Two Layers of Attention Stability: A Koopman-Operator Analysis of Linear and Softmax Attention

Zhenglong Chen
Research Center for Mathematics and
Interdisciplinary Sciences
Shandong University
Qingdao, China
chenzhenglong@mail.sdu.edu.cn

Abstract

We study long-horizon attention stability by separating the $n \times n$ token-mixing matrix from the $d \times d$ projected dynamical operator. Under a fixed shared encoder, full rank, one head with full-width square projections, and an unmasked full window, the optimal dual operator equals the EDMD estimator, giving a Koopman/Galerkin interpretation and an approximation-error floor. Causal masking generally precludes this global identity. Independently, softmax and absolute-row normalization supply an input-uniform induced-norm bound for compatible token-mixing products, regardless of entry signs. Per-step spectral radius alone does not provide this bound, and neither certificate guarantees complete-network rollout stability. Non-affine LayerNorm also makes the full-window least-squares solution non-identifiable under a generic-rank assumption, whereas non-affine RMSNorm preserves rank. Across 120 synthetic runs, RN-LA and softmax meet a joint rollout-and-training-fit criterion in $18/20$ and $17/20$ runs, with no non-finite rollout. Exact spectral-radius and spectral-norm controls do not close this joint gap. Free-running ETTh1 and input-scale diagnostics complement the ETT/Weather forecasting experiments. The evaluated sign-general normalizers cost $O(n^2)$; structured spectral-radius constraints can be linear in sequence length, while the non-negative row-certified special case recovers existing normalized linear attention. The framework thus connects scoped projected-dynamics theory with token-mixing design criteria and finite-horizon empirical diagnostics.

1 Introduction

Attention mechanisms without a normalized token-mixing matrix can become unstable or collapse on long horizons, while softmax attention [17] is often comparatively robust. Kernel-method viewpoints [16, 2], expressivity gaps, and rank-collapse analyses [10] each capture part of the picture, but none fully accounts for the gap in *long-horizon prediction stability*.

Which object we analyse. Since "linear attention" names a family, we fix the object of study at the outset. *Unnormalized bilinear mixing* means $\hat{Z}_+ = (ZW_Q)(ZW_K)^\top(ZW_V)$, with no feature map and no row-normalizing denominator. Its global dual-operator identity is exact only in the fixed-encoder, full-rank, single-head, full-width-square-projection, unmasked full-window idealization. It is *not* the normalized linear attention of Katharopoulos et al. [4], which applies a non-negative ϕ and divides by $\sum_{j\le i}\phi(q_i)^\top\phi(k_j)$. That denominator already supplies the token-mixing certificate at

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

issue, so our contribution is to identify and test this mechanism, not to present it as a new linear-time architecture.

Viewing self-attention through encoder-induced observables, we separate two objects. **Architectural stability** is a bound on the $n \times n$ token-mixing matrix A: the matrix depends on the input, but a suitable normalization makes the *bound* hold uniformly over inputs. **Dynamical stability** concerns the spectrum of the $d \times d$ dual operator C in the fixed-encoder idealization and describes the dynamics projected onto that encoder.

In the unmasked full-window idealization, unnormalized bilinear mixing trained for next-step prediction solves an unconstrained EDMD problem: $C^* = K_{\mathrm{EDMD}}$. A causal mask instead introduces position-dependent prefix Gram matrices, so the causal experiments below test the token-mixing criterion rather than instantiate one global EDMD operator. Causal softmax makes A^{soft} row-stochastic and lower-triangular, hence $\|A^{\mathrm{soft}}\|_\infty = 1$ for every input, but W_V remains unconstrained. Thus $\|A_t\|_\infty \le 1$ certifies non-expansion of compatible token-mixing compositions; it does not guarantee bounded rollout of the complete model.

Section 4 instantiates this criterion with sign-general spectral and row-normalized controls. Both evaluated forms remain $O(n^2)$, although structured exact-ρ variants are linear in sequence length; the non-negative $O(nd^2)$ row-normalized special case recovers existing normalized linear attention. As a separate consequence, we show under a generic-rank assumption that non-affine LayerNorm makes the EDMD solution non-unique, non-affine RMSNorm rescaling preserves rank, and token-wise GroupNorm loses at least one rank direction per group.

Contributions.

- **Two-object framework** (Section 3): an input-uniform certificate for token mixing is separated from the fixed-encoder projected dynamics; it controls compatible token-mixing compositions on a common normed space but not the complete network.
- **Scoped EDMD interpretation** (Section 2): under the full-window assumptions of Theorem 1, the dual optimum equals K_{EDMD}. Its population limit is the Koopman/Galerkin projection with prediction floor $\sigma^2_{\mathrm{ale}} + \epsilon^2_d$ (App. D). These dynamical and approximation conclusions complement the elementary token-mixing bounds.
- **Criterion and controls** (Sections 3.5–4): $\|A_t\|_\infty \le 1$ is a sufficient certificate for token-mixing products regardless of sign (Prop. 5). RN-LA enforces it; estimated SN-LA and two exact controls diagnose what spectral normalization retains. These are not presented as a new $O(n)$ method.
- **Empirical diagnostic** (Section 5): a 120-run synthetic sweep separates numerical finiteness from training fit and prediction accuracy, with exact-normalization controls, free-running ETTh1 rollouts, and a retraining scale ablation. On ETT and Weather the normalized variants close 85–96% of the unnormalized-to-softmax gap at the longest horizon.

Roadmap. Section 2 fixes notation and establishes the restricted full-window EDMD identity; Section 3 develops the framework and its certificates; Section 4 gives the two controls; Section 5 reports experiments; Sections 6–7 cover related work and limitations.

2 Background: Full-Window Bilinear Mixing and EDMD

This section fixes notation and states the restricted full-window identity used for the projected-dynamics interpretation.

2.1 Notation and dual view of attention

We use row-major conventions, so rows are observation vectors and operators act by right multiplication. Let $\{x_t\} \subset \mathbb{R}^{d_x}$ be a state trajectory. The finite-sample identities below do not require stationarity; only the population-limit results in Appendix D assume a stationary ergodic Markov process with distribution μ. For a length-$(n{+}1)$ trajectory, define

$$Z, Z_+ \in \mathbb{R}^{n \times d}, \quad G := Z^\top Z, \quad M := Z^\top Z_+, \tag{1}$$

the input matrix and its one-step shift, and the empirical Gram and cross-correlation matrices. We assume (A1) local Markovianity within a window; (A2) Z has full column rank, so $G \succ 0$ (this is relaxed under LayerNorm; see Appendix F); (A3) square-integrable observables.

Definition 1 (Shared encoder)**.** *A* shared encoder *is a map $\mathcal{E} : \mathbb{R}^{d_x} \to \mathbb{R}^d$, fixed throughout the analysis, that plays two roles simultaneously: it is the dictionary of observables spanning the Koopman subspace $\mathcal{V}_d := \operatorname{span}\{\mathcal{E}_1, \dots, \mathcal{E}_d\}$ on which EDMD is performed,* and *it is the map producing the input tokens of the attention layer, so that $z_t = \mathcal{E}(x_t)$ and $Z = [z_1; \dots; z_n]$. Queries and keys are then linear in the same $\mathcal{E}$, i.e. $Q = ZW_Q$ and $K = ZW_K$ share the encoder rather than each learning their own. This is what makes the two objects in* (2) *comparable; Section 7 discusses relaxing it.*

Throughout, A with a superscript (A^{lin}, A^{soft}, $\tilde{A}$) always denotes an $n \times n$ *token-mixing* matrix, while M, G, C denote $d \times d$ matrices on the state space. For any square matrix B, its spectral radius is $\rho(B) := \max\{|\lambda| : \lambda \in \sigma(B)\}$.

Two views of attention. Every attention layer defines an $n \times n$ token-mixing operation, but a single global $d \times d$ dual operator exists only under additional assumptions. With a single head, full-width square projections $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$, and an *unmasked full window*, *unnormalized bilinear mixing*—the object studied here, distinct from the normalized linear attention of Katharopoulos et al. [4]—computes

$$\hat{Z}_+ \;=\; \underbrace{(ZW_Q)(ZW_K)^\top}_{A^{\mathrm{lin}} \in \mathbb{R}^{n\times n}} (ZW_V) \;=\; Z\,\big[\underbrace{W_Q W_K^\top G\, W_V}_{C \in \mathbb{R}^{d\times d}}\big], \tag{2}$$

admitting both forms simultaneously. Equation (2) is not a causal identity: after applying a causal mask, row i instead uses

$$\hat{z}_{+,i} = z_i\big(W_Q W_K^\top G_i W_V\big)\,, \qquad G_i := \sum_{j\le i} z_j^\top z_j, \tag{3}$$

so the prefix-dependent G_i generally precludes one global C shared by all rows. Our causal experiments therefore test token-mixing criteria, not a direct realization of a global EDMD operator. Softmax attention [17] similarly computes $\hat{z}_{+,i} = \sum_{j\le i} \alpha_{ij}(z_j W_V)$ with $\alpha_{ij} \propto \exp(\langle z_i W_Q, z_j W_K\rangle/\sqrt{d})$; it is a token-mixing operation $\hat{Z}_+ = A^{\mathrm{soft}}(ZW_V)$ but does not admit a global linear dual form.

2.2 One-paragraph primer on EDMD

The stochastic Koopman operator [8, 5, 1] is the conditional-expectation operator $(\mathcal{K}\phi)(x) := \mathbb{E}[\phi(x_+) \mid x]$. Its standard finite-sample approximation, *Extended Dynamic Mode Decomposition* (EDMD) [18], solves

$$K_{\mathrm{EDMD}} \;=\; \arg\min_{K\in\mathbb{R}^{d\times d}} \; \|ZK - Z_+\|_F^2 \;=\; G^{-1}M. \tag{4}$$

As $n \to \infty$, K_{EDMD} converges to the Galerkin truncation $K_\infty := \mathcal{G}^{-1}\mathcal{A}$ of $\mathcal{K}$ on the encoder subspace $\mathcal{V}_d := \operatorname{span}\{\mathcal{E}_1, \dots, \mathcal{E}_d\}$ [6], and the prediction error of any linear-in-$\mathcal{E}$ model is bounded below by the closed-form floor $L^\star = \sigma^2_{\mathrm{ale}} + \epsilon_d^2$ (Appendix D). This population convergence, rather than the finite-sample least-squares identity alone, supplies the Koopman interpretation of the restricted full-window dual.

2.3 Full-window bilinear mixing recovers K_{EDMD} at the optimum

Theorem 1 (Full-window bilinear mixing recovers EDMD)**.** *Let the shared encoder in Definition 1 be fixed, let Z have full column rank, and consider single-head, unmasked full-window bilinear mixing with full-width square projections $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$. For the induced variable $C = W_Q W_K^\top G W_V$, the objective $\mathcal{L}_C(C) := \|ZC - Z_+\|_F^2$ is strictly convex and has the unique minimizer $C^* = G^{-1}Z^\top Z_+ = K_{\mathrm{EDMD}}$. The factorization is surjective onto $\mathbb{R}^{d\times d}$, so the parameterized loss can attain C^*; this does not assert convexity in (W_Q, W_K, W_V).*

Proof sketch. As a function of the induced C, the loss is a convex quadratic whose first-order condition is $GC = M$, with unique solution $C^* = G^{-1}M$ because $G \succ 0$. Surjectivity follows by

choosing $W_Q W_K^\top = G^{-1}$ and $W_V = C_0$ for any target C_0. The multilinear weight parameterization remains nonconvex. Full proof in Appendix C. □

A canonical realization $C^* = K_{\mathrm{EDMD}}$ assigns

$$W_Q W_K^\top = G^{-1} \text{ (whitening)}, \qquad W_V = K_{\mathrm{EDMD}} \text{ (Koopman transition)}, \tag{5}$$

giving the cleanest interpretation of one reachable optimum. After reparameterizing by the induced C, EDMD and full-window bilinear mixing share the convex objective (4); the original optimization over (W_Q, W_K, W_V) is not the same convex problem. Tsai et al. [16] give an early kernel-method connection for softmax. Our analysis below treats this restricted equivalence as a starting point rather than a contribution.

2.4 The hidden cost: long-horizon divergence

Within its fixed-encoder, unmasked full-window idealization, Theorem 1 identifies the induced optimum with K_{EDMD}. The chosen encoder and sample determine this projected operator, whose spectral radius can exceed one.

Proposition 1 (Divergence under the fixed-operator idealization)**.** *Under the fixed-operator idealization $Z_{t+1} = Z_t C^*$, suppose $\rho(C^*) = \rho(K_{\mathrm{EDMD}}) > 1$ and Z_0 has a nonzero projection onto a generalized eigenspace associated with some $|\lambda| > 1$ (for the right action of C^*). Then $\|Z_t\|_F \to \infty$ with exponential rate, up to polynomial Jordan factors.*

Proof sketch. $Z_t = Z_0 (C^*)^t$. In a Jordan decomposition, the assumed nonzero unstable component scales as λ^t times a nonzero polynomial in t; norm equivalence then gives divergence. The projection assumption is essential, since an initial state confined to the complementary invariant subspace need not excite an unstable mode. Full proof in Appendix C. □

Projected operators with $\rho > 1$ can arise for unstable manifolds, non-stationary processes, and unstable directions of chaotic systems. Proposition 1 isolates the consequence only for a fixed operator and an excited unstable subspace; it does not claim that a causal attention rollout is governed by powers of one global C.

3 Token-Mixing Certificates and Projected Dynamics

Under the restricted full-window assumptions, $C^* = K_{\mathrm{EDMD}}$ is a projected operator whose spectrum should remain free to fit unstable modes. We instead seek an input-uniform *token-mixing nonexpansion certificate* for products of $n \times n$ mixing matrices. Such a certificate does not bound the complete model.

3.1 The Two-Layer Decomposition

Any attention layer admits the factorization

$$\hat{Z}_+ \;=\; \underbrace{A}_{\substack{n\times n \\ \text{token-mixing}}} \cdot \underbrace{Z\,W_V}_{\substack{n\times d \\ \text{value-transformed tokens}}}, \tag{6}$$

where $A^{\mathrm{lin}} = (ZW_Q)(ZW_K)^\top$ before masking and $A^{\mathrm{soft}} = \mathrm{softmax}(QK^\top/\sqrt{d})$ after masking. An induced-norm bound is *input-uniform* when it holds for every input. It controls any composable sequence of fixed-dimensional token-mixing suboperators on a common normed space, but not interleaved value/output projections, residual paths, MLPs, or normalization. Full causal rollout is instead a state-dependent composition of nonlinear updates, not generally a product of attention matrices; A is recomputed from the evolving context. The global $d \times d$ dual C exists only in the restricted unmasked setting, not generally under causal prefix operators.

Remark 1 (Design principle)**.** *A bound on A is a token-mixing certificate in a specified norm, not a complete-model stability guarantee; we therefore do not directly impose $\rho(C) \le 1$ where a global C exists.*

3.2 Softmax: A Token-Mixing Certificate Only

Causal softmax has an input-uniform token-mixing certificate, while its value map remains unconstrained.

Remark 2 (Structure of A^{soft})**.** *Immediately from the definition of softmax and of the causal mask, $A^{\mathrm{soft}} \in \mathbb{R}^{n\times n}$ is non-negative, row-stochastic and lower-triangular: $A^{\mathrm{soft}}_{ij} \geq 0$, $\sum_j A^{\mathrm{soft}}_{ij} = 1$, and $A^{\mathrm{soft}}_{ij} = 0$ for $j > i$. Consequently $\|A^{\mathrm{soft}}\|_\infty = 1$, and, the matrix being triangular, its eigenvalues are its diagonal entries $A^{\mathrm{soft}}_{ii} \in (0, 1]$, so also $\rho(A^{\mathrm{soft}}) \leq 1$ (Appendix C).*

Proposition 2 (Softmax value map is unconstrained)**.** *Although $\|A^{\mathrm{soft}}\|_\infty = 1$ certifies its token-mixing products, $\|\hat{z}_{+,i}\|_2 \leq \|W_V\|_2 \max_{j\leq i} \|z_j\|_2$ and $\|W_V\|_2$ is unconstrained. Hence it does not certify the complete attention map.*

3.3 Unnormalized Bilinear Mixing: No Uniform Certificate

Proposition 3 (No token-mixing certificate without normalization)**.** *For $A^{\mathrm{lin}} = (ZW_Q)(ZW_K)^\top$, neither an induced norm nor the spectral radius has a finite input- and weight-uniform upper bound: scaling W_Q scales A^{lin}.*

3.4 Design Principle: Certify Token Mixing, Free the Dynamics

The design rule is to constrain A^{lin} in a chosen induced norm without directly capping W_V or the restricted projected operator. A spectral-radius cap gives only a weaker single-matrix statement.

3.5 Which Certificate? Spectral Radius Does Not Control Products

The spectral radius cannot provide a compositional token-mixing certificate, even within the causal lower-triangular class.

Proposition 4 (Spectral radius does not control time-varying products)**.** *Per-step $\rho(A_t) \leq 1$ does not bound $\|A_T \cdots A_1\|$. Indeed, the causal lower-triangular shear*

$$S = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

satisfies $\rho(S) = 1$, while its T-fold product is $S^T = \left(\begin{smallmatrix} 1 & 0 \\ T & 1 \end{smallmatrix}\right)$, whose norm grows with T. A state-dependent sequence is therefore not controlled by per-step spectral radii.

Proposition 5 (Induced norms control token-mixing products)**.** *If $\|A_t\| \leq 1$ for all t in an induced operator norm, then $\|A_T \cdots A_1\| \leq 1$ for every horizon and every sequence of mixing matrices, in that same norm. This statement concerns the A-product only. For the ℓ_∞-induced norm,*

$$\|A\|_\infty \;=\; \max_i \sum_j |A_{ij}| \;=\; \sup_{v\neq 0} \frac{\|Av\|_\infty}{\|v\|_\infty}, \tag{7}$$

so $\|A\|_\infty \leq 1$ is equivalent to pointwise non-expansion, $\|Av\|_\infty \leq \|v\|_\infty$.

Remark 3 (Which norm to use)**.** *Non-negativity is irrelevant to* (7)*; absolute-row normalization works for signed entries. Exact $\|A_t\|_2 \leq 1$ likewise certifies the A-product in ℓ_2, but not W_V or the complete model. Its weaker empirical performance below does not invalidate that certificate.*

For complete block maps with $\mathrm{Lip}(T_\ell) \leq \gamma_\ell$, composition gives $\mathrm{Lip}(T_L \circ \cdots \circ T_1) \leq \prod_\ell \gamma_\ell$, while a residual block satisfies $\mathrm{Lip}(I + \alpha_\ell F_\ell) \leq 1 + |\alpha_\ell| \,\mathrm{Lip}(F_\ell)$ (Proposition 6). Thus even a stricter per-layer spectral-radius bound cannot certify a deep network unless the projections, head aggregation, residual path, MLP, normalization, and nonlinearities are bounded compatibly.

Causal softmax and RN-LA give input-uniform ℓ_∞ certificates; exact spectral-norm normalization gives an ℓ_2 certificate; exact spectral-radius normalization does not control products; and finite-iteration SN-LA has no strict certificate. Separately, non-affine LayerNorm generically makes the restricted full-window least-squares solution non-identifiable, non-affine RMSNorm preserves rank, and token-wise GroupNorm admits a group-dependent rank bound (Appendix F).

4 Stabilized Linear Attention via Token-Mixing Constraints

Section 3 motivates constraints on the $n \times n$ token-mixing matrix without directly capping W_V or, where defined, the restricted projected operator. Such constraints certify only the specified token-mixing operation or product; they do not bound the complete attention layer or guarantee that a model cannot diverge. We study spectral normalization (SN-LA) and row normalization (RN-LA) under this scope.

4.1 Unified framework: constrained optimization on the token-mixing matrix

We formalize the design rule as a constrained next-step prediction problem:

$$\min_{W_Q, W_K, W_V} \left\| \tilde{A}^{\mathrm{lin}} Z W_V \; - \; Z_+ \right\|_F^2 \qquad \text{s.t.} \qquad \mathcal{B}\big(\tilde{A}^{\mathrm{lin}}\big) \; \le \; 1. \tag{8}$$

Here $\tilde{A}^{\mathrm{lin}} = \mathcal{N}(A^{\mathrm{lin}})$, with $\mathcal{B} = \rho$ for ideal exact-ρ normalization and $\mathcal{B} = \|\cdot\|_\infty$ for RN-LA. In causal experiments, $A^{\mathrm{lin}} = ((ZW_Q)(ZW_K)^\top) \odot L$, where $L_{ij} = \mathbb{1}[j \le i]$. The constraint acts on the realized mixing matrix, not on parameter norms, and imposes no direct norm or spectral cap on W_V. We make no general claim that the normalized model admits a global C; causally, the prefix-dependent operators in (3) already preclude one in general.

4.2 Spectral-Normalized Linear Attention (SN-LA)

The ideal exact-ρ construction divides A^{lin} by its spectral radius whenever the latter exceeds one:

$$\tilde{A}^{\mathrm{lin}}_{\mathrm{SN}} \; := \; \frac{A^{\mathrm{lin}}}{\max\big(1,\, \rho(A^{\mathrm{lin}})\big)}, \qquad \hat{Z}_+ \; = \; \tilde{A}^{\mathrm{lin}}_{\mathrm{SN}}\,(ZW_V). \tag{9}$$

The evaluated implementation is different: it uses five power-iteration steps with a persistent buffer [9]. Its Rayleigh-type estimate converges to the spectral norm $\sigma_{\max}(A^{\mathrm{lin}})$, not to $\rho(A^{\mathrm{lin}})$. At convergence, $\rho \le \sigma_{\max}$ would make spectral-norm scaling conservative relative to exact-ρ scaling, but a *finite* iteration estimate need not upper-bound either quantity. Hence the five-step implementation has no strict certificate, and Theorem 2 below does not apply to it. Implementation pseudocode appears in Appendix A.

To separate the estimator from the constraint, Section 5 also evaluates two diagnostic controls in which no estimation occurs: normalization by the *exact* $\rho(A^{\mathrm{lin}})$ and by the *exact* $\|A^{\mathrm{lin}}\|_2$. The evaluated controls still materialize A. However, the spectral-radius constraint itself can exploit additional structure, as formalized below; exact $\|\cdot\|_2$ remains a dense diagnostic.

Theorem 2 (Ideal exact-ρ property)**.** *When* (9) *uses the exact spectral radius, it satisfies* $\rho(\tilde{A}^{\mathrm{lin}}_{\mathrm{SN}}) \le 1$ *for every realized* A^{lin}*. This is a single-matrix spectral statement, not a token-mixing product or full-model non-expansion certificate.*

Proof. For any matrix M and scalar $c > 0$, $\rho(M/c) = \rho(M)/c$. With $c = \max(1, \rho(A^{\mathrm{lin}}))$, we have $\rho(\tilde{A}^{\mathrm{lin}}_{\mathrm{SN}}) = \rho(A^{\mathrm{lin}})/c \le 1$ in both cases ($c = 1$ when $\rho \le 1$, $c = \rho$ when $\rho > 1$). □

Structured linear-in-n exact-ρ controls. Although our controls materialize A, the constraint itself need not: for unmasked $QK^\top$, the nonzero eigenvalues are those of $K^\top Q$; causally, a running maximum of $|q_i^\top k_i|$ gives a prefixwise exact-ρ scaling. Both have cost linear in n for fixed d (Proposition 7). With a separate token bound $\|z_i\|_2 \le B$, the conservative parameter condition $\|W_Q\|_2 \|W_K\|_2 \le B^{-2}$ also suffices causally (Remark 5). These untested alternatives still lack an induced-norm product certificate.

Remark 4 (Single matrix vs product)**.** *Exact* $\rho \le 1$ *does not bound* $\|\tilde{A}_T \cdots \tilde{A}_1\|$ *because* ρ *is not submultiplicative. By contrast, exact* $\|\tilde{A}_t\|_2 \le 1$ does *bound the corresponding token-mixing product in* ℓ_2*, though it neither implies an* ℓ_∞ *certificate nor controls* W_V *and the complete model. Experimentally, exact-*ρ *improves raw numerical finiteness, but neither exact control closes the full model's joint rollout-and-training-fit gap; estimation error is therefore not the sole explanation.*

4.3 Row-Normalized Linear Attention (RN-LA)

A second realization mimics the row-normalization of softmax structurally, enforcing an at-most-unit row-ℓ_1 norm and causal masking, but unlike softmax allows negative entries. Define

$$\tilde{A}^{\mathrm{lin}}_{\mathrm{RN},ij} \;:=\; \frac{A^{\mathrm{lin}}_{ij} \cdot \mathbb{1}[j \le i]}{\sum_{j' \le i} \big|A^{\mathrm{lin}}_{ij'}\big| \;+\; \varepsilon}, \tag{10}$$

with $\varepsilon > 0$ a small numerical safeguard. The implementation is parameter-free and one-shot (Appendix A).

Theorem 3 (RN-LA token-mixing certificate)**.** *$\tilde{A}^{\mathrm{lin}}_{\mathrm{RN}}$ in* (10) *satisfies $\|\tilde{A}^{\mathrm{lin}}_{\mathrm{RN}}\|_\infty \le 1$ for every realized causal mixing matrix, and consequently $\rho(\tilde{A}^{\mathrm{lin}}_{\mathrm{RN}}) \le 1$. The claim concerns token mixing only.*

Proof. Writing $D_i = \sum_{j' \le i} |A^{\mathrm{lin}}_{ij'}|$, each absolute row sum is $D_i/(D_i + \varepsilon) \le 1$. Thus $\|\tilde{A}^{\mathrm{lin}}_{\mathrm{RN}}\|_\infty \le 1$, and $\rho(M) \le \|M\|_\infty$ gives the spectral consequence. □

Corollary 1 (RN-LA token-mixing products are non-expansive)**.** *For all v, $\|\tilde{A}^{\mathrm{lin}}_{\mathrm{RN}} v\|_\infty \le \|v\|_\infty$, and $\|\tilde{A}_T \cdots \tilde{A}_1\|_\infty \le 1$ for every product of RN-LA mixing matrices. This input-uniform certificate excludes the interleaved W_V, output projections, residual paths, and all other model components, so it does not guarantee bounded complete-model rollout.*

RN-LA permits negative entries, yet absolute-row normalization still gives induced-ℓ_∞ non-expansion. Non-negativity is relevant instead to the prefix factorization discussed below; with standard positive feature maps, that factorization recovers existing normalized linear attention. Sander et al. [14] enforce doubly-stochastic non-negative attention via Sinkhorn iteration.

4.4 Comparison and complexity

RN-LA gives an input-uniform ℓ_∞ certificate for token-mixing products. Ideal exact-ρ normalization gives only a single-matrix spectral statement, whereas exact-$\|A\|_2$ normalization validly certifies the A-product in ℓ_2; the five-step SN implementation has no strict certificate. The experiments assess how these distinctions predict the complete model's joint rollout-and-training-fit criterion, not whether an induced norm is mathematically valid. All variants leave W_V and the rest of the layer uncontrolled.

The sign-general SN-LA and RN-LA *evaluated here* explicitly materialize $A^{\mathrm{lin}} \in \mathbb{R}^{n \times n}$ and cost $O(n^2)$ in sequence length. Proposition 7 shows that an exact spectral-radius cap can instead be imposed in time linear in n by exploiting low-rank or causal-triangular structure, but it retains only the weak single-matrix ρ statement. For the stronger row certificate, a non-negative feature map ϕ permits prefix sums $s_i = s_{i-1} + \phi(k_i)$ and $S_i = S_{i-1} + \phi(k_i) v_i^\top$, with $y_i = \phi(q_i)^\top S_i / (\phi(q_i)^\top s_i + \varepsilon)$. This costs $O(nd^2)$ time and $O(d^2)$ state, but with standard choices such as $\phi = \mathrm{elu} + 1$ it is precisely the existing normalized linear attention of Katharopoulos et al. [4]. Signed absolute-row normalization does not generally factorize. Our contribution is the certificate, the structured distinction between spectral and induced-norm controls, and their diagnostic use.

5 Numerical Experiments

We test the framework in two settings. The first is a controlled linear-Gaussian system in which the ground-truth spectral radius is a knob, so that the architectural claim of Section 3.5 can be isolated. The second is the standard ETT/Weather long-horizon forecasting suite. Two further experiments tied to the population-limit theorems, the LayerNorm solution manifold (Theorem 6) and the saturation of the linear-attention lower bound (Theorem 4), are reported in Appendix G.

5.1 Synthetic verification: finite-horizon token-mixing diagnostics

Setup. We generate trajectories $z_{t+1} = z_t K_{\mathrm{true}} + \eta_t$, $\eta_t \sim \mathcal{N}(0, \sigma^2 I_d)$, with $d = 8$, $\sigma = 0.05$, and sweep $\rho(K_{\mathrm{true}}) \in \{0.90, 0.95, 1.00, 1.05\}$. We train six architectures on next-step prediction (Adam, 2000 epochs, 5 seeds per architecture–radius cell, hence $6 \times 4 \times 5 = 120$ trained runs): vanilla LA, causal softmax, RN-LA, SN-LA, and two *controls* in which SN-LA's power-iteration estimate is

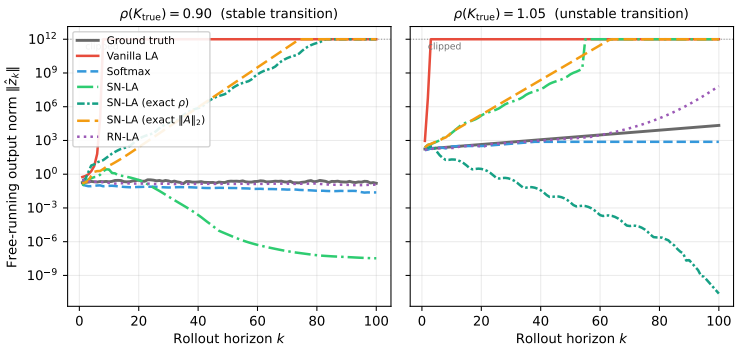


Figure 1: **Free-running output norm** $\|\hat{z}_k\|$, five-seed median with ground truth in grey; non-finite values are displayed at 10^{12}. The two sweep extremes are shown (all radii: Fig. 2). At $\rho = 0.90$ the transition is mean-square stable, so catastrophic output growth is model-generated. These curves measure magnitude, not prediction accuracy.

Table 1: **Runs meeting the joint criterion**, out of 5 seeds per cell (120 total): rollout stays below the stated threshold, training-window one-step relative error is below 0.5, and fitted $\rho(C_{\text{eff}})$ is within 30% of the target. The last two are training-fit screens, not held-out accuracy or non-collapse guarantees (App. G.3).

Architecture	ρ=0.90	0.95	1.00	1.05	**total**
RN-LA (ours)	**5/5**	**5/5**	**5/5**	3/5	**18/20**
Causal softmax	3/5	**5/5**	4/5	**5/5**	17/20
SN-LA, exact ρ	1/5	2/5	2/5	2/5	7/20
SN-LA (estimated)	3/5	0/5	0/5	0/5	3/20
SN-LA, exact $\|A\|_2$	2/5	3/5	1/5	0/5	6/20
Vanilla LA	1/5	2/5	2/5	0/5	5/20

replaced by the exact $\rho(A^{\text{lin}})$ and by the exact $\|A^{\text{lin}}\|_2$. The rollout context length equals the training sequence length, since A^{lin} is an unnormalized bilinear form whose row sums grow with sequence length, so a model trained at one length is not the same operator at another.

What to measure. We report numerical behavior separately from prediction accuracy. The free-running output norm $\|\hat{z}_k\|$ itself does not reference the target; the finite-horizon threshold test additionally uses the target scale and requires $\|\hat{z}_k\| < 10^6 \max\{1, \|z_k\|\}$ throughout the 100-step rollout, with all outputs finite. This is an empirical test, not a stability theorem. At $\rho(K_{\text{true}}) < 1$, even the oracle conditional-mean predictor approaches relative error 1 as stochastic trajectories lose memory, while at $\rho = 1.05$ growing target magnitude can inflate absolute error without increasing relative mismatch. We therefore include both stable and unstable transitions and report oracle-calibrated errors separately (App. G.1).

The row-bounded methods lead, with distinct exact controls. Table 1 shows that causal softmax and RN-LA, the two architectures with an input-uniform $\|A\|_\infty \le 1$ certificate, satisfy the joint criterion in $17/20$ and $18/20$ runs with no non-finite rollout across their 40 runs; the others score at most $7/20$. RN-LA permits negative entries, confirming that the row certificate does not require non-negativity. Estimated SN-LA's median pre-breakdown $\|A\|_\infty$ reaches 5–6. The exact-ρ and exact-$\|A\|_2$ controls improve on its $3/20$ but do not close the joint gap; exact-ρ improves raw finiteness, while exact $\|A\|_2$ remains a valid token-product certificate (App. G.1).

Full-step gain. The first-step diagnostic $\|A\|_\infty \|W_V\|_2$ broadly tracks the aggregate ordering, but already exceeds 1 for finite methods and is not a stability theorem (App. G.1).

Real-data and retraining-scale diagnostics. Two further studies appear in Appendix G.2–G.3. On ETTh1, 5 trained models per architecture evaluated on 20 held-out windows each give 100

free-running 96-step trajectories: softmax/RN-LA meet the threshold in all 100, while vanilla LA becomes non-finite in 92. Neither row-bounded method beats persistence in prediction error. In a retraining scale sweep, vanilla $\|A\|_\infty$ follows the c^2 law for $c \geq 0.5$ but remains near 9–12 at smaller tested scales, consistent with parameter compensation. This protocol-specific trend does not make input magnitude irrelevant.

Caveats. Table 1 does not eliminate long-horizon collapse; all models are underfitted relative to the oracle; and at $\rho = 1.05$ neither row-bounded method tracks the unstable mode (App. G.3). Separately, a nonlinear dictionary example illustrates how Koopman-invariant features remove a projected-dynamics approximation floor, whereas increasing dictionary size alone need not do so (App. D).

5.2 Long-horizon forecasting on real benchmarks

Setup. ETTh1/h2/m1/m2 [21] and Weather [19], 96-token lookback, $H \in \{96, 192, 336, 720\}$, standard splits as in Zeng et al. [20]; PatchTST backbone [11] with the attention block swapped, 5 seeds per cell. The decoder is patch-based rather than unrolled token-by-token, so the rollout failure mode of Section 5.1 does not directly apply; this benchmark tests whether the architectural constraint costs accuracy, not whether it prevents divergence.

Results. On ETTh1, vanilla LA's deficit to softmax widens from 0.042 at $H = 96$ to 0.079 at $H = 720$. On the five-dataset aggregate at $H = 720$, SN-LA and RN-LA close 85% and 96% of that gap, but softmax remains best. Both variants are numerically ahead of softmax on $3/5$ datasets at $H = 720$; treating the sub-0.001 Weather differences as ties gives $2/5$. We flag two counterexamples: on ETTh2 both are worse than softmax at every horizon, and on ETTm2 at $H = 720$ vanilla LA is best (0.350 versus 0.363 for softmax). Full results are in App. G.7; no significance or data-volume claim is made.

6 Related Work

Koopman-augmented Transformers [3, 7, 12] append spectrally-constrained Koopman modules or learn Koopman embeddings via autoencoders. We instead modify token mixing without directly capping W_V or, where the restricted full-window dual exists, $\rho(C)$. Nait Saada et al. [10] study rank collapse via spectral gaps. Tsai et al. [16] formalize softmax as Nadaraya–Watson kernel regression (used in Section 3.2; lower bound $\sigma^2_{\mathrm{ale},\mathcal{E}} + \epsilon^2_{\mathrm{soft}}$ in App. E).

Relation to normalized linear attention. The efficient-attention literature [4, 15, 2, 13] is usually read as pursuing kernel-approximation efficiency, with its denominator keeping kernel weights normalized. With a non-negative feature map, that denominator makes the token-mixing matrix row-stochastic, hence $\|A\|_\infty = 1$ for every input. By Proposition 5, this is a sufficient certificate for the time-varying *token-mixing product*, not for the complete model. Thus Katharopoulos et al. [4] already has the mechanism at issue, while RN-LA shows that the certificate itself does not require non-negativity. The criterion identifies variants that lose it by dropping the denominator, using a signed rather than absolute row normalizer, or otherwise failing to bound absolute row sums.

7 Discussion

The framework separates an input-uniform certificate for token mixing from the projected dynamics. Across 120 synthetic runs and 100 ETTh1 trajectories per architecture (600 total), causal softmax and RN-LA lead the joint synthetic criterion and are the only methods meeting the ETTh1 threshold in all 100 trajectories; exact-ρ is intermediate on ETTh1.

Limitations. (i) $\|A_t\|_\infty \leq 1$ is sufficient for non-expansion of compatible token-mixing compositions and correlates with finite-horizon behavior here; it is neither necessary nor sufficient for stability of the complete model, since W_V, output projections, residual paths, and MLPs remain uncontrolled. (ii) The evaluated sign-general normalizers materialize A and cost $O(n^2)$. Structured exact-ρ variants are linear in sequence length but retain only a single-matrix spectral statement; the non-negative row-normalized special case is also linear in sequence length but equivalent to existing normalized linear attention. (iii) The global EDMD identity assumes a fixed shared encoder, full rank, one

head with full-width square projections, and an unmasked full window; causal masking introduces prefix-dependent operators, and end-to-end encoder co-adaptation is outside the analysis. (iv) We have no architecture-specific transfer theorem for multi-head/deep Transformers, Mamba, or RWKV. For the latter two, the analogous diagnostic object would be an input-dependent recurrent transition or full-step Jacobian, not the attention matrix analyzed here. Our GroupNorm result is limited to token-wise coordinate groups, and affine LayerNorm or other normalization axes require separate analyses. **Impacts:** this theoretical study has no direct societal risks.

Acknowledgments and Disclosure of Funding

No specific research grant was received for this work. The author thanks the Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, for providing laboratory computing resources.

References

[1] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022. doi: 10.1137/21M1401243.

[2] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with Performers. In *International Conference on Learning Representations (ICLR)*, 2021.

[3] Ali Forootani, Mohammad Khosravi, and Masoud Barati. DeepKoopFormer: A Koopman enhanced transformer based architecture for time series forecasting. *npj Artificial Intelligence*, 2026. doi: 10.1038/s44387-026-00085-3. arXiv:2508.02616.

[4] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020.

[5] Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the Koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416, 2020.

[6] Milan Korda and Igor Mezić. On convergence of extended dynamic mode decomposition to the Koopman operator. *Journal of Nonlinear Science*, 28:687–710, 2018.

[7] Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018.

[8] Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

[9] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[10] Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. *arXiv preprint arXiv:2410.07799*, 2024.

[11] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

[12] Samuel E. Otto and Clarence W. Rowley. Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, 18(1):558–593, 2019.

[13] Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. Random feature attention. In *International Conference on Learning Representations (ICLR)*, 2021.

[14] Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[15] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning (ICML)*, 2021.

[16] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[18] Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

[19] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *NeurIPS*, 2021.

[20] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, 2023.

[21] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021.

A Pseudocode for SN-LA and RN-LA

We give reference Python implementations of the two token-mixing controls from Section 4.

Spectral-Normalized Linear Attention. The persistent buffer `u_buffer` is registered once per layer and updated across forward passes. The implementation used in our experiments applies five alternating power-iteration steps to the causally masked matrix. The resulting `sigma_hat` estimates the leading singular value, not the spectral radius, and a finite-step estimate is not a certified upper bound; hence estimated SN-LA has no strict induced-norm or spectral-radius guarantee.

```
def sn_linear_attention(z, w_q, w_k, w_v, u_buffer):
    """Apply causally masked estimated spectral normalization."""
    q, k, values = z @ w_q, z @ w_k, z @ w_v
    a_lin = causal_mask(q @ k.T)                # zero out j > i
    u_iter = u_buffer
    for _ in range(5):
        v_iter = a_lin.T @ u_iter
        v_iter = v_iter / v_iter.norm()
        u_iter = a_lin @ v_iter
        u_iter = u_iter / u_iter.norm()
    sigma_hat = (u_iter @ a_lin @ v_iter).abs()
    a_stab = a_lin / torch.clamp(sigma_hat.detach(), min=1.0)
    u_buffer.copy_(u_iter.detach())
    return a_stab @ values
```

Row-Normalized Linear Attention. RN-LA is parameter-free and one-shot.

```
def rn_linear_attention(z, w_q, w_k, w_v, eps=1e-6):
    """Apply causal absolute-row normalization."""
    q, k, values = z @ w_q, z @ w_k, z @ w_v
    a_lin = causal_mask(q @ k.T)                # zero out j > i
    row_l1 = a_lin.abs().sum(-1, keepdim=True)
    a_stab = a_lin / (row_l1 + eps)
    return a_stab @ values
```

B Notation Summary

Symbol	Meaning
$x_t \in \mathbb{R}^{d_x}$	raw state at time t
$\mathcal{E} : \mathbb{R}^{d_x} \to \mathbb{R}^d$	shared encoder (observable lift)
$Z \in \mathbb{R}^{n \times d}$	rows $\mathcal{E}(x_1), \ldots, \mathcal{E}(x_n)$
$Z_+ \in \mathbb{R}^{n \times d}$	one-step shift of Z
$G = Z^\top Z$, $M = Z^\top Z_+$	Gram and cross-correlation matrices
$K_{\mathrm{EDMD}} = G^{-1} M$	EDMD Koopman estimator
$K_\infty = \mathcal{G}^{-1} \mathcal{A}$	population EDMD limit (Galerkin truncation)
$W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$	attention projections (single head)
$A^{\mathrm{lin}} = (ZW_Q)(ZW_K)^\top$	unmasked bilinear token-mixing matrix
$A^{\mathrm{soft}} = \mathrm{softmax}(QK^\top/\sqrt{d})$	softmax-attention token-mixing matrix
$C = W_Q W_K^\top G W_V$	unmasked full-window dual operator
$\rho(\cdot)$	spectral radius
$\mathbf{1} \in \mathbb{R}^d$	all-ones vector
$H = I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$	centering projector
$\mathrm{LN}, \mathrm{RMS}$	row-wise LayerNorm, RMSNorm
$\mathcal{S}_{\mathrm{LN}}$	solution manifold under LayerNorm
$\sigma^2_{\mathrm{ale}}, \epsilon^2_d$	aleatoric variance, encoder truncation error

C Detailed Proofs from the Main Text

C.1 Proof of Theorem 1 (Section 2)

Proof of Theorem 1. (i) Under the fixed-shared-encoder, full-rank, single-head, full-width-square-projection, unmasked full-window assumptions, the loss $\mathcal{L}(C) = \|ZC - Z_+\|_F^2$ is convex quadratic in C. The first-order condition $\nabla_C \mathcal{L} = 2Z^\top(ZC - Z_+) = 0$ gives the normal equation $GC = M$, hence $C^* = G^{-1}M = K_{\mathrm{EDMD}}$. Uniqueness follows from $G \succ 0$ (assumption (A2)).

(ii) For any target $C_0 \in \mathbb{R}^{d\times d}$, choose $W_Q W_K^\top = G^{-1}$ and $W_V = C_0$. Then $C = G^{-1}\cdot G \cdot C_0 = C_0$, so the factorization is surjective onto $\mathbb{R}^{d\times d}$. □

C.2 Proof of Proposition 1 (Section 2)

Proof of Proposition 1. Write $C^* = PJP^{-1}$ in Jordan form. By assumption, $Z_0 P$ has a nonzero component in a Jordan block associated with some $|\lambda| > 1$. On that block, J^t contains terms $\lambda^t p(t)$ for a nonzero polynomial p, so norm equivalence under the fixed change of basis gives exponential divergence of $\|Z_0(C^*)^t\|_F$, up to polynomial Jordan factors. The projection assumption is essential. This proof concerns the stated fixed-C recursion, not causal masked attention, which recomputes prefix-dependent operators. □

C.3 Proofs from Section 3

Proof of Remark 2. Structure. (i) $\exp(\cdot) > 0$ and division by a positive normalizer preserves non-negativity. (ii) The softmax denominator equals $\sum_{j'\le i} \exp(\langle q_i, k_{j'}\rangle/\sqrt{d})$, so each row sums to one. (iii) Causal masking sets the pre-softmax logits at positions $j > i$ to $-\infty$, mapping to zero after exp.

Norm. By (i)–(ii), $\|A^{\mathrm{soft}}\|_\infty = \max_i \sum_j |A_{ij}^{\mathrm{soft}}| = \max_i \sum_j A_{ij}^{\mathrm{soft}} = 1$.

Spectrum. By (iii) the matrix is lower triangular, so its eigenvalues are its diagonal entries. Each diagonal entry

$$A_{ii}^{\mathrm{soft}} \;=\; \frac{e^{\langle q_i, k_i\rangle/\sqrt{d}}}{\sum_{j\le i} e^{\langle q_i, k_j\rangle/\sqrt{d}}} \tag{11}$$

is one component of a softmax output, hence lies in $(0, 1]$, and $\rho(A^{\mathrm{soft}}) = \max_i A_{ii}^{\mathrm{soft}} \le 1$. (This also follows from $\rho \le \|\cdot\|_\infty$; we record the diagonal form because it identifies *which* entries control the spectrum.) □

Proof of Proposition 2. For each row, non-negativity and row-stochasticity make $\hat{z}_{+,i} = \sum_{j\le i} A_{ij}^{\mathrm{soft}}(z_j W_V)$ a convex combination. Hence

$$\|\hat{z}_{+,i}\|_2 \le \max_{j\le i} \|z_j W_V\|_2 \le \|W_V\|_2 \max_{j\le i} \|z_j\|_2.$$

No architectural mechanism in softmax, the causal mask, or normalization of the token-mixing matrix constrains W_V. Therefore the bound can grow arbitrarily with $\|W_V\|_2$ and does not certify the complete attention map. □

Proof of Proposition 3. The matrix $A^{\mathrm{lin}} = (ZW_Q)(ZW_K)^\top$ has no normalization. For any nonzero realization, replacing W_Q by αW_Q scales A^{lin} and every induced norm by $|\alpha|$; suitable choices also make its spectral radius arbitrarily large. Hence no finite bound can hold uniformly over weights and inputs. □

Proposition 6 (Conditional propagation through depth)**.** *Let complete block maps T_ℓ act on a common normed space and satisfy $\mathrm{Lip}(T_\ell) \le \gamma_\ell$. Then*

$$\mathrm{Lip}(T_L \circ \cdots \circ T_1) \le \prod_{\ell=1}^{L} \gamma_\ell.$$

For a residual block $T_\ell = I + \alpha_\ell F_\ell$, one has $\mathrm{Lip}(T_\ell) \le 1 + |\alpha_\ell| \mathrm{Lip}(F_\ell)$. Consequently, a bound on the attention mixing matrix alone—including any stricter per-layer spectral-radius bound—does not certify a deep network unless the other block components are bounded in compatible norms.

Proof. The first inequality follows by repeated composition. For the residual form, $\|T_\ell(x) - T_\ell(y)\| \le \|x - y\| + |\alpha_\ell| \|F_\ell(x) - F_\ell(y)\|$. □

Proposition 7 (Exact-ρ constraints linear in sequence length)**.** *Writing $V := ZW_V$, the condition $\rho(\tilde{A}) \le 1$ can be imposed without forming an $n \times n$ matrix in two structured cases.*

1. *For unmasked $A = QK^\top$, the nonzero eigenvalues equal those of $K^\top Q$, so $\rho(A) = \rho(K^\top Q)$ and the normalized output is $Q(K^\top V)/\max(1, \rho(K^\top Q))$, evaluated in $O(nd^2 + d^3)$ time.*

2. *For causal prefixes, write $q_i, k_i, v_i \in \mathbb{R}^d$ for the transposes of row i of Q, K, V, respectively, and initialize $S_0 = 0$. Let $c_i := \max(1, \max_{s \le i} |q_s^\top k_s|)$, $S_i := S_{i-1} + k_i v_i^\top$, and $y_i := q_i^\top S_i / c_i$. This defines $\tilde{A} = D_c^{-1}((QK^\top) \odot L)$ with diagonal magnitude at most one, so $\rho(\tilde{A}) \le 1$, in $O(nd^2)$ time and $O(d^2)$ recurrent state.*

The second construction is prefixwise rather than the globally scaled SN-LA evaluated in Section 5; neither construction supplies an induced-norm product certificate.

Proof. The matrices $QK^\top$ and $K^\top Q$ have the same nonzero eigenvalues. Causally, $\tilde{A}$ is lower triangular with diagonal $\tilde{A}_{ii} = q_i^\top k_i / c_i$, proving the spectral bound. □

Remark 5 (What parameter regularization can certify)**.** *For causal triangular mixing,*

$$\rho(A) = \max_i |z_i W_Q W_K^\top z_i^\top| \le \left(\max_i \|z_i\|_2^2\right) \|W_Q\|_2 \|W_K\|_2.$$

Thus $\|z_i\|_2 \le B$ and $\|W_Q\|_2 \|W_K\|_2 \le B^{-2}$ suffice for $\rho(A) \le 1$, independently of sequence length. Without the token bound this is not input-uniform. For unmasked full-window mixing, the bound $\rho(QK^\top) \le \|Z\|_2^2 \|W_Q\|_2 \|W_K\|_2$ can grow with sequence length. Neither condition yields an induced-norm product certificate.

D Population Limit and Approximation Error Lower Bound

This appendix establishes the population counterpart of Theorem 1 for its fixed-encoder, unmasked full-window idealization: as $n \to \infty$, the optimum converges to the Galerkin truncation of the stochastic Koopman operator on the encoder subspace, and any linear-in-$\mathcal{E}$ model has an aleatoric-plus-truncation prediction floor.

D.1 Stochastic Koopman Operator

For a stochastic dynamical system with transition kernel $P(x_{t+1} \mid x_t)$, the *stochastic Koopman operator* $\mathcal{K} : L^2(\mu) \to L^2(\mu)$ is defined by

$$(\mathcal{K}\phi)(x) \;:=\; \mathbb{E}\big[\phi(x_{t+1}) \,\big|\, x_t = x\big]. \tag{12}$$

For deterministic systems this reduces to the classical Koopman operator since $\mathbb{E}[\phi(x_{t+1}) \mid x_t] = \phi(f(x_t))$.

D.2 Population Limit

Assume stationary ergodic sampling and a nonsingular population Gram matrix $\mathcal{G}$. As $n \to \infty$, the empirical matrices converge to their population counterparts:

$$\tfrac{1}{n} G \;\to\; \mathcal{G} := \mathbb{E}_\mu[\mathcal{E}(x)^\top \mathcal{E}(x)], \qquad \tfrac{1}{n} M \;\to\; \mathcal{A} := \mathbb{E}_\mu[\mathcal{E}(x)^\top (\mathcal{K}\mathcal{E})(x)]. \tag{13}$$

The EDMD estimator K_{EDMD} converges to $K_\infty := \mathcal{G}^{-1}\mathcal{A}$, the *Galerkin truncation* of $\mathcal{K}$ onto the encoder subspace $\mathcal{V}_d := \mathrm{span}\{\mathcal{E}_1, \ldots, \mathcal{E}_d\}$ [6]: K_∞ is the matrix representation of the orthogonal projection $\Pi_{\mathcal{V}_d} \mathcal{K}|_{\mathcal{V}_d}$.

D.3 Two Sources of Prediction Error

The fundamental prediction error decomposes into two components:

- **Aleatoric variance** (σ^2_{ale}): for stochastic systems, the next state is uncertain even given x_t. This noise is intrinsic to the data-generating process and cannot be reduced by any model.
- **Truncation error** (ϵ_d^2): if $\mathcal{K}$ maps observables outside $\mathcal{V}_d$, this creates an irreducible approximation error governed by how the chosen encoder span aligns with Koopman-invariant directions; it is not generally monotone in dictionary size.

A standard bias–variance argument (the cross-term vanishes because $\hat{g}(x) = \mathcal{E}(x)C$ is $\sigma(x)$-measurable) yields:

$$\mathbb{E}\left\|\mathcal{E}(x_+) - \mathcal{E}(x)C\right\|^2 \;=\; \sigma^2_{\text{ale}} \;+\; \mathbb{E}\left\|(\mathcal{K}\mathcal{E})(x) - \mathcal{E}(x)C\right\|^2, \tag{14}$$

where $\sigma^2_{\text{ale}} := \mathbb{E}\,\mathrm{tr}(\mathrm{Cov}[\mathcal{E}(x_+) \mid x])$. Infimizing the second term over $C \in \mathbb{R}^{d\times d}$ gives the encoder truncation error $\epsilon_d^2 := \|\Pi_{\mathcal{V}_d^\perp}\mathcal{K}\mathcal{E}\|^2_{L^2(\mu)}$.

Theorem 4 (Error floor for a fixed-encoder linear dual)**.** *For any $C \in \mathbb{R}^{d\times d}$ acting linearly on the fixed encoder,*

$$\mathbb{E}\left\|\mathcal{E}(x_+) - \mathcal{E}(x)\,C\right\|^2 \;\geq\; \sigma^2_{\text{ale}} \;+\; \epsilon_d^2. \tag{15}$$

The bound is attained when $C = K_\infty$, the population optimum of the linear-in-encoder class. Furthermore, $\epsilon_d^2 = 0$ if and only if $\mathcal{V}_d$ is $\mathcal{K}$-invariant.

Proof. The lower bound is immediate from (14): the aleatoric term is irreducible, and the parametric term is bounded below by ϵ_d^2 by definition. Attainment at $C = K_\infty$ follows because K_∞ is the matrix representation of $\Pi_{\mathcal{V}_d}\mathcal{K}|_{\mathcal{V}_d}$, which is the $L^2(\mu)$-orthogonal projection minimizer of $\|(\mathcal{K}\mathcal{E})(x) - \mathcal{E}(x)C\|^2_{L^2(\mu)}$. The invariance characterization follows from $\epsilon_d^2 = 0 \iff \Pi_{\mathcal{V}_d^\perp}\mathcal{K}\mathcal{E} = 0 \iff \mathcal{K}\mathcal{V}_d \subset \mathcal{V}_d$. □

Why this matters. At finite n, the restricted identity $C^* = K_{\text{EDMD}}$ is a least-squares fact. The population limit identifies that fixed-encoder full-window optimum with a specific Koopman-theoretic object, the Galerkin truncation, and makes its prediction floor depend on how the encoder span aligns with $\mathcal{K}$'s action.

Example: $\epsilon_d^2 = 0$ vs $\epsilon_d^2 > 0$. For an AR(1) process $x_{t+1} = \alpha x_t + \eta_t$ with mean-zero innovations and identity encoder $\mathcal{E}(x) = x$, the span of the coordinate observables is $\mathcal{K}$-invariant and $\epsilon_d^2 = 0$. By contrast, for a nonlinear system whose Koopman operator maps low-degree polynomial observables to higher-degree polynomials, a low-degree polynomial encoder can yield $\epsilon_d^2 > 0$. Changing or enriching the dictionary can reduce the floor only insofar as its span better approximates an invariant subspace; increasing dimension alone is not sufficient. Numerical verification is reported in Appendix G.

What ϵ_d depends on: invariance, not dictionary size. It is tempting to read ϵ_d^2 as a capacity term that shrinks as the dictionary grows. It is not. Consider the system with an exactly finite-dimensional Koopman lift,

$$x_1' \;=\; \lambda x_1, \qquad x_2' \;=\; \mu x_2 + c\,x_1^2, \tag{16}$$

with $\lambda = 0.9$, $\mu = 0.8$, $c = 1$. We draw $4{\cdot}10^5$ states from a trajectory with additive Gaussian noise of standard deviation 0.1, then evaluate the *deterministic* map in (16) at those states. Least squares uses empirically centered input and output features, equivalently an unpenalized intercept; dictionary sizes below count nonconstant features. The reported squared residuals illustrate deterministic closure under this sampling measure, rather than estimate the full stochastic Koopman prediction floor:

dictionary	**size**	ϵ_d^2
$\{x_1, x_2\}$	2	$4.0 \cdot 10^{-3}$
$\{x_1, x_2, x_1^2\}$	3	$\mathbf{5.4 \cdot 10^{-29}}$
$\{x_1, x_2, x_1^2, x_1^3\}$	4	$5.4 \cdot 10^{-29}$
all nonconstant monomials of degree ≤ 2	5	$4.2 \cdot 10^{-3}$

The three-dimensional dictionary $\{x_1, x_2, x_1^2\}$ spans an invariant subspace — x_1', x_2' and $(x_1')^2 = \lambda^2 x_1^2$ all lie in the span — so ϵ_d^2 is numerically zero. The *larger* five-dimensional dictionary of all quadratic monomials does not, because it contains x_2^2, whose image $(\mu x_2 + c x_1^2)^2$ requires degree-four terms; its truncation error is comparable to that of the two-dimensional linear dictionary. Enlarging the dictionary is therefore not the same as making it invariant, and ϵ_d measures the latter.

This delimits the scope of the whole framework. Where ϵ_d^2 dominates the floor, *no* model linear in the encoder can improve, however its token-mixing matrix is normalized: the two-layer analysis speaks to the architectural layer and to numerical stability, not to approximation power. For a linear-Gaussian system the two are far apart: in a sweep of 200 random rank-d encoders in ambient dimension 12, the *mean* truncation share is below 9% of the floor at every d (individual draws reach about 16%), and it is exactly zero at $d = 12$ where the subspace is trivially invariant. This is why truncation is a secondary, rather than absent, effect in the stability experiments of Section 5.1.

E Softmax Attention as Nadaraya–Watson Kernel Regression

This appendix complements the token-mixing analysis of Section 3.2 with an approximation-theoretic view. In the population limit, the softmax class has a parametric Nadaraya–Watson form and a bias–variance decomposition parallel to Theorem 4. The result enables a conditional comparison of approximation errors; it does not by itself prove that softmax is uniformly better than the linear-in-encoder class.

E.1 Softmax Attention as Kernel Regression

Softmax attention with shared input $Z \in \mathbb{R}^{n\times d}$ computes, for each query row $z_i = \mathcal{E}(x_i)$,

$$\hat{z}_{+,i} \;=\; \sum_{j=1}^{n} \alpha_{ij}\,(z_j W_V), \qquad \alpha_{ij} \;=\; \frac{\exp\big(\langle z_i W_Q,\, z_j W_K\rangle/\sqrt{d}\big)}{\sum_{j'=1}^{n}\exp\big(\langle z_i W_Q,\, z_{j'} W_K\rangle/\sqrt{d}\big)}. \tag{17}$$

Defining the parametric *exponential bilinear kernel*

$$\kappa_\theta(z, z') \;:=\; \exp\big(\langle z W_Q,\, z' W_K\rangle/\sqrt{d}\big), \qquad \theta := (W_Q, W_K), \tag{18}$$

equation (17) is a Nadaraya–Watson kernel regression with values $z_j W_V$:

$$\hat{z}_+(z;\,\theta, W_V) \;=\; \frac{\sum_{j=1}^{n}\kappa_\theta(z, z_j)\, z_j}{\sum_{j=1}^{n}\kappa_\theta(z, z_j)}\, W_V. \tag{19}$$

E.2 Population Limit

Proposition 8 (Population limit of softmax attention)**.** *Assume stationary ergodic sampling, $\mathcal{E}_j \in L^2(\mu)$ for each j, and bounded $\mathcal{E}$ so that $\kappa_\theta(z, \mathcal{E}(\cdot))$ is bounded for every θ and z. Then for μ_Z-almost every z with $\mathbb{E}_\mu[\kappa_\theta(z, \mathcal{E}(x'))] > 0$,*

$$\hat{z}_+(z;\,\theta, W_V) \;\xrightarrow{a.s.}\; \mathcal{F}_\theta(z)\, W_V \qquad (n \to \infty), \tag{20}$$

where the kernel-weighted mean operator *is the row vector*

$$\mathcal{F}_\theta(z) \;:=\; \frac{\mathbb{E}_{x'\sim\mu}\big[\kappa_\theta(z, \mathcal{E}(x'))\,\mathcal{E}(x')\big]}{\mathbb{E}_{x'\sim\mu}\big[\kappa_\theta(z, \mathcal{E}(x'))\big]} \;\in\; \mathbb{R}^{1\times d}. \tag{21}$$

Proof. Under boundedness of $\mathcal{E}$, both $x' \mapsto \kappa_\theta(z, \mathcal{E}(x'))\,\mathcal{E}(x')$ and $x' \mapsto \kappa_\theta(z, \mathcal{E}(x'))$ are bounded measurable. Birkhoff's theorem gives a.s. convergence of the empirical numerator and denominator in (19) to the corresponding μ-expectations; the continuous mapping theorem applied to division (positive denominator) gives (20). □

E.3 The Encoder-Conditional Koopman Expectation

Because the population output $\mathcal{F}_\theta(\mathcal{E}(x))W_V$ depends on x only through $z = \mathcal{E}(x)$, the relevant Bayes baseline is the conditional expectation given the encoder image rather than given x itself.

Definition 2 (Encoder-conditional Koopman expectation)**.** *Define* $g^\star : \mathbb{R}^d \to \mathbb{R}^d$ *by*

$$g^\star(z) \;:=\; \mathbb{E}\big[\mathcal{E}(x_+) \;\big|\; \mathcal{E}(x) = z\big]. \tag{22}$$

When $\mathcal{E}$ is injective, $\sigma(x) = \sigma(\mathcal{E}(x))$ and $g^\star(\mathcal{E}(x)) = (\mathcal{K}\mathcal{E})(x)$. When $\mathcal{E}$ is not injective, $g^\star$ is the projection of $\mathcal{K}\mathcal{E}$ onto the smaller σ-algebra $\sigma(\mathcal{E}(x))$, i.e., the best $\mathcal{E}$-measurable approximation of the Koopman expectation.

E.4 Aleatoric–Approximation Decomposition for Softmax

For any measurable $\hat{g} : \mathbb{R}^d \to \mathbb{R}^d$, an analogue of the linear bias–variance decomposition (14) holds with $g^\star$ replacing $\mathcal{K}\mathcal{E}$:

$$\mathbb{E}\left\|\mathcal{E}(x_+) - \hat{g}(\mathcal{E}(x))\right\|^2 \;=\; \sigma^2_{\mathrm{ale},\mathcal{E}} \;+\; \mathbb{E}\left\|g^\star(\mathcal{E}(x)) - \hat{g}(\mathcal{E}(x))\right\|^2, \tag{23}$$

where

$$\sigma^2_{\mathrm{ale},\mathcal{E}} \;:=\; \mathbb{E}\,\mathrm{tr}\big(\mathrm{Cov}[\mathcal{E}(x_+) \mid \mathcal{E}(x)]\big). \tag{24}$$

By the law of total variance, $\sigma^2_{\mathrm{ale},\mathcal{E}} \geq \sigma^2_{\mathrm{ale}}$, with equality when $\mathcal{E}$ is injective.

Specializing $\hat{g}(z) = \mathcal{F}_\theta(z)\, W_V$ and infimizing over the softmax parameters yields the population error of the softmax class.

Theorem 5 (Softmax-attention error lower bound)**.** *Under the assumptions of Proposition 8,*

$$\inf_{W_Q, W_K, W_V} \mathbb{E}\left\|\mathcal{E}(x_+) - \mathcal{F}_\theta(\mathcal{E}(x))\, W_V\right\|^2 \;=\; \sigma^2_{\mathrm{ale},\mathcal{E}} \;+\; \epsilon^2_{\mathrm{soft}}, \tag{25}$$

where the softmax approximation error *is*

$$\epsilon^2_{\mathrm{soft}} \;:=\; \inf_{W_Q, W_K, W_V} \mathbb{E}_{x\sim\mu}\left\|g^\star(\mathcal{E}(x)) - \mathcal{F}_\theta(\mathcal{E}(x))\, W_V\right\|^2 \;\geq\; 0. \tag{26}$$

Proof. Apply (23) with $\hat{g}(z) = \mathcal{F}_\theta(z)\, W_V$ and infimize the second term over (W_Q, W_K, W_V). □

What the comparison establishes. Theorems 4 and 5 are structurally parallel: each decomposes population loss into a conditional-variance term and a class-dependent approximation term. The fixed linear dual approximates $\mathcal{K}\mathcal{E}$ in the linear-in-$\mathcal{E}$ class with closed-form error ϵ^2_d; softmax approximates $g^\star$ in the stated parametric Nadaraya–Watson class with error $\epsilon^2_{\mathrm{soft}}$. Because the classes and their conditioning baselines differ, neither theorem supplies a universal ordering. Softmax is better only when the complete quantity $\sigma^2_{\mathrm{ale},\mathcal{E}} + \epsilon^2_{\mathrm{soft}}$ is below $\sigma^2_{\mathrm{ale}} + \epsilon^2_d$.

F Normalization Choice and Full-Window EDMD Identifiability

This appendix analyzes algebraic effects of LayerNorm, RMSNorm, and a specifically defined token-wise GroupNorm on the restricted full-window EDMD problem. It is separate from the token-mixing certificate and is not a complete-network stability result.

The preceding analysis concerns token-mixing certificates and a restricted projected operator. A separate algebraic issue is whether that full-window EDMD solution is unique. Under the assumptions below, non-affine LayerNorm destroys uniqueness whereas non-affine RMSNorm rescaling preserves rank.

F.1 The Hidden Cost of LayerNorm

Let $H := I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ denote the centering projector. We analyze non-affine row-wise LayerNorm, $\mathrm{LN}(z) := zH/\sigma_\varepsilon(z)$ with $\sigma_\varepsilon(z) := \sqrt{\frac{1}{d}\|zH\|^2 + \varepsilon}$ and $\varepsilon > 0$. Learned affine scale and bias can change the rank argument, so this result and the experiment in Section G.5 do not directly cover modern affine LayerNorm.

Lemma 1 (Null vector of Z_{LN})**.** $Z_{\mathrm{LN}}\mathbf{1} = \mathbf{0}_n$.

Proof. For each row i, $(Z_{\mathrm{LN}})_i \mathbf{1} = z_i H \mathbf{1}/\sigma_\varepsilon(z_i)$, and $H\mathbf{1} = \mathbf{1} - \frac{1}{d}\mathbf{1}(\mathbf{1}^\top \mathbf{1}) = \mathbf{1} - \mathbf{1} = \mathbf{0}$. □

Proposition 9 (Rank deficiency of G_{LN})**.** *The Gram matrix $G_{\mathrm{LN}} := Z_{\mathrm{LN}}^\top Z_{\mathrm{LN}}$ satisfies $G_{\mathrm{LN}}\mathbf{1} = \mathbf{0}$, hence $\mathrm{rank}(G_{\mathrm{LN}}) \le d - 1$. Generically (when no other linear constraint is satisfied by the rows of Z_{LN}), $\ker(G_{\mathrm{LN}}) = \mathrm{span}(\mathbf{1})$.*

Proof. $G_{\mathrm{LN}}\mathbf{1} = Z_{\mathrm{LN}}^\top (Z_{\mathrm{LN}}\mathbf{1}) = Z_{\mathrm{LN}}^\top \mathbf{0}_n = \mathbf{0}_d$ by Lemma 1. □

Theorem 6 (Solution manifold under a LayerNorm design)**.** *Let $Y \in \mathbb{R}^{n\times d}$ be any fixed target and suppose the generic condition $\ker(G_{\mathrm{LN}}) = \mathrm{span}(\mathbf{1})$ in Proposition 9 holds. The least-squares normal equation $G_{\mathrm{LN}} C = M_Y$, where $M_Y := Z_{\mathrm{LN}}^\top Y$, is consistent and admits a d-parameter family of solutions:*

$$\mathcal{S}_{\mathrm{LN}}(Y) \;=\; \left\{ K_Y^{(\mathrm{LN})} + \mathbf{1}v^\top \;:\; v \in \mathbb{R}^d \right\}, \tag{27}$$

where $K_Y^{(\mathrm{LN})} := G_{\mathrm{LN}}^+ M_Y$ is the minimum-Frobenius-norm solution. All elements of $\mathcal{S}_{\mathrm{LN}}(Y)$ produce identical predictions on the LayerNorm data manifold ($Z_{\mathrm{LN}}C$ is constant), but their spectra generically differ. Taking $Y = Z_{+,\mathrm{LN}}$ gives the LayerNorm–EDMD special case; the diagnostic in Section G.5 instead uses the unnormalized target $Y = Z_+$.

Proof. **Consistency:** $M_Y = Z_{\mathrm{LN}}^\top Y$ lies in the column space of $Z_{\mathrm{LN}}^\top$, which equals the column space of G_{LN}.

Solution structure: The general solution of a consistent linear system is one particular solution plus the kernel of the linear map $C \mapsto G_{\mathrm{LN}}C$. The kernel consists of all $D \in \mathbb{R}^{d\times d}$ with every column in $\ker(G_{\mathrm{LN}}) = \mathrm{span}(\mathbf{1})$, i.e., $D = \mathbf{1}v^\top$ for some $v \in \mathbb{R}^d$.

Loss-equivalence on the data manifold: $Z_{\mathrm{LN}}(C + \mathbf{1}v^\top) = Z_{\mathrm{LN}}C + (Z_{\mathrm{LN}}\mathbf{1})v^\top = Z_{\mathrm{LN}}C$ by Lemma 1.

Spectral divergence: By the matrix-determinant lemma,

$$\det(\lambda I_d - C - \mathbf{1}v^\top) = \det(\lambda I_d - C) - v^\top \,\mathrm{adj}(\lambda I_d - C)\,\mathbf{1},$$

so the eigenvalues of $C + \mathbf{1}v^\top$ depend continuously and generically nontrivially on v. □

Corollary 2 (Optimization landscape)**.** *Under the generic condition of Theorem 6, $\mathcal{L} = \|Z_{\mathrm{LN}}C - Z_{+,\mathrm{LN}}\|_F^2$ has a d-dimensional flat valley of global minima. The loss does not identify the component $\mathbf{1}v^\top$; which element an optimizer selects depends on its parameterization, initialization, and implicit bias. No generic departure from the minimum-norm solution is claimed.*

F.2 Non-Affine RMSNorm Preserves Rank

Proposition 10 (Full rank under RMSNorm)**.** *Any non-affine row-wise RMS normalization that divides each row by a strictly positive scalar—including $\tau_\varepsilon(z) := \sqrt{\frac{1}{d}\|z\|^2 + \varepsilon}$ and the additive-safeguard form $\sqrt{\frac{1}{d}\|z\|^2} + \varepsilon$, with $\varepsilon > 0$—acts as $Z_{\mathrm{RMS}} = D_\tau Z$, where $D_\tau \succ 0$ is diagonal. Hence $\mathrm{rank}(Z_{\mathrm{RMS}}) = \mathrm{rank}(Z)$, and if Z has full column rank then $G_{\mathrm{RMS}} \succ 0$ and the full-window EDMD solution is unique. Learned feature-wise scales require an additional nonzero-scale assumption and are not covered here.*

Proof. RMSNorm rescales each row by a positive scalar; it does not subtract any vector. Left-multiplication by an invertible diagonal matrix preserves rank. □

F.3 Conditional Extension to Token-Wise GroupNorm

Partition the d embedding coordinates into g nonempty groups $\mathcal{I}_1, \ldots, \mathcal{I}_g$. Here *token-wise GroupNorm* means that each token is centered and rescaled separately within these coordinate groups. This definition does not cover normalizers whose groups also span positions, batches, or other axes. Let $R \in \mathbb{R}^{d\times g}$ have as its r-th column the indicator of $\mathcal{I}_r$.

Proposition 11 (Rank bound for token-wise GroupNorm)**.** *For non-affine token-wise GroupNorm,* $Z_{\mathrm{GN}}R = 0$ *and therefore*

$$\operatorname{rank}(Z_{\mathrm{GN}}) \le d - g, \qquad \operatorname{rank}(Z_{\mathrm{GN}}^{\top} Z_{\mathrm{GN}}) \le d - g.$$

LayerNorm is the special case $g = 1$*. A fixed feature-wise affine scale with all entries nonzero and no bias preserves the* g *transformed null directions. Adding a shared feature-wise bias gives only the weaker bound* $\operatorname{rank}(Z_{\mathrm{GN,aff}}) \le d - g + 1$.

Proof. Centering within group $\mathcal{I}_r$ makes the coordinates in that group sum to zero for every token; subsequent division by a positive within-group scale preserves that equality. Hence $Z_{\mathrm{GN}}R = 0$. Because the group indicators have disjoint nonempty supports, $\operatorname{rank}(R) = g$, proving the first two bounds. An invertible diagonal feature scale preserves rank and maps the null directions accordingly. A shared bias adds a rank-one matrix $\mathbf{1}_n\beta^{\top}$, so the final bound follows from rank subadditivity. □

Pre-norm versus post-norm. In a pre-norm block, the normalized design directly defines the queries, keys, and Gram matrix of the current attention sublayer, so the preceding rank statements apply to that design. In a post-norm block, the current attention may use the unnormalized design, while normalization acts on the residual output and therefore on the next block's input. Neither placement yields complete-block non-expansion from a token-mixing bound alone: the residual map and the normalization Jacobian must also be controlled, as in Proposition 6. Affine LayerNorm and normalizers over different axes remain outside the exact solution-manifold theorem above.

F.4 Three Distinct Architectural Properties

The analysis distinguishes three properties with different scopes:

Property	Condition	Scope/result
Token-product non-expansion	$\|A_t\|_\infty \le 1$	causal softmax and RN-LA
Causal geometry	A_t lower triangular	all evaluated causal variants
Full-window dual uniqueness	$\operatorname{rank}(Z) = d$	RMS preserves; LN violates generically

Remark 6 (Takeaway)**.** *Unnormalized bilinear mixing has no input-uniform version of the first property, while non-affine LayerNorm generically violates the rank condition for the third. RN-LA supplies the first; estimated SN-LA has no strict certificate because its finite-iteration estimate need not upper-bound the realized norm. The first and third properties concern different mathematical objects, and neither guarantees stability of a complete network.*

G Numerical Verification of Theoretical Predictions

This appendix collects supporting empirical material referenced from Sections 3–5: (§G.1) additional diagnostics for the six-architecture synthetic sweep; (§G.4) training-time spectral diagnostics; (§G.5) verification of the LayerNorm solution-manifold theorem (Theorem 6); (§G.6) verification of the linear-attention lower bound (Theorem 4); (§G.7) per-dataset breakdown of the real-benchmark results from Section 5.2.

G.1 Additional diagnostics for the synthetic sweep

Tables 2–3 and Figures 3–4 expand Section 5.1. All numbers are over 5 seeds per cell; the full sweep is 6 architectures $\times$ 4 values of $\rho(K_{\mathrm{true}})$ $\times$ 5 seeds $=$ 120 runs, reproducible via `train_synthetic.py` followed by `plot_rebuttal.py`.

Training and rollout both use 100 context tokens, with Adam learning rate 10^{-3} for 2000 epochs and a 100-step rollout horizon. The three joint-criterion components are reported together in Table 1; the training-window fit screens do not certify held-out prediction accuracy.

The exact controls answer different questions. Exact-ρ does not certify products of changing matrices, but it substantially improves raw numerical finiteness; exact-$\|A\|_2$ does provide a valid induced-2-norm certificate for token-mixing products. Neither closes the joint rollout-and-training-fit gap in

Table 2: **Induced ℓ_∞ norm** $\|A\|_\infty = \max_i \sum_j |A_{ij}|$, measured (a) on training data at the end of training (5-seed mean), (b) at the first rollout step (5-seed median), and (c) as the largest value over the rollout *truncated at blow-up* (5-seed median). Truncation matters: once a run goes non-finite, the matrix computed from the corrupted context is meaningless. Softmax and RN-LA stay at or below 1 (reported as 1.000 at table precision). For estimated SN-LA, (b) is below 1 at $\rho = 0.90/0.95$ but above 1 at $\rho = 1.00/1.05$, while (c) is approximately 5–6.

Architecture	ρ=0.90	0.95	1.00	1.05
(a) training data				
Softmax / RN-LA	1.000	1.000	1.000	1.000
SN-LA (estimated)	0.894	0.825	1.513	2.991
SN-LA, exact $\|A\|_2$	0.812	0.926	1.626	3.052
SN-LA, exact ρ	2.671	3.894	7.087	17.616
Vanilla LA	5.667	6.776	8.393	104.613
(b) rollout step $k = 1$ (out of distribution)				
Softmax / RN-LA	1.000	1.000	1.000	1.000
SN-LA (estimated)	0.916	0.961	1.589	3.297
SN-LA, exact ρ	3.510	4.983	9.968	17.741
Vanilla LA	6.584	6.547	9.044	140.873
(c) rollout maximum, truncated at blow-up				
Softmax / RN-LA	1.000	1.000	1.000	1.000
SN-LA (estimated)	5.368	5.835	6.374	4.646
SN-LA, exact ρ	10.906	22.507	28.316	40.206
Vanilla LA	1981.1	313.7	407.8	70734.7

Table 3: **Single-step gain $\|A\|_\infty \|W_V\|_2$ at the first rollout step**, reported as the 5-seed median of the per-seed product. It broadly correlates with the aggregate stability ordering, although cells overlap and the cellwise median $\|W_V\|_2$ varies over $[0.58, 3.18]$. This is an empirical diagnostic rather than a sufficient condition: the gains already exceed 1 even for the methods that remain finite.

Architecture	ρ=0.90	0.95	1.00	1.05
RN-LA	1.66	1.78	1.82	3.18
Softmax	2.17	2.08	1.86	1.98
SN-LA (estimated)	2.27	2.71	4.67	5.84
SN-LA, exact $\|A\|_2$	2.38	2.54	4.64	6.19
SN-LA, exact ρ	6.76	7.91	17.08	10.34
Vanilla LA	12.81	9.14	9.11	80.77

Table 1, so estimation error is not the sole explanation, but these results neither make estimation irrelevant nor invalidate the exact 2-norm bound. Both controls leave W_V and the rest of the model unconstrained.

G.2 Token-by-token autoregressive rollout on ETTh1

Each model uses the seven ETTh1 variables directly as tokens and is trained with Adam (learning rate 10^{-3}) for 1500 updates, averaging next-step MSE over eight randomly sampled training windows per update. The supplementary script `train_ett_rollout.py` implements this protocol.

The forecasting benchmark of Section 5.2 uses a patch-based decoder that emits a whole horizon at once and never consumes its own output, so it cannot exhibit—or rule out—the rollout failure mode the framework concerns. This appendix supplies the missing diagnostic on real data: the same six architectures trained as single-layer next-step predictors on ETTh1 and then run *free-running*, token by token.

This is a stability diagnostic, not a benchmark comparison: there is no patching, feed-forward block, residual stream, or positional encoding, so its errors are not comparable to Table 6. We use the standard $12/4/4$-month split ($8640/2880/2880$ rows), training-split standardization, lookback 96, and horizon 96. Each architecture has 5 independently trained models (seeds), each evaluated from the same 20 test-window starts, yielding 100 trajectories rather than 100 independently trained models. Persistence has median relative error 0.878. The reported model error retains all finite predictions and

omits only a non-finite suffix, whereas $\|A\|_\infty$ is truncated immediately before the threshold failure; because failure-prone methods have different usable horizons, their median errors are not directly comparable.

Table 4: **ETTh1, free-running token-by-token rollout**: 5 trained models $\times$ 20 windows per architecture. "Stable" means finite output below $10^6 \max\{1, \|z_k\|\}$. Both row-bounded methods meet this threshold in all 100 trajectories. Error omits only non-finite suffixes, while $\|A\|_\infty$ is truncated before the first threshold failure; error is therefore not directly comparable across methods with different failure profiles.

Architecture	stable	non-finite	median rel. err	$\|A\|_\infty$	mean 1-step rel. err
RN-LA (ours)	**100/100**	**0**	**0.974**	**1.000**	0.619
Causal softmax	**100/100**	**0**	1.511	**1.000**	**0.472**
SN-LA, exact ρ	73/100	0	45.2	23.774	0.769
Vanilla LA	8/100	92	21.2	1018.5	0.818
SN-LA (estimated)	8/100	35	$9.8{\cdot}10^3$	4.702	0.716
SN-LA, exact $\|A\|_2$	7/100	0	$3.6{\cdot}10^4$	5.065	0.697

Among the two methods stable in all 100 trajectories, RN-LA has the lower median error, but the censored error column must not rank all six methods. Estimated SN-LA is stable in $8/100$ trajectories with 35 non-finite trajectories, whereas exact-$\|A\|_2$ is stable in $7/100$ with none non-finite. Both have pre-failure $\|A\|_\infty \approx 5$ and similar one-step fit, but their different censoring profiles preclude an accuracy ranking. Exact-ρ is markedly more finite ($73/100$ stable and no non-finite trajectory). Thus estimation affects raw numerical behavior, but neither spectral control closes the joint gap. Neither row-bounded method beats persistence over 96 steps (0.974 and 1.511 versus 0.878), so this experiment isolates numerical behavior rather than forecasting skill.

G.3 Input-scale ablation

At fixed W_Q, W_K, rescaling the input by c rescales $A^{\text{lin}} = (ZW_Q)(ZW_K)^\top$ by c^2. We test how retraining interacts with this scale dependence by retraining every architecture from scratch for each c, at $\rho(K_{\text{true}}) = 1.05$ with 5 seeds.

Table 5: **Input-scale ablation with retraining.** Top: 5-seed median $\|A\|_\infty$ on training data. Middle: the same median divided by c^2; a constant row indicates quadratic scaling of the median *retrained* solution. Bottom: runs that are stable and not collapsed. For fixed weights and an active homogeneous denominator, normalization cancels the input scale; the table records distinct solutions learned after retraining.

	c=0.05	0.1	0.25	0.5	1.0	2.0	4.0
$\|A\|_\infty$ *on training data*							
Vanilla LA	9.69	9.12	11.53	31.04	117.6	470.4	1882.7
Softmax / RN-LA	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SN-LA / exact $\|A\|_2$	3.059	3.059	3.059	3.059	3.059	3.059	3.059
SN-LA, exact ρ	13.15	14.40	14.31	15.95	15.95	15.95	15.95
$\|A\|_\infty/c^2$ *for vanilla LA (constant $\Leftrightarrow$ exactly quadratic)*							
Vanilla LA	3875	912	184	124	**117.6**	**117.6**	**117.7**
stable and not collapsed, out of 5 seeds							
Vanilla LA	0/5	0/5	0/5	0/5	0/5	0/5	0/5
Softmax	5/5	5/5	5/5	5/5	5/5	5/5	5/5
RN-LA	3/5	3/5	3/5	3/5	3/5	3/5	3/5
SN-LA, exact ρ	0/5	0/5	0/5	0/5	0/5	0/5	0/5

For $c \geq 0.5$, the median retrained vanilla solution follows the quadratic law closely; the displayed median ratio is constant to four significant figures only for $c \geq 1$. Below 0.5, the median $\|A\|_\infty$ remains 9–12, consistent with learned-parameter compensation; we do not attribute this to particular parameter matrices or claim it for every seed. Vanilla LA has no successful run in this sweep. The medians for softmax/RN-LA and estimated SN-LA/exact-$\|A\|_2$ are unchanged at the shown precision, whereas the exact-ρ median varies at the smaller scales. These observations apply only to the stated

retraining protocol and $c \in [0.05, 4]$: they do not show that magnitude is irrelevant or that sufficiently shrinking the input of a *fixed-weight* model could never yield a bound.

Two further readings of the bottom block. The exact-ρ control scores $0/5$ everywhere once collapse is excluded, even though its raw stability rate is 4–5 of 5: every one of those runs collapsed to a rollout norm below 10^{-3} of the ground-truth scale. RN-LA's $3/5$, by contrast, contains no collapsed run at any scale.

Caveats on the synthetic results

(i) Finite output and training fit do not rule out rollout collapse. At $\rho = 1.05$, where the oracle magnitude grows, the exact-ρ control's rollout norm falls to $2 \cdot 10^{-11}$ against a ground-truth scale of $2.2 \cdot 10^4$. All five runs are finite, and two still pass the training-window fit screens in Table 1. Thus that table is not a long-horizon accuracy or non-collapse certificate; the fit terms only reject some degenerate solutions. We do not label decay toward zero at $\rho < 1$ as collapse solely from its magnitude, because the oracle conditional mean also decays there.

(ii) All models are underfitted. Averaged over five seeds on the test trajectory at $k = 1$, the oracle's relative error is 0.607, 0.487, 0.098 and 0.001 for $\rho = 0.90, 0.95, 1.00, 1.05$. The best model mean is 0.110 (RN-LA) at $\rho = 1.05$, two orders of magnitude above the oracle. These are deliberately minimal single-head models without FFNs, residual connections, or positional encoding, isolating token-mixing differences. Causal masking still excludes the global EDMD identity. The comparison is between six imperfect models, and the numerical findings are relative, finite-horizon statements.

(iii) Neither stable architecture tracks an unstable ground truth. At $\rho = 1.05$, the five-seed median final rollout norm is 756 for softmax and $6.8 \cdot 10^7$ for RN-LA, against a median truth of $2.2 \cdot 10^4$: by the ratio of medians, one undershoots by $30\times$ and the other overshoots by $3000\times$. A descriptive least-squares fit to the training-window model outputs gives five-seed mean $\rho(C_{\text{eff}}) = 1.028$ and 1.024 versus 1.050, while vanilla LA gives 95.9. For these causal models, C_{eff} is a training-fit diagnostic, not the global EDMD operator of Theorem 1.

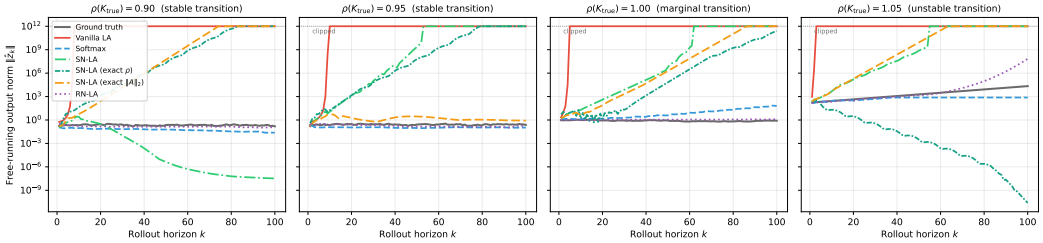


Figure 2: **Free-running output norm, all four ground-truth spectral radii.** Curves are five-seed medians; non-finite values are displayed at the fixed plotting ceiling 10^{12}. The two-panel version in Figure 1 shows the extremes. Softmax and RN-LA remain finite and avoid catastrophic non-finiteness over the plotted horizon; this does not assert that either follows the ground-truth magnitude accurately.

G.4 Training-time spectral diagnostics

We track $\rho(A)$ and $\|W_V\|_2$ per epoch for the four architectures trained on the synthetic linear-Gaussian system of Section 5.1 ($d = 8$, $\rho(K_{\text{true}}) = 1.05$, 5 seeds).

G.5 LayerNorm solution manifold (Theorem 6)

Setup. Linear-Gaussian system $Z_{i+1} = Z_i K_{\text{true}} + \epsilon_i$, $d = 4$, $n = 100$, $\sigma = 0.05$, with K_{true} stable ($\rho = 0.95$). For each of three design normalizations (None, RMSNorm, LayerNorm) we normalize Z only, retain the target $Y = Z_+$, (a) compute $K_Y = G^+ Z^\top Y$, and (b) train W_Q, W_K, W_V by Adam (lr$= 3 \cdot 10^{-3}$, 20,000 epochs, 3 seeds, float64). The implementation uses an additive denominator safeguard 10^{-6}; both normalizers therefore apply strictly positive row scalings after the relevant centering choice. Metrics: $m_1 = \|C - K_Y\|_F / \|K_Y\|_F$ and $m_3 = \|ZC - ZK_Y\|_F / \|ZK_Y\|_F$.

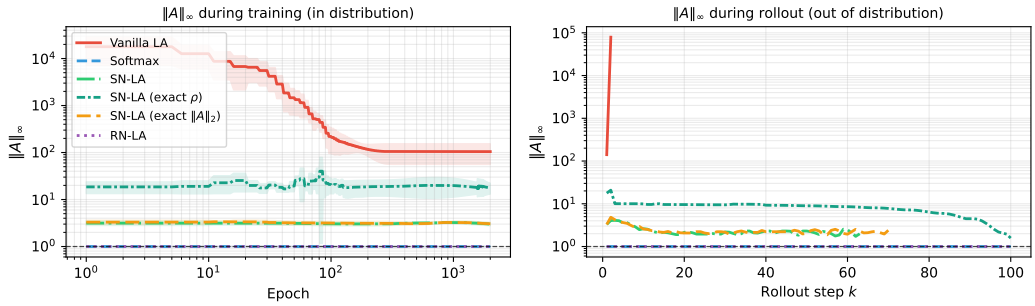
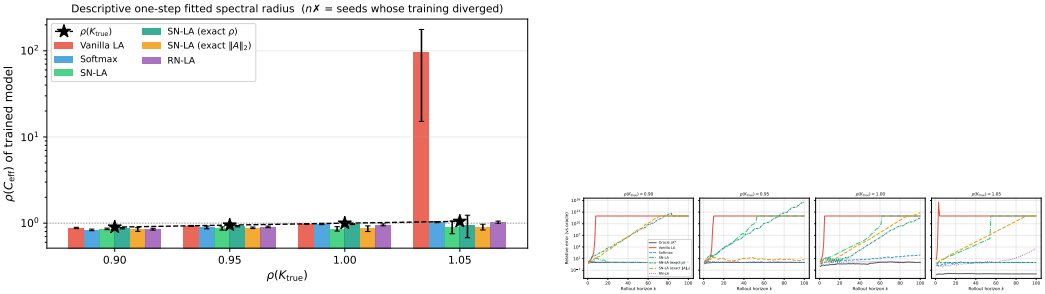


Figure 3: $\|A\|_\infty$ **in distribution (left) and during rollout (right).** Values after blow-up are masked. Softmax and RN-LA remain at or below 1 (reported as 1.000), since their row bound holds by construction. The finite-step SN estimate is not a certified upper bound; its realized $\|A\|_\infty$ rises to about 5–6 before breakdown.



(a) fitted $\rho(C_{\text{eff}})$.

(b) relative error against the oracle.

Figure 4: **(a)** Five-seed mean $\pm$ one standard deviation of the spectral radius of the descriptive one-step least-squares fit $C_{\text{eff}} = \arg\min_C \|ZC - \text{model}(Z)\|_F$ on the training window; Z has full column rank here. This is not a global operator for the causal models. At target 1.050, the means are 1.028 for softmax, 1.024 for RN-LA, and 95.9 for vanilla LA; annotations mark divergent seeds. **(b)** Relative rollout error with the oracle $z_{n-1}K_{\text{true}}^k$ overlaid. For $\rho < 1$, even the oracle approaches relative error 1 within a few steps, so this metric does not isolate numerical stability in that regime.

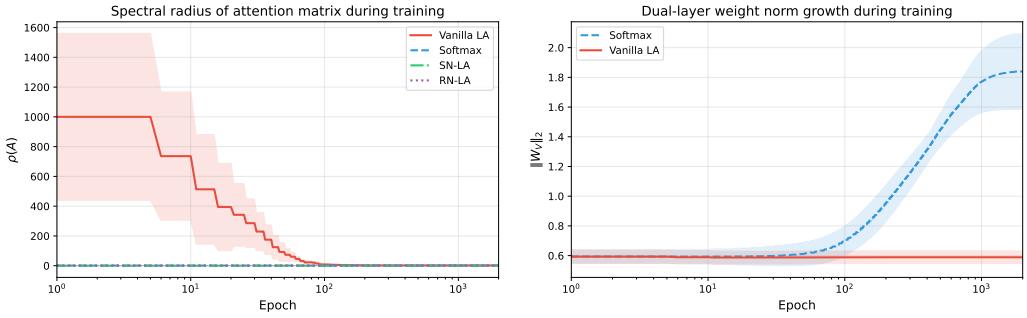
For LayerNorm, the end-of-training kernel-direction residual $\|\Delta - \mathbf{1}\hat{v}^\top\|_F$, with $\Delta = C - K_Y^{(\text{LN})}$ and $\hat{v} = \frac{1}{d}\Delta^\top\mathbf{1}$, ranges from $6.7{\cdot}10^{-7}$ to $2.5{\cdot}10^{-3}$ across seeds (two of three are below 10^{-3}). Thus the residual is close to the predicted kernel direction, with visible finite-optimization error in one seed.

G.6 Saturation of the fixed-encoder linear lower bound (Theorem 4)

Setup. Low-rank linear-Gaussian system, ambient $D = 4$, encoder dim $d = 2$ chosen so that the encoder span is *not* $\mathcal{K}$-invariant (hence $\epsilon_d^2 > 0$). Theorem 4 uses the vector risk $L^\star = \sigma_{\text{ale}}^2 + \epsilon_d^2 \approx 0.00505$; the figure reports per-coordinate MSE, whose matching floor is $L^\star_{\text{coord}} := L^\star/d \approx 0.00253$. We sweep training-sample size $n \in \{500, 1000, 2000, 5000, 10{,}000, 20{,}000\}$, train the full-window bilinear model on whitened data (3 seeds), and estimate risk on a held-out trajectory of length 200,000.

G.7 Per-dataset breakdown of the real-benchmark results

Table 6 gives the five-dataset aggregate referenced from Section 5.2; Table 7 expands it per dataset and per horizon. Bold marks best per row.



(a) $\rho(A)$ vs epoch.

(b) $\|W_V\|_2$ vs epoch.

Figure 5: **Training-time spectral diagnostics.** (a) Vanilla LA's realized token-mixing spectral radius settles above 1; softmax and RN-LA stay at or below 1 by construction, and estimated SN-LA is observed below 1 in these training traces despite lacking a finite-step certificate. (b) $\|W_V\|_2$ in softmax grows by $2.5\times$–$3.9\times$, consistent with the absence of a direct architectural cap on the value map (Prop. 2).

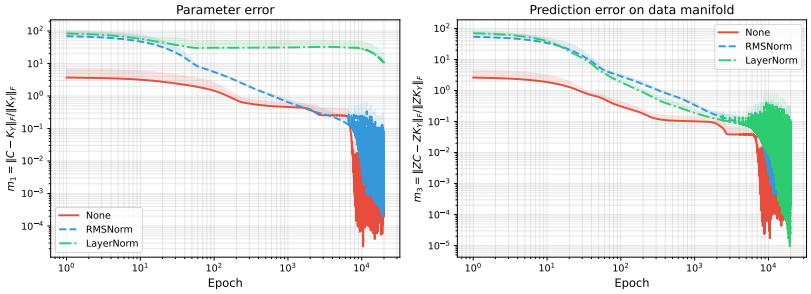


Figure 6: **LayerNorm solution-manifold diagnostic.** Under **None** and **RMSNorm**, most seeds drive both m_1 and m_3 near zero, although one RMSNorm seed remains at $m_1 = 0.048$ and $m_3 = 0.039$ after the fixed training budget. Under **LayerNorm**, parameter error remains $\mathcal{O}(1)$ (4.90–19.86) while prediction error is much smaller ($1.4{\cdot}10^{-6}$–$8.1{\cdot}10^{-3}$), consistent with the non-identifiability in Theorem 6; this finite-optimization experiment is not an exact convergence claim.

Table 6: **Test MSE averaged across the five datasets (ETTh1/h2/m1/m2, Weather), per horizon.** Lower is better. Softmax is best at every horizon on this aggregate; the stabilized variants close most of the vanilla-to-softmax gap. Per-dataset values, including cases where this ordering reverses, appear in Table 7.

Horizon H	Vanilla LA	Softmax	SN-LA	RN-LA
96	0.314	**0.286**	**0.286**	0.288
192	0.361	**0.328**	0.332	0.333
336	0.394	**0.371**	0.377	0.375
720	0.481	**0.455**	0.459	0.456

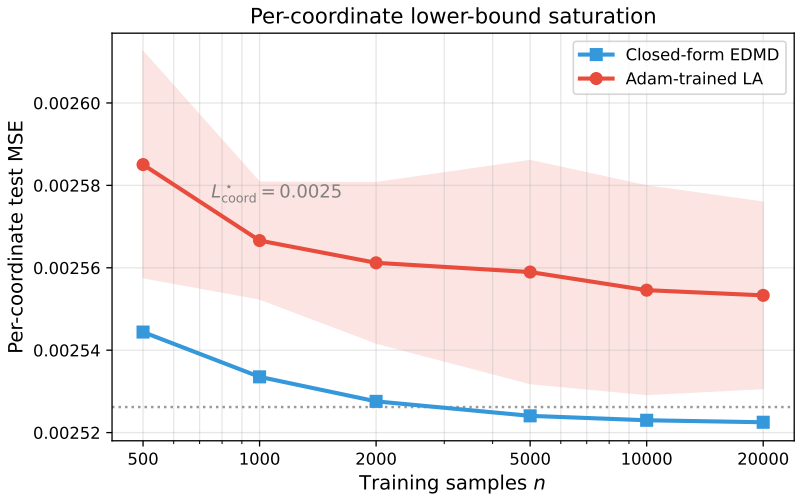


Figure 7: **Saturation of the per-coordinate population floor.** Closed-form EDMD and the Adam-trained full-window bilinear model approach the vicinity of $L^\star_{\text{coord}} = L^\star/d$ as n grows. The finite held-out estimate for EDMD crosses slightly below $L^\star_{\text{coord}}$ (by at most $3.7{\cdot}10^{-6}$), which is compatible with Monte Carlo test error and is not a claim that an empirical estimate obeys the population inequality pointwise.

Table 7: **Per-dataset test MSE (5-seed mean).** ETTh1 displays the clearest horizon-dependent gap. Method differences are usually small on ETTm and Weather, but ETTm2 at $H = 720$ is a clear exception; the table reports all reversals rather than asserting a uniform ordering.

Dataset	H	Vanilla LA	Softmax	SN-LA	RN-LA
ETTh1	96	0.483	0.441	**0.432**	0.435
	192	0.540	0.489	**0.486**	0.490
	336	0.572	0.525	**0.523**	0.529
	720	0.736	0.657	0.652	**0.645**
ETTh2	96	0.339	**0.248**	0.249	0.255
	192	0.379	**0.275**	0.293	0.288
	336	0.354	**0.287**	0.311	0.302
	720	0.415	**0.354**	0.363	0.365
ETTm1	96	0.400	**0.396**	0.400	0.402
	192	0.443	0.447	0.442	**0.441**
	336	**0.493**	0.499	0.504	0.499
	720	0.553	0.552	**0.542**	0.549
ETTm2	96	0.174	**0.170**	0.171	0.171
	192	0.222	**0.212**	0.216	0.222
	336	0.273	**0.270**	**0.270**	0.271
	720	**0.350**	0.363	0.389	0.372
Weather	96	0.177	**0.173**	0.177	0.177
	192	0.220	**0.219**	0.221	0.223
	336	0.276	**0.272**	0.276	0.274
	720	0.351	0.351	**0.350**	0.351

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The abstract and Introduction present the contribution as a token-mixing criterion and diagnostic. They distinguish the sufficient induced-norm certificate for products of token-mixing matrices from stability of the complete model, which is neither implied nor required by that certificate. They also scope the global EDMD identity to a fixed shared encoder, full rank, one head with full-width square projections, and an unmasked full window; state that causal experiments do not instantiate one global dual operator; and identify the evaluated sign-general methods as $O(n^2)$ controls rather than new linear-time architectures. The synthetic and real-data evidence is reported under finite-horizon, explicitly stated criteria in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes. Section 7 states that (i) $\|A_t\|_\infty \leq 1$ certifies only the token-mixing product and is neither necessary nor sufficient for stability of a complete model containing value/output projections, residual paths, and MLPs; (ii) the evaluated sign-general normalizers materialize A and cost $O(n^2)$, while the non-negative $O(nd^2)$ prefix-sum special case recovers existing normalized linear attention; (iii) the global EDMD identity requires a fixed shared encoder, full rank, a single head with full-width square projections, and an unmasked full window; and (iv) there is no general transfer theorem for deep or multi-head models, encoder co-adaptation, or affine LayerNorm. Appendix F provides a restricted non-affine token-wise GroupNorm extension and discusses pre/post-norm placement; Proposition 6 gives a conditional depth bound. Section 5 and Appendix G distinguish finite-horizon diagnostics from complete-network or forecasting guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be

used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes. Section 2.1 states the trajectory assumptions and Theorem 1 explicitly requires the five structural conditions used by the EDMD identity: a fixed shared encoder, full-column-rank Z, a single head, full-width square projections, and unmasked full-window mixing. The causal prefix construction is explicitly excluded from the global-C claim. Proposition 5 certifies only products of token-mixing matrices; ideal exact-ρ normalization has only a single-matrix spectral property, exact-$\|A\|_2$ validly certifies the token product in ℓ_2, and the finite five-step SN estimate is stated to have no strict guarantee. Appendix F treats non-affine LayerNorm and non-affine RMSNorm only, with the generic-kernel/full-rank assumptions stated; learned affine parameters are outside its scope. Proofs and scoped proof sketches appear in Appendices C–F.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes. The main synthetic sweep specifies six architectures—vanilla LA, causal softmax, RN-LA, five-step estimated SN-LA, exact-ρ SN-LA, and exact-$\|A\|_2$ SN-LA—over four values of $\rho(K_{\mathrm{true}})$ and five seeds per architecture–radius cell, for $6 \times 4 \times 5 = 120$ trained runs. Appendix A gives the causal five-step SN and RN implementations. Appendix G reports the ETTh1 design as five trained models times twenty windows per architecture, the five-seed scale sweep, and three seeds each for the LayerNorm and lower-bound diagnostics. ETT/Weather benchmark cells use five seeds. Data generation, splits, training budgets, metrics, censoring at non-finiteness, and the exact-control definitions are given in the paper and supplementary implementation materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets are publicly available, and we plan to release our experimental code. No public code-release link is currently included; Appendix A provides the attention implementations and Appendix G describes the experimental protocols, but a complete public reproduction package is not yet available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes. The main synthetic sweep uses $d = 8$, $\sigma = 0.05$, $\rho(K_{\text{true}}) \in \{0.90, 0.95, 1.00, 1.05\}$, Adam for 2000 epochs, and five seeds per cell for all six architectures, including the exact-ρ and exact-$\|A\|_2$ controls. Appendix G specifies the ETTh1 lookback/horizon and five-model-by-twenty-window evaluation, the seven-scale retraining sweep with five seeds, the LayerNorm diagnostic ($d = 4$, $n = 100$, 20,000 epochs, float64, three seeds), and the lower-bound sample-size sweep with three seeds and a 200,000-step held-out trajectory. The ETT/Weather benchmark uses standard chronological splits, 96-token lookback, horizons $\{96, 192, 336, 720\}$, a PatchTST backbone, and five seeds per cell; the corresponding implementation scripts are the authoritative source for optimizer and backbone details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports the number of seeds and defines each aggregation: the 120-run sweep gives counts out of five seeds per cell; rollout curves use a five-seed median; auxiliary tables distinguish means from medians; and the LayerNorm/lower-bound diagnostics provide seed ranges or mean–standard-deviation bands. However, the main synthetic and ETT/Weather comparisons do not include confidence intervals or formal significance tests. We therefore avoid significance claims, treat sub-0.001 Weather differences as tie-like, and answer No rather than implying more statistical resolution than the available runs support.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Partially. The local synthetic, LayerNorm, and lower-bound diagnostics used GPU acceleration or CPU computation, while the 400 ETT/Weather training runs used a remote high-performance GPU. Specific hardware models, per-run wall-clock time, peak memory, and sufficiently precise total runtime for every experiment are not provided, so the full checklist requirement is not met.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. This work is methodological and simulation-based, with no human-subject data collection or ethically sensitive intervention in deployment settings.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Section 7 identifies this as a theoretical study with no direct societal risks. It does not provide a detailed assessment of positive and negative downstream impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate

to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable. The present work does not release high-risk generative models or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes. We credit the original ETT [21], Weather [19], and PatchTST [11] sources and identify their applicable licenses: ETDataset, CC BY-ND 4.0; Weather, CC BY 4.0; and PatchTST, Apache-2.0. These public assets are used for the cited forecasting experiments under their respective terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Not applicable. No new dataset, pretrained model, or software package is released with this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable. The work is based on dynamical-system simulation and does not involve crowdsourcing or human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable because no human-subject study is conducted in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Not applicable. LLMs are not part of the proposed method, experiments, or scientific claims.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.