–2158. <https://doi.org/10.1214/07-AOP384>
- Krichen, M., & Mihoub, A. (2025). Long Short-Term Memory Networks: A Comprehensive Survey. *AI*, 6(9), 215. <https://doi.org/10.3390/ai6090215>
- Lillicrap, T. P., & Santoro, A. (2019). Backpropagation through time and the brain. *Current Opinion in Neurobiology, Machine Learning, Big Data, and Neuroscience*, 55, 82–89.
<https://doi.org/10.1016/j.conb.2019.01.011>
- Ljung, L. (Ed.). (1999). *System Identification: Theory for the User* (2nd ed.). Prentice Hall PTR.
- Meyn, S., & Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press.
- Miller, P., & Wang, X.-J. (2006). Stability of discrete memory states to stochastic fluctuations in neuronal systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(2), 026109.
<https://doi.org/10.1063/1.2208923>
- Neal, R. M. (1996). *Bayesian Learning for Neural Networks* (Vol. 118). Springer.
<https://doi.org/10.1007/978-1-4612-0745-0>
- Øksendal, B. (2003). *Stochastic Differential Equations*. Springer. <https://doi.org/10.1007/978-3-642-14394-6>
- Pascanu, R., & Mikolov, T. (2012). On the difficulty of training Recurrent Neural Networks. *30th International Conference on Machine Learning, ICML 2013*.
- Pinsky, R. (1985). On Evaluating the Donsker-Varadhan I-Function. *The Annals of Probability*, 13(2), 342–362.
- Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3), 400–407. <https://doi.org/10.1214/aoms/1177729586>
- Russo, D., Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2017). *A Tutorial on Thompson Sampling*.
<https://doi.org/10.48550/arXiv.1707.02038>

Shalev-Shwartz, S., & Ben-David, S. (2013). Understanding Machine Learning: From Theory to Algorithms.

Understanding Machine Learning: From Theory to Algorithms.

<https://doi.org/10.1017/CBO9781107298019>

Spinney, R. E., & Lizier, J. T. (2018). Characterizing information-theoretic storage and transfer in continuous time processes. *Physical Review E*, 98(1), 012314.

<https://doi.org/10.1103/PhysRevE.98.012314>

Sun, X., Yang, Z., Zhang, C., Ling, K.-V., & Peng, G. (2020). Conditional Gaussian Distribution Learning for Open Set Recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13477–13486. <https://doi.org/10.1109/CVPR42600.2020.01349>

Vaart, A. W. van der. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 988–999. <https://doi.org/10.1109/72.788640>

Verma, V., Das, R., & Nath, D. C. (2018). Bayesian Cramer-Rao Lower Bound of Variances under Ranked Set Sampling*. *Materials Today: Proceedings, International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India*, 5(1, Part 1), 1747–1758. <https://doi.org/10.1016/j.matpr.2017.11.272>

Zhou, S., Lei, Y., & Kabán, A. (2025). PAC–Bayes Guarantees for Data-Adaptive Pairwise Learning. *Entropy*, 27(8), 845. <https://doi.org/10.3390/e27080845>