Why Do Transformers Fail to Forecast Time Series In-Context?

Yufa Zhou[1*], Yixiao Wang[1*], Surbhi Goel[2], Anru R. Zhang[1]

[1]Duke University, [2]University of Pennsylvania

Why This Matters

Time Series Forecasting (TSF) underpins critical applications in healthcare, finance, energy, and transportation. Despite their sweeping success in language, vision, and multimodal tasks, **Transformers surprisingly underperform simple linear models** in long-horizon TSF—an effect observed across numerous empirical studies.

Contributions

Our work provides a **direct, rigorous explanation**: even when equipped with optimal parametrized Linear Self-Attention (LSA), Transformers **cannot outperform classical linear predictors on AR(p) processes**. Attention inevitably *compresses historical information*, giving LSA no structural advantage over OLS.

We identify an unavoidable **finite-sample excess-risk gap**, revealing the fundamental representational limits of attention-based architectures in TSF. These insights offer clear design guidance: when forecasting long-horizon time series, Transformer architectures may offer no performance benefits relative to classical **linear or frequency-domain models**.

Definition: AR(p) Process

We consider **AR(p) process** as our input data distribution:

A real-valued stochastic process $\{x_i\}_{i=1}^T$ follows an autoregressive model of order p, denoted AR(p), if there exist coefficients $\rho_1, \dots, \rho_p \in \mathbb{R}$ and white noise $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ such that for all $i > 0$,

$$x_{i+1} = \sum_{j=1}^{p} \rho_j x_{i-j+1} + \varepsilon_{i+1},$$

with fixed initial values $\{x_{-p+1}, \dots, x_0\}$.

Assuming the characteristic polynomial $1 - \rho_1 z - \cdots - \rho_p z^p$ has all roots outside the unit circle, i.e. $|z| > 1$, to ensure weak stationarity, the process satisfies: (1) $\mathbb{E}[x_i] = 0$, (2) $\mathbb{E}[x_i^2] = \gamma_0$, and (3) $\mathbb{E}[x_i x_{n+1}] = \gamma_{n+1-i}$, where $\gamma_k := \mathbb{E}[x_i x_{i+k}]$ and $r_k := \gamma_k / \gamma_0$.

Definition: Linear Self-Attention

We use Linear Self-Attention (LSA). Let $H \in \mathbb{R}^{(d+1)\times(m+1)}$ be the input matrix and define the causal mask $M := \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$. We denote the attention weights $P, Q \in \mathbb{R}^{(d+1)\times(d+1)}$. Then the linear self-attention output is defined as

$$\mathsf{LSA}(H) := H + \frac{1}{m} PHM(H^\top Q H) \in \mathbb{R}^{(d+1)\times(m+1)}.$$

Definition: Hankel Matrix

For $(x_1, \dots, x_n) \in \mathbb{R}^n$ and $p \le n$, define

$$H_n := \begin{bmatrix} x_1 & x_2 & \cdots & x_{n-p} & x_{n-p+1} \\ x_2 & x_3 & \cdots & x_{n-p+1} & x_{n-p+2} \\ \vdots & \vdots & & \vdots & \vdots \\ x_p & x_{p+1} & \cdots & x_{n-1} & x_n \\ x_{p+1} & x_{p+2} & \cdots & x_n & 0 \end{bmatrix} \in \mathbb{R}^{(p+1)\times(n-p+1)},$$

where each column is a sliding window of length $p+1$, with the last zero marking the prediction.

Our Setting & Build

Setup: univariate *AR(p)* with Hankelized context. We feed the matrix to an LSA-only Transformer and read the forecast from a masked **label slot** $\hat{x}_{n+1} := [(H_n)]_{(p+1,\, n-p+1)} \in \mathbb{R}$. Hankel layout embeds p lags and order, replacing positional encodings and enabling **in-context forecasting**.

Core Separation: Why Attention Loses

We establish a **strict representational gap**: Linear Self-Attention (LSA) compresses history into a restricted cubic feature space, whereas Linear Regression (LR) accesses the exact autoregressive lags.

- **The Result:** For any finite context n, LSA suffers a strict excess risk of $\rho^\top \Delta_n \rho$ with $\Delta_n \succ 0$, meaning it is provably worse than the optimal linear predictor.
- **The Mechanism:** The gap is **structural**, not due to estimation error. Attention inherently "blurs" the autoregressive signal through its reweighting scheme, creating a positive definite Schur-complement deficit that LR does not suffer.

Closed-form Gap: The $1/n$ Rate

The separation between Transformer models and linear predictors is structural and strictly positive for finite n.

- **Explicit Expansion:** We derive the gap as $\Delta_n = \frac{1}{n} B_p + o(1/n)$ with leading constant $B_p \succ 0$. This decay arises from the **overlap of sliding Hankel windows**, which acts as an implicit regularization mechanism.
- **Robustness:** The $\mathcal{O}(1/n)$ rate persists for general linear stationary processes; non-Gaussianity affects only the constant B_p through **higher-order cumulants**.
- **Depth Monotonicity:** Stacking layers enlarges the *Kronecker feature span*, yielding strictly monotone risk reduction. However, predictions remain bounded by the linear models, approaching it only asymptotically.

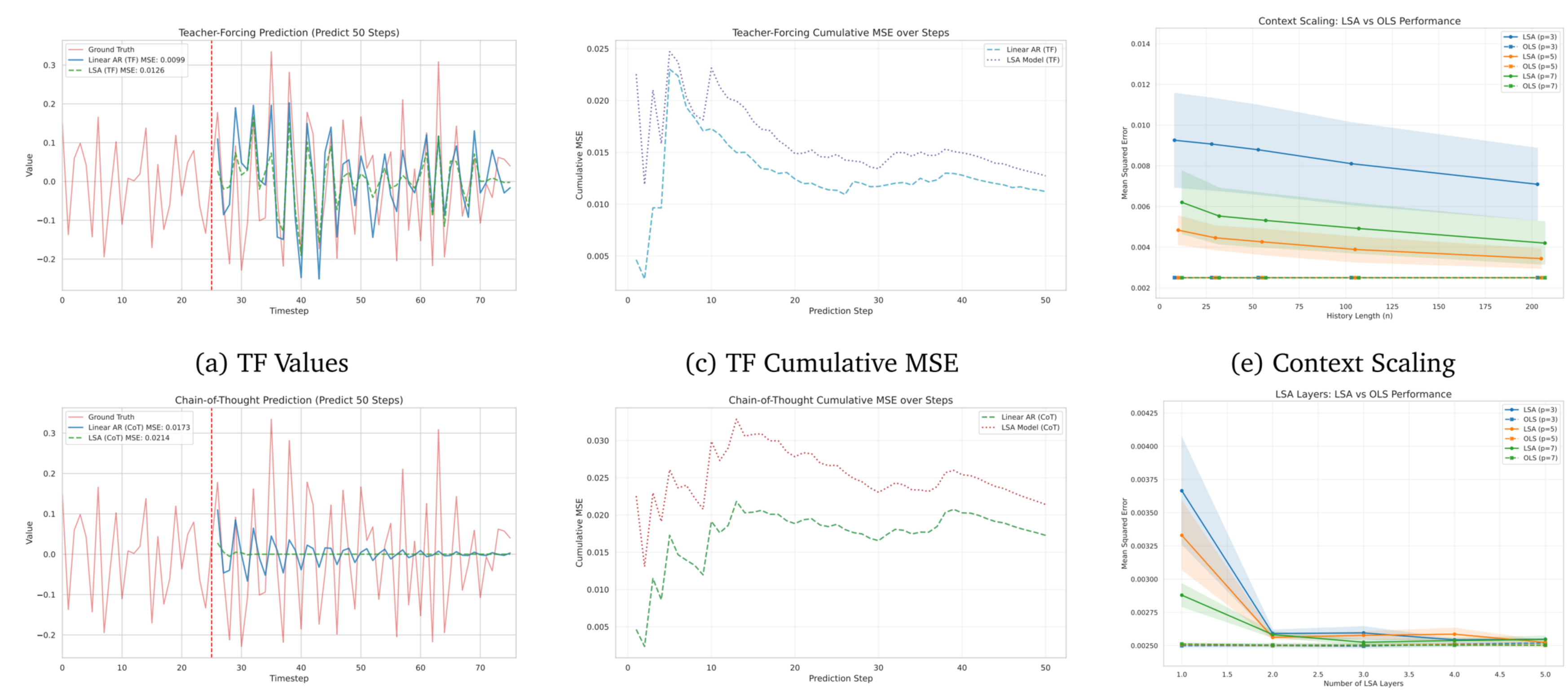
The CoT Trap: Exponential Collapse

Contrary to NLP, Chain-of-Thought (CoT) rollout is harmful for Time Series Forecasting.

- **Error Compounding:** Autoregressive feedback loops amplify prediction errors into input noise. The forecast rapidly collapses to the mean, with error growing **exponentially** toward the unconditional variance $\mathrm{Var}(x_t)$.
- **Uniform Dominance:** The Bayes linear predictor remains horizon-optimal, whereas LSA-based CoT is uniformly dominated at every step and fails significantly earlier under long-horizon rollout.

Experiments Confirm Theory

We validate on synthetic *AR(p)* (e.g., 50-step TF/CoT). Under Teacher-Forcing, LSA tracks dynamics yet **never surpasses** OLS. Under CoT, both collapse-to-mean; LSA fails earlier. Scaling context length n and layers reduces error but **saturates** at OLS. Softmax attention is slightly better, still below OLS.



(a) TF Values (c) TF Cumulative MSE (e) Context Scaling

(b) CoT Values (d) CoT Cumulative MSE (f) LSA Layer Scaling

Why Transformers Struggle

Our theory highlights a fundamental **inductive bias mismatch**. Attention mechanisms enforce a "contextual reweighting" that acts as lossy compression. While powerful for long-range semantic dependencies in language, this compression obscures the **exact local signals** required for low-order autoregressive dynamics.

Future Work

Extend strict gap analysis to **VAR/ARMA** processes and optimization dynamics. Pivot to models like Diffusion/Flow-Matching. Examine the role of MLP that might escape the attention bottleneck.

Yufa Zhou is looking for Research Scientist Summer Internship 2026 `yufa.zhou@duke.edu` Yixiao Wang is looking for CS PhD opportunities 2026 fall `yixiao.wang@duke.edu`