Residual Koopman Spectral Profiling for Predicting and Preventing Transformer Training Instability

Bum Jun Kim[*1] **Shohei Taniguchi**[1] **Makoto Kawano**[1] **Yusuke Iwasawa**[1] **Yutaka Matsuo**[1]

[1]Graduate School of Engineering, The University of Tokyo, Japan

Abstract

Training divergence in transformers wastes compute, yet practitioners discover instability only after expensive runs begin. They therefore need an initialization-time signal for identifying risky training configurations. We study Residual Koopman Spectral Profiling (RKSP), a fixed, single-forward-pass spectral risk score. RKSP extracts Koopman spectral features by applying whitened dynamic mode decomposition to layer-wise residual snapshots. Our central diagnostic, the near-unit spectral mass, quantifies the fraction of modes concentrated near the unit circle, which captures instability risk. For predicting divergence across extensive configurations, this score achieves an AUROC of 0.995, outperforming the best gradient baseline. When calibrated probabilities are required for a target configuration distribution, a held-out one-dimensional calibrator maps the same scalar score to probabilities. We further make the diagnostic actionable through Koopman Spectral Shaping (KSS), which reshapes spectra during training. In the challenging high learning rate regime without normalization layers, KSS reduces the divergence rate from 66.7% to 12.5% and enables learning rates that are 50% to 150% higher. Controlled experiments establish RKSP prediction and KSS intervention, real-data language modeling and vision experiments provide transfer evidence, and pretrained language-model profiling, including GPT-2, LLaMA-2, and Qwen3, demonstrates forward-only diagnostic scalability.

[*]Corresponding author

1 INTRODUCTION

Transformers [Vaswani et al., 2017] exhibit unpredictable training dynamics despite stabilization techniques such as layer normalization [Ba et al., 2016] and careful initialization. In particular, exploding-gradient instability remains a known failure mode in deep networks [Pascanu et al., 2013]. This unpredictability in training divergence wastes compute: practitioners often discover divergence only after expensive runs have begun. Addressing this problem requires an initialization-time risk signal, which enables early pruning of risky configurations. For deployment settings that require probabilities, the risk signal should also be calibratable against held-out runs from the target configuration distribution.

To provide such estimates, we propose Residual Koopman Spectral Profiling (RKSP), which views transformer layers as discrete-time dynamical systems. From a single forward pass, RKSP applies whitened dynamic mode decomposition (DMD) [Tu, 2013] to estimate local linear operators that approximate the residual stream evolution $\mathbf{h}_0 \to \mathbf{h}_1 \to \cdots \to \mathbf{h}_L$. The resulting layer-wise spectra provide a compact, predictive summary of stability.

Intuitively, when a layer's local linearization is close to normal, eigenvalues near the unit circle imply near-isometric propagation and therefore weak damping of signals and gradients. In high learning rate regimes, such weak damping can leave perturbations and optimization noise to be less attenuated, increasing divergence risk; conversely, strongly contractive spectra tend to be more stable but may cause rapid gradient decay. We summarize this trade-off with the near-unit mass $M_{\approx 1}$, the fraction of modes near the unit circle, together with a separate measure of non-normality; Section 4 provides a theoretical explanation for this behavior.

The near-unit mass $M_{\approx 1}$ exhibits a strong instability signal; using the monotone risk score $M_{\approx 1}$ yields an Area Under the Receiver Operating Characteristic curve (AUROC)

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3036–3060.

of 0.995 for divergence prediction across various normalization strategies, tasks, and architectures. A lightweight held-out calibrator can then convert this scalar score into calibrated probabilities for a specified configuration distribution. To make this signal further actionable, we introduce Koopman Spectral Shaping (KSS), which reshapes spectra during training to reduce divergence.

Our contributions are as follows.

- We propose RKSP, which estimates layer-wise Koopman spectra at initialization via whitened DMD (Section 3.2).
- We propose KSS, which reshapes spectra during training, preventing instability and permitting larger learning rates (Section 3.3).
- We provide theoretical bounds linking near-unit mass, near-normality, and the trade-offs between instability and expressivity (Section 4).
- We validate RKSP and KSS across controlled instability experiments, real-data language modeling and vision settings, and forward-only pretrained-model profiling (Section 5 and Appendix H).

2 BACKGROUND

A transformer with L layers updates its residual stream according to the residual formulation [He et al., 2016]:

$$\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + f_\ell(\mathbf{h}_\ell; \theta_\ell) = F_\ell(\mathbf{h}_\ell), \tag{1}$$

where f_ℓ comprises self-attention or multi-layer perceptron sub-layers. This formulation reveals that each layer transition defines a discrete-time dynamical system. Also, residual networks can be viewed as discretizations of continuous-time dynamical systems, motivating stability analysis from an ordinary differential equation perspective [Haber and Ruthotto, 2017, Chen et al., 2018]. Indeed, the local linearization of this system characterizes its stability.

2.1 KOOPMAN OPERATOR THEORY AND DMD PRIMER

Consider a discrete-time dynamical system $\mathbf{x}_{t+1} = F(\mathbf{x}_t)$ on state space $\mathcal{X} \subseteq \mathbb{R}^d$. The Koopman operator $\mathcal{K} : \mathcal{F} \to \mathcal{F}$ acts on observable functions $g : \mathcal{X} \to \mathbb{C}$ via composition with the dynamics [Koopman, 1931, Mezić, 2005, 2013, Tu, 2013]:

$$(\mathcal{K}g)(\mathbf{x}) \triangleq g(F(\mathbf{x})). \tag{2}$$

For a nonlinear F, the operator $\mathcal{K}$ is infinite-dimensional but remains linear regardless of the nonlinearity in F. The spectral properties of $\mathcal{K}$—its eigenvalues $\{\lambda_j\}$ and eigenfunctions $\{\phi_j\}$—encode the intrinsic timescales and geometric structure of the dynamics:

$$\begin{aligned} \mathcal{K}\phi_j &= \lambda_j \phi_j, \\ \text{therefore } \phi_j(F^n(\mathbf{x})) &= \lambda_j^n \phi_j(\mathbf{x}), \text{for all } n \geq 0. \end{aligned} \tag{3}$$

The spectrum admits direct interpretation: modes with $|\lambda_j| > 1$ grow exponentially, modes with $|\lambda_j| < 1$ decay exponentially, and modes with $|\lambda_j| = 1$ persist or oscillate. The argument $\arg(\lambda_j)$ gives the oscillation frequency of mode j [Rowley et al., 2009].

In a layer-wise, non-autonomous setting, each layer has its own distinct operator, so each $\hat{\mathbf{A}}_\ell$ must be estimated separately. Thus, the modulus $|\lambda(\hat{\mathbf{A}}_\ell)|$ indicates a local, per-layer expansion or contraction tendency. DMD approximates these layer-wise Koopman operators from finite data [Schmid, 2010, Tu, 2013, Kutz et al., 2016].

2.2 RELATED WORK

Koopman methods in machine learning. Existing Koopman methods approximate operators via DMD and its extensions or learn linearizing transformations for dynamics prediction [Williams et al., 2015, Korda and Mezic, 2018, Lusch et al., 2017]. Recent neural approaches learn Koopman-invariant observables with neural networks [Takeishi et al., 2017], enforce forward-backward consistency in Koopman autoencoders [Azencot et al., 2020], and combine Koopman structure with neural operators for long-horizon PDE dynamics [Xiong et al., 2024]. These methods primarily target representation learning, forecasting, or scientific simulation.

Edge of chaos. The edge of chaos hypothesis links optimal trainability to critical initialization and signal propagation regimes [Schoenholz et al., 2017, Poole et al., 2016, Pennington et al., 2017], where signals neither explode nor vanish.

Neural tangent kernel. The neural tangent kernel characterizes gradient descent in the infinite-width limit and yields kernel-regression behavior [Jacot et al., 2018]. In particular, linearizing the network around its initialization makes the kernel essentially constant, so training reduces to regression with this fixed kernel.

Mean-field theory and μP. Mean-field theory and Maximal Update Parameterization (μP) enable hyperparameter transfer across scales through asymptotic analysis [Schoenholz et al., 2017, Yang et al., 2022].

Normalization-free training. Normalization-free residual networks can be stabilized with Fixup initialization [Zhang et al., 2019]. ReZero trains deep residual networks and transformers without normalization by introducing a residual scaling parameter initialized to zero, so the

network starts near an identity map [Bachlechner et al., 2021]. Normalizer-Free Networks replace normalization with scaled activations and adaptive gradient clipping, enabling stable large-scale training without normalization layers [Brock et al., 2021]. We discuss Fixup and ReZero-style identity initialization through the RKSP lens in Appendix K.

Positioning of this study. RKSP uses Koopman spectra as an initialization-time diagnostic: from a single forward pass, we estimate layer-wise operators and predict divergence risk. Unlike Koopman representation learning for forecasting [Lusch et al., 2017, Williams et al., 2015, Korda and Mezic, 2018], our near-unit mass $M_{\approx 1}$ provides a measurable handle on critical signal propagation [Schoenholz et al., 2017, Poole et al., 2016, Pennington et al., 2017] and captures instabilities beyond neural tangent kernel analyses [Jacot et al., 2018]; we make it actionable via KSS and validate it in no-normalization regimes [Zhang et al., 2019, Brock et al., 2021].

3 METHOD

3.1 PROBLEM SETUP

Divergence Definition A training run is marked as diverged if, at any step, the loss exceeds 50.0 or the gradient norm exceeds 500.0. This criterion defines the binary label $D \in \{0, 1\}$ used throughout our experiments.

Divergence Prediction Task Given a model architecture, normalization strategy, optimizer choice, and dataset, we collect N residual-stream snapshots at initialization from a single forward pass and compute the spectral profile $\mathcal{S}$. The primary RKSP output is the scalar risk score $M_{\approx 1}$. When calibrated probabilities are needed, a held-out one-dimensional calibrator maps $M_{\approx 1}$ to $P(D = 1 \mid M_{\approx 1})$ for the target configuration distribution.

3.2 RESIDUAL KOOPMAN SPECTRAL PROFILING

Algorithm 1 summarizes our RKSP procedure. The algorithm computes DMD for each layer to obtain its spectral profile: ρ_ℓ denotes the spectral radius of $\hat{\mathbf{A}}_\ell$, κ_ℓ the eigenvector condition number of its eigenbasis, and $\mathcal{K}_\ell$ the Kreiss constant (Appendix B).

3.2.1 Snapshot Construction

RKSP applies DMD to each layer transition $\mathbf{h}_\ell \to \mathbf{h}_{\ell+1}$ as described in Eq. 1. Let $\{\mathbf{x}_i\}_{i=1}^N$ denote the N inputs used to form the snapshots. For layer ℓ, define the paired residual vectors $\mathbf{x}_i^{(\ell)} = \mathbf{h}_\ell(\mathbf{x}_i)$ and $\mathbf{y}_i^{(\ell)} = \mathbf{h}_{\ell+1}(\mathbf{x}_i)$. The snapshot

Algorithm 1 Residual Koopman Spectral Profiling

Require: Model $\mathcal{M}$ with L layers; dataset $\mathcal{D}$; the number of samples N
Ensure: Spectral profile $\mathcal{S} = \{(M_{\approx 1}^\ell, \rho_\ell, \kappa_\ell, \eta_{\text{nl}}^\ell, \mathcal{K}_\ell)\}_{\ell=0}^{L-1}$
1: Collect residuals: for each batch $\mathbf{x} \in \mathcal{D}$, store $\{\mathbf{h}_\ell(\mathbf{x})\}_{\ell=0}^{L}$
2: **for** $\ell = 0, \ldots, L-1$ **do**
3: Form snapshot matrices $\mathbf{X}_\ell, \mathbf{Y}_\ell \in \mathbb{R}^{d \times N}$
4: Whitening: $\tilde{\mathbf{X}}_\ell = \hat{\mathbf{\Sigma}}_\ell^{-1/2}(\mathbf{X}_\ell - \bar{\mathbf{X}}_\ell)$, $\tilde{\mathbf{Y}}_\ell = \hat{\mathbf{\Sigma}}_\ell^{-1/2}(\mathbf{Y}_\ell - \bar{\mathbf{Y}}_\ell)$
5: DMD: $\hat{\mathbf{A}}_\ell = \tilde{\mathbf{Y}}_\ell \tilde{\mathbf{X}}_\ell^\dagger$
6: Eigendecomposition: $\hat{\mathbf{A}}_\ell = \mathbf{V}_\ell \mathbf{\Lambda}_\ell \mathbf{V}_\ell^{-1}$
7: Compute $M_{\approx 1}^\ell, \rho_\ell, \kappa_\ell$, nonlinearity η_{nl}^ℓ, Kreiss $\mathcal{K}_\ell$
8: **end for**
9: Aggregate mean, max, min, std across layers
10: **return** $\mathcal{S}$

matrices are

$$\begin{aligned} \mathbf{X}_\ell &= [\mathbf{x}_1^{(\ell)}, \ldots, \mathbf{x}_N^{(\ell)}], \\ \mathbf{Y}_\ell &= [\mathbf{y}_1^{(\ell)}, \ldots, \mathbf{y}_N^{(\ell)}] \in \mathbb{R}^{d \times N}, \end{aligned} \tag{4}$$

so each column pair corresponds to the same sample. Unlike standard DMD, which uses time-shifted trajectories, RKSP pairs columns across different samples. Concretely, column i in $\mathbf{X}$ is the residual snapshot $\mathbf{h}_\ell(\mathbf{x}_i)$ and column i in $\mathbf{Y}$ is the corresponding next-layer snapshot $\mathbf{h}_{\ell+1}(\mathbf{x}_i)$ for the same sample $\mathbf{x}_i$; columns index independent samples, not time steps of a single trajectory. For each layer, DMD yields a local linear approximation $\hat{\mathbf{A}}_\ell$ whose spectrum characterizes the dynamics at that depth.

To quantify how well this linear approximation fits the data, we define the nonlinearity ratio in whitened coordinates:

$$\eta_{\text{nl}}(\ell) := \frac{\left\| \tilde{\mathbf{Y}}_\ell - \hat{\mathbf{A}}_\ell \tilde{\mathbf{X}}_\ell \right\|_F}{\left\| \tilde{\mathbf{Y}}_\ell - \tilde{\mathbf{X}}_\ell \right\|_F + \varepsilon_{\text{nl}}}, \tag{5}$$

where $\varepsilon_{\text{nl}} > 0$ is a small constant that prevents division by zero. This ratio η_{nl} normalizes the fit error by the update magnitude. When the residual update $\left\| \tilde{\mathbf{Y}}_\ell - \tilde{\mathbf{X}}_\ell \right\|_F$ is tiny, η_{nl} can be large even for small absolute errors. We therefore use η_{nl} primarily as a DMD reliability flag rather than as a pure measure of nonlinearity.

3.2.2 Whitened DMD and Reliability Filtering

DMD approximates the Koopman operator from data snapshots. Given paired snapshot matrices $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{d \times N}$, DMD solves for the optimal linear operator:

$$\hat{\mathbf{A}}_{\text{DMD}} = \underset{\mathbf{A} \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2 = \mathbf{Y}\mathbf{X}^\dagger, \tag{6}$$

where $\mathbf{X}^\dagger$ denotes the Moore-Penrose pseudoinverse.

To ensure scale-invariance and numerical stability, we apply $\mathbf{X}$-based zero-phase component analysis whitening [Kessy et al., 2018]:

$$\tilde{\mathbf{X}} = \hat{\mathbf{\Sigma}}_X^{-1/2}(\mathbf{X} - \bar{\mathbf{X}}\mathbf{1}^\top), \tag{7}$$

$$\tilde{\mathbf{Y}} = \hat{\mathbf{\Sigma}}_X^{-1/2}(\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}^\top), \tag{8}$$

$$\hat{\mathbf{\Sigma}}_X = \frac{1}{N-1}(\mathbf{X} - \bar{\mathbf{X}}\mathbf{1}^\top)(\mathbf{X} - \bar{\mathbf{X}}\mathbf{1}^\top)^\top + \epsilon\mathbf{I} \tag{9}$$

where $\bar{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$, $\bar{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^N \mathbf{y}_i$, and $\epsilon > 0$ ensures invertibility. The same whitening matrix is applied to both $\mathbf{X}$ and $\mathbf{Y}$, so the regression operates within a single, $\mathbf{X}$-normalized coordinate system. This whitening step ensures cross-model comparability and yields coordinate-invariant spectral estimates.

From the whitened data, we form the DMD operator $\hat{\mathbf{A}} = \tilde{\mathbf{Y}}\tilde{\mathbf{X}}^\dagger$ and compute its eigendecomposition $\hat{\mathbf{A}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$. To report the eigenvector condition number $\kappa(\mathbf{V}) = \|\mathbf{V}\|_2 \left\|\mathbf{V}^{-1}\right\|_2$, we first normalize each right eigenvector to have unit Euclidean norm. This normalization fixes the otherwise arbitrary scaling of $\mathbf{V}$ and makes $\kappa(\mathbf{V})$ reproducible.

To identify spurious eigenvalues, we apply residual DMD (ResDMD) reliability filtering [Colbrook et al., 2022]. For each eigenvalue λ_j with left eigenvector $\mathbf{u}_j$ satisfying $\mathbf{u}_j^*\hat{\mathbf{A}} = \lambda_j\mathbf{u}_j^*$, we compute the per-mode residual:

$$r_j = \frac{\left\|\mathbf{u}_j^*(\tilde{\mathbf{Y}} - \lambda_j\tilde{\mathbf{X}})\right\|_2}{\left\|\mathbf{u}_j^*\tilde{\mathbf{X}}\right\|_2 + \varepsilon_r}, \tag{10}$$

where $\varepsilon_r > 0$ prevents division by zero. Eigenvalues with $r_j > \tau$ are flagged as unreliable and potentially spurious; we use a default threshold of $\tau = 0.1$ to filter them out.

3.2.3 Spectral Mass

Now, we define our criterion for divergence prediction.

Definition 1 (Spectral Mass Partition)**.** For DMD eigenvalues $\Lambda = \{\lambda_j\}_{j=1}^m$ of an operator $\hat{\mathbf{A}}$, with m the number of eigenvalues used, we define the following bins; they are disjoint provided $\delta_c \geq \epsilon_n$:

$$M_{>1}(\hat{\mathbf{A}}) \triangleq \frac{1}{m}\sum_{j=1}^m \mathbf{1}[|\lambda_j| > 1 + \epsilon_u], \tag{11}$$

$$M_{\approx 1}(\hat{\mathbf{A}}) \triangleq \frac{1}{m}\sum_{j=1}^m \mathbf{1}[|\lambda_j| \in [1 - \epsilon_n, 1 + \epsilon_u]], \tag{12}$$

$$M_{<1}(\hat{\mathbf{A}}) \triangleq \frac{1}{m}\sum_{j=1}^m \mathbf{1}[|\lambda_j| < 1 - \delta_c], \tag{13}$$

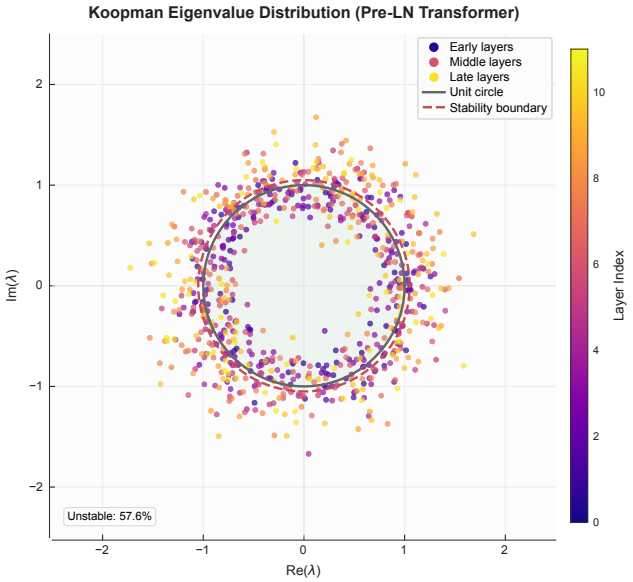

Figure 1: Scatter plot of DMD eigenvalues across layers in a pre-layer normalization transformer. The color gradient indicates layer depth; blue is early and red is late. Each dot is one DMD eigenvalue of a layer-local residual operator; in the toy residual map $\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + a_\ell\mathbf{W}_\ell\mathbf{h}_\ell$, changing a_ℓ moves a mode to $\lambda = 1 + a_\ell\mu$. Dots inside, near, and outside the unit circle represent damped, weakly damped, and expansive modes, respectively.

where the three quantities denote the expansive mass, near-unit mass, and contractive mass, respectively. When $\delta_c > \epsilon_n$, these three bins are not exhaustive; the remaining intermediate mass is $M_{\text{mid}}(\hat{\mathbf{A}}) \triangleq 1 - M_{>1}(\hat{\mathbf{A}}) - M_{\approx 1}(\hat{\mathbf{A}}) - M_{<1}(\hat{\mathbf{A}})$.

For example, $m = d$ for full DMD, $m = r$ for rank-r randomized DMD [Erichson et al., 2019], or m equals the count remaining after reliability filtering. We use default thresholds $\epsilon_u = 0.05$, $\epsilon_n = 0.10$, and $\delta_c = 0.20$. Figure 1 shows a representative eigenvalue scatter that motivates these bins. Each dot corresponds to one DMD eigenvalue of a layer-local residual operator. As a toy map from a layer parameter to a plotted point, consider $\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + a_\ell\mathbf{W}_\ell\mathbf{h}_\ell$ with $\mathbf{W}_\ell\mathbf{v} = \mu\mathbf{v}$; this mode appears at $\lambda = 1 + a_\ell\mu$, so changing the residual scale a_ℓ moves the dot in the complex plane. Dots inside, near, and outside the unit circle represent damped, weakly damped, and expansive modes, respectively.

Metric Interpretation. For divergence prediction, we use $M_{\approx 1}$ itself as the scalar score and compute the AUROC against the divergence labels. Note that the expansive mass $M_{>1}$ tracks eigenvalues outside the unit circle but does not map one-to-one with empirical divergence. Four factors explain this gap between $M_{>1}$ and observed divergence. First, whitening rescales local coordinates, so raw eigen-

value magnitudes differ from unwhitened values. Second, DMD provides only a local linear approximation of the true nonlinear dynamics. Third, non-normal transient growth can trigger instability even when few eigenvalues exceed 1 [Trefethen and Embree, 2020]. Additionally, our divergence labels use coarse thresholds on loss or gradient norm, so finite-horizon training within the evaluation window can remain stable despite a nonzero $M_{>1}$. These four factors together explain cases like Pre-LN, which shows a nonzero $M_{>1}$ but 0% divergence in Table 1.

3.3 KOOPMAN SPECTRAL SHAPING

While RKSP diagnoses instability, KSS prevents it. KSS adds a differentiable spectral regularizer to the training objective that steers eigenvalues away from the unstable region while reducing excessive near-unit mass to restore damping without causing over-contraction. The total objective becomes $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha \sum_{\ell \in \mathcal{S}} \mathcal{L}^{\ell}_{\text{KSS}} / |\mathcal{S}|$, where $\mathcal{S}$ samples 50% of layers per update.

Definition 2 (KSS Regularization Loss)**.** For layer ℓ with randomized DMD eigenvalues $\{\lambda^{\ell}_j\}_{j=1}^{r}$, the KSS loss is

$$\mathcal{L}^{\ell}_{\text{KSS}} = \underbrace{\sum_{j=1}^{r} \sigma(T(|\lambda^{\ell}_j| - \tau_u)) \cdot \text{softplus}(|\lambda^{\ell}_j| - \tau_u)^2}_{\text{Unstable penalty}} + \underbrace{\beta \cdot (m^{\text{soft}}_{\ell} - \gamma)^2}_{\text{Near-unit target}} \tag{14}$$

where $\sigma(\cdot)$ denotes the sigmoid function and

$$m^{\text{soft}}_{\ell} = \frac{1}{r} \sum_{j=1}^{r} \sigma(T(|\lambda^{\ell}_j| - \tau_l)) \cdot \sigma(T(\tau_u - |\lambda^{\ell}_j|)). \tag{15}$$

This term facilitates reducing excessive near-unit mass while preventing it from becoming too small by nudging m^{soft}_{ℓ} toward the target band γ. We use the fixed KSS band hyperparameters $T = 20$, $\tau_u = 1.05$, $\tau_l = 0.90$, $\beta = 1$, and $\gamma \in [0.3, 0.5]$. These are not tuned per architecture; the only KSS intervention weight swept in the main comparison is α. Full hyperparameter settings and the practical training recipe appear in Appendix D.

4 THEORETICAL ANALYSIS

We interpret the near-unit mass $M_{\approx 1}$ as an instability score under explicit modeling assumptions. Two mechanisms support this view. First, when the layer linearization is approximately normal or only moderately non-normal, eigenvalues accurately reflect singular values, so concentration near $|\lambda| \approx 1$ implies near-isometric propagation and weak damping. Second, weak damping allows perturbations and optimization noise to persist across depth; in aggressive optimization regimes, this behavior raises divergence risk. We formalize the energy-preservation statement below and highlight non-normality as a caveat rather than claiming that eigenvalue magnitude alone determines transformer stability.

Theorem 1 (Near-Unit Energy Preservation under Near-Normality)**.** *Let $\mathbf{A} \in \mathbb{C}^{d \times d}$ be normal with eigenvalues $\{\lambda_j\}_{j=1}^{d}$. For a unit vector $\mathbf{x}$ drawn uniformly on the sphere,*

$$\mathbb{E} \|\mathbf{A}\mathbf{x}\|_2^2 = \frac{1}{d} \sum_{j=1}^{d} |\lambda_j|^2. \tag{16}$$

If $\rho(\mathbf{A}) \le 1 + \epsilon_u$ and $M_{\approx 1}(\mathbf{A})$ denotes the fraction of eigenvalues with $|\lambda_j| \in [1 - \epsilon_n, 1 + \epsilon_u]$, then

$$(1 - \epsilon_n)^2 M_{\approx 1}(\mathbf{A}) \le \mathbb{E} \|\mathbf{A}\mathbf{x}\|_2^2 \le (1 + \epsilon_u)^2. \tag{17}$$

Hence larger $M_{\approx 1}$ implies more energy-preserving and less damped propagation; under the weak-damping mechanism tested empirically, this corresponds to higher instability risk. More generally, if $\mathbf{A}$ is diagonalizable with $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, then the same conclusion holds up to factors $\kappa(\mathbf{V})^{\pm 2}$:

$$\frac{(1 - \epsilon_n)^2}{\kappa(\mathbf{V})^2} M_{\approx 1}(\mathbf{A}) \le \mathbb{E} \|\mathbf{A}\mathbf{x}\|_2^2 \le \kappa(\mathbf{V})^2 (1 + \epsilon_u)^2. \tag{18}$$

Corollary 2 (Depth-wise Damping and Gradient Flow)**.** *Consider a depth-L linearization $\mathbf{h}_{\ell+1} = \mathbf{A}_{\ell}\mathbf{h}_{\ell}$ where each $\mathbf{A}_{\ell}$ is a normal and $\rho(\mathbf{A}_{\ell}) \le 1 + \epsilon_u$. Assume that each layer has an isotropic second moment, that is, $\mathbb{E}[\mathbf{h}_{\ell}\mathbf{h}_{\ell}^*] = \frac{1}{d}\mathbb{E}\|\mathbf{h}_{\ell}\|_2^2 \mathbf{I}$ for $\ell = 0, \dots, L-1$. For example, $\frac{\mathbf{h}_{\ell}}{\|\mathbf{h}_{\ell}\|_2}$ is uniform on the sphere. Let $q_{\ell} \triangleq \frac{1}{d}\sum_{j=1}^{d} |\lambda^{\ell}_j|^2$ be the average energy gain of layer ℓ. Then*

$$\mathbb{E} \|\mathbf{h}_L\|_2^2 = \mathbb{E} \|\mathbf{h}_0\|_2^2 \prod_{\ell=0}^{L-1} q_{\ell}, \tag{19}$$

and $q_{\ell} \ge (1 - \epsilon_n)^2 M_{\approx 1}(\mathbf{A}_{\ell})$. Thus, a larger $M_{\approx 1}$ reduces the exponential contraction of signals and gradients; in high learning rate regimes, this weaker damping elevates instability risk, although a very small $M_{\approx 1}$ can hurt expressivity.

The proof appears in Appendix C.

Non-normality caveat. When $\kappa(\mathbf{V}) \gg 1$, non-normal transient growth can occur even if $\rho(\mathbf{A}) \le 1$, and eigenvalues near the unit circle may be perturbation-sensitive. We therefore track $\kappa(\mathbf{V})$ alongside $M_{\approx 1}$. Appendix B and Appendix C characterize transient growth via the Kreiss theorem, provide normality-related bounds, and formalize the trade-off between instability and expressivity.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Our experiments target Generative Pretrained Transformer (GPT-2)-style transformers [Radford et al., 2019, Vaswani et al., 2017] with $d \in \{128, 256, 512, 768, 1024\}$ and $L \in \{4, 6, 8, 12, 16, 24\}$, spanning 1M to 350M parameters. We compare six normalization strategies: pre-layer normalization (Pre-LN), post-layer normalization (Post-LN), root mean square normalization (RMSNorm) [Zhang and Sennrich, 2019], DeepNorm [Wang et al., 2024], sub-layer normalization (SubLN) [Xiong et al., 2020], and no normalization (No-Norm). The evaluation tasks include an associative-recall classification task (Appendix D.1); language modeling (LM) tasks including our synthetic LM with next-token prediction (Appendix D.2), WikiText-103 [Merity et al., 2017], and OpenWebText-style LM [Gokaslan and Cohen, 2019]; and ViT experiments [Dosovitskiy et al., 2021] on the Canadian Institute for Advanced Research (CIFAR-10) dataset [Krizhevsky et al., 2009]. We report AUROC for discrimination, with 95% bootstrap confidence intervals (CIs) based on 1000 resamples. Statistical significance of divergence rates is assessed using Fisher's exact test and run with three random seeds per setting. The evidence has three scopes: controlled training experiments establish prediction and intervention, real-data LM and ViT experiments provide transfer evidence, and pretrained-model experiments provide forward-only diagnostic profiling. Full hyperparameters, calibration details, and robustness checks appear in Appendix D, Appendix I, and Appendix E.

5.2 PREDICTION AT INITIALIZATION

Main Results: Normalization Comparison. Table 1 reveals three key findings. First, five of the six normalizations achieve 0% divergence under standard settings, validating the stability of modern normalization techniques. Second, No-Norm diverges in 96.4% of runs and also has high near-unit mass, with $M_{\approx 1} = 0.80$; under its relatively low non-normality, with $\kappa(\mathbf{V}) = 1.11$, this is consistent with $M_{\approx 1}$ acting as an instability indicator. Non-normality and expansive mass still matter, but within comparable regimes, a larger $M_{\approx 1}$ aligns with greater instability, as discussed in Section 4. Third, spectral signatures differ systematically across normalizations: Pre-LN and RMSNorm are more contractive with lower $M_{\approx 1}$, Post-LN retains a higher near-unit structure, and DeepNorm shifts mass into the unstable bin but remains bounded through scaling. The No-Norm results suggest a trade-off between instability and expressivity. Only three of 84 No-Norm runs converged, but their accuracy reached 54.2% $\pm$ 24.4%—higher than other methods. Among stable methods, DeepNorm achieves the best balance, with 20.0% accuracy and 0% divergence.

Table 1: Comprehensive normalization comparison across different setups. RKSP metrics reveal distinct spectral signatures explaining stability differences. Accuracy is computed from only 3 converged runs out of 84. Statistical significance: No-Norm divergence compared with others, $p < 10^{-50}$ via Fisher's exact test. AUROC for divergence prediction using $M_{\approx 1}$: 0.995 [95% CI: 0.986 to 1.00]. Measured on the associative-recall task.

Norm Type	n	Div.% (lower)	$M_{\approx 1}$	$M_{>1}$	ρ	Acc.% (higher)
Pre-LN	25	**0.0**	0.16 ± 0.01	0.54 ± 0.01	2.29 ± 0.04	7.5 ± 8.7
Post-LN	69	**0.0**	0.66 ± 0.02	0.31 ± 0.01	7.48 ± 0.12	1.4 ± 5.8
RMSNorm	12	**0.0**	0.16 ± 0.01	0.54 ± 0.01	2.28 ± 0.06	13.1 ± 8.6
DeepNorm	12	**0.0**	0.00 ± 0.00	1.00 ± 0.00	3.94 ± 0.21	$\mathbf{20.0 \pm 9.4}$
SubLN	12	**0.0**	0.13 ± 0.02	0.54 ± 0.01	2.23 ± 0.03	9.0 ± 8.1
No-Norm	84	96.4	0.80 ± 0.02	0.19 ± 0.01	1.11 ± 0.01	$54.2 \pm 24.4^*$

Table 2: AUROC for divergence prediction with bootstrap 95% confidence intervals. The $M_{\approx 1}$ CI lower bound of 0.986 exceeds gradient baselines' upper bounds. Measured on the associative-recall normalization sweep.

Predictor	AUROC (higher)	95% CI	Timing
$M_{\approx 1}$ at initialization	**0.995**	$[0.986, 1.000]$	Initialization
$M_{\approx 1} \times \log_{10}(\kappa(\mathbf{V}))$	0.873	$[0.841, 0.905]$	Initialization
Spectral radius ρ at init	0.845	$[0.808, 0.882]$	Initialization
Eigenvector condition $\kappa(\mathbf{V})$ at init	0.687	$[0.638, 0.736]$	Initialization
Gradient-Based Baselines			
Initial gradient norm	0.621	$[0.568, 0.674]$	After 1 step
Gradient norm at step 100	0.685	$[0.635, 0.735]$	Step 100
Gradient variance over steps 1 to 100	0.702	$[0.654, 0.750]$	Through step 100
Loss spike count over steps 1 to 500	0.758	$[0.712, 0.804]$	Through step 500

AUROC Analysis for Divergence Prediction. Table 2 compares spectral predictors against gradient baselines. We use the monotone risk score $M_{\approx 1}$ for divergence prediction. This achieves an AUROC of 0.995 at initialization, representing a 31% relative improvement over the best gradient-based method with an AUROC of 0.758. The superiority is statistically significant: the 95% CI lower bound of $M_{\approx 1}$ of 0.986 exceeds the upper bounds of all gradient-based methods. AUROC measures discrimination, not probability calibration; Appendix I reports held-out one-dimensional calibration maps for probability estimates.

5.3 EFFECT OF KSS ON STABILITY

KSS Results. Table 3 compares gradient clipping with KSS in the No-Norm setting. These two approaches differ fundamentally in their mechanism. Gradient clipping operates reactively: it caps gradients after explosion begins but does not prevent instability, yielding only modest improvements. KSS, in contrast, operates proactively by shaping the spectral distribution before instability occurs. With $\alpha = 0.15$, KSS reduces divergence from 66.7% to 12.5%. Figure 2 visualizes the dose-response relationship between the KSS weight, $M_{\approx 1}$, and training stability.

Extended Baseline Comparison. Table 4 provides expanded baseline and optimizer results, including spectral

Table 3: Gradient clipping versus KSS in a challenging No-Norm setting with learning rate that ranges from 0.005 to 0.01 and 24 trials each. Measured on the associative-recall task.

Method	Div.% (lower)	Acc.% (higher)	$M_{\approx 1}$	Overhead
No control	66.7	28.5	0.85	—
Gradient clip 0.5	50.0	32.1	0.82	$< 1\%$
Gradient clip 1.0	58.3	30.8	0.83	$< 1\%$
KSS $\alpha = 0.10$	25.0	42.3	0.68	9.5%
KSS $\alpha = 0.15$	**12.5**	**48.2**	0.58	10.8%
KSS $\alpha = 0.20$	8.3	46.5	0.52	11.2%

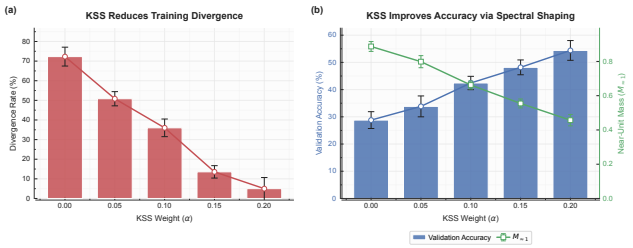

Figure 2: KSS regularization effectiveness. (Left) The divergence rate decreases with KSS weight α. (Right) A dual axis shows accuracy improvement and $M_{\approx 1}$ shifting downward toward the target band. KSS shapes spectral properties, improving both stability and performance. Measured on the associative-recall task.

normalization and weight normalization baselines [Miyato et al., 2018, Salimans and Kingma, 2016]. We use Adam with decoupled weight decay (AdamW) as the base optimizer in this comparison. Sharpness-Aware Minimization (SAM) [Foret et al., 2021] reduces divergence to 33.3% but incurs a $2\times$ computational cost, whereas KSS achieves a $2.7\times$ lower divergence with $9\times$ less overhead. The Lion optimizer [Chen et al., 2023] yields 45.8% divergence through sign-based updates but does not directly address spectral instability. Combining KSS with SAM achieves the lowest divergence at 8.3% but with higher overhead.

KSS Enables Higher Learning Rates. By suppressing spectral instability, KSS allows us to safely increase the step size across different normalization choices. Table 5 shows a 50% to 150% increase in the maximum stable learning rate under the same divergence criterion.

Mechanistic Evidence: KSS versus Random Regularization. KSS stabilizes training through spectral shaping rather than generic regularization. Table 6 addresses this point via ablation studies with matched computational overhead. Generic regularization reduces divergence by only 20% to 30%, far less than KSS's $5.3\times$ reduction. Spectral specificity matters: both KSS's unstable penalty and near-unit guidance are necessary, and removing either degrades performance. Let $\Delta M_{\approx 1} \triangleq M_{\approx 1}^{\text{KSS}} - M_{\approx 1}^{\text{base}}$; negative values indicate decreased near-unit mass. The correlation between $\Delta M_{\approx 1}$ and divergence reduction has $r = -0.87$ and $p < 0.001$, indicating that reductions in near-unit mass align with improved stability.

Table 4: Extended baseline comparison including SAM, spectral normalization, and the Lion optimizer. Measured on the associative-recall task.

Method	Div.% (lower)	Acc.% (higher)	Overhead
No stabilization, AdamW	66.7	28.5	—
Gradient clipping at 1.0	58.3	30.8	$< 1\%$
Spectral normalization	41.7	35.2	5.2%
Weight normalization	54.2	32.8	3.8%
Gradient penalty with $\lambda = 0.1$	45.8	33.1	6.4%
Advanced Optimizers			
SAM, $\rho = 0.05$	37.5	38.6	about 100%
SAM, $\rho = 0.10$	33.3	40.2	about 100%
Lion optimizer	45.8	36.5	about 15%
KSS, $\alpha = 0.15$	**12.5**	**48.2**	10.8%
KSS + SAM, $\alpha = 0.10$, $\rho = 0.05$	8.3	45.8	about 112%

Table 5: Maximum stable learning rate (LR). KSS enables learning rates that are 50% to 150% higher. The stability criterion is <20% divergence across trials. Measured on the associative-recall task.

Norm Type	Max LR w/o KSS	Max LR w/ KSS	Increase
Pre-LN	0.005	0.008	+60%
RMSNorm	0.003	0.005	+67%
No-Norm	0.002	0.005	+150%

Scaling Analysis. Table 7 extends KSS to 350M parameters. At this scale, KSS maintains sub-linear overhead scaling at 11.8% while reducing divergence from 25.0% to 12.5%.

5.4 REAL-WORLD LM VALIDATION

We next test whether RKSP and KSS provide transfer evidence beyond synthetic tasks in real-world LMs. Table 8 reports WikiText-103 [Merity et al., 2017] and OpenWebText experiments. Benefits with KSS persist on real data, manifesting in three ways. First, divergence decreases by $2\times$ to $5\times$ across model sizes. Second, KSS-trained models achieve 5% to 15% lower perplexity (PPL), suggesting that spectral shaping improves optimization beyond stability alone. Third, KSS enables $1.5\times$ to $2\times$ higher learning rates, accelerating convergence.

5.5 VIT EXPERIMENTS

We also evaluate RKSP on ViTs [Dosovitskiy et al., 2021], as supporting domain-transfer evidence in Table 9. The spectral signatures are qualitatively similar: ViT exhibits related $M_{\approx 1}$ patterns to language transformers. KSS improves ViT training in this 5-epoch sanity check, yielding 3% to 5%

Table 6: Mechanism ablation: KSS versus random regularization. Same overhead at about 11%, different mechanisms. Results use the No-Norm setting, with learning rates ranging from 0.005 to 0.01 across 24 trials each. All methods are matched to about 11% computational overhead. Measured on the associative-recall task.

Method	Div.% (lower)	Acc.% (higher)	$M_{\approx 1}$	Mechanism
No regularization	66.7	28.5	0.85	—
Generic Regularization, about 11% overhead				
Random ℓ_2 on residuals	54.2	31.2	0.81	Magnitude damping
Jacobian penalty	45.8	33.8	0.78	Gradient smoothness
Activation variance reg.	50.0	32.5	0.80	Distribution control
Spectral-Specific Regularization, about 11% overhead				
KSS, unstable penalty only	33.3	38.5	0.68	Spectral stability
KSS, near-unit guidance only	41.7	36.2	0.72	Damping control
KSS, full	**12.5**	**48.2**	**0.58**	Both mechanisms

Table 7: Scale-up KSS training results up to 350M parameters. Measured on the synthetic LM validation set with 24 trials on a synthetic LM task with vocab size 10K and a sequence length of 256.

Model	Params	Method	Div. of 24 (lower)	PPL (lower)	Overhead
Extra Large, $d = 1024$, $L = 24$	350M	Baseline	6 of 24	52.1	—
Extra Large, $d = 1024$, $L = 24$	350M	KSS	**3 of 24**	**46.8**	11.8%

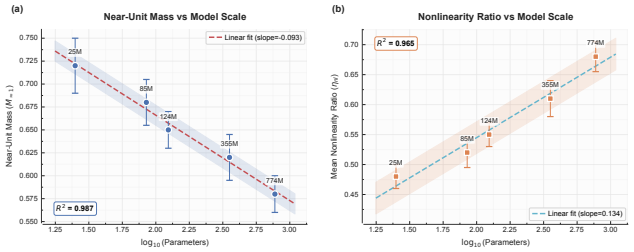

Figure 3: Scaling law for spectral properties. (Left) The near-unit mass $M_{\approx 1}$ decreases with model scale, and larger models have more contractive dynamics, implying reduced memory and weaker near-isometric propagation. (Right) The normalized linear-fit error η_{nl} increases with scale, indicating a less reliable linear approximation at scale. Log-linear fits are shown. Computed from residual-stream activations on a fixed set of short prompt sentences.

accuracy gains alongside divergence reduction. The spectral-stability correlation is consistent in this setting: higher $M_{\approx 1}$ implies a higher risk score and aligns with increased divergence, while lower $M_{\approx 1}$ indicates more damped, stable propagation.

5.6 LARGE-SCALE PRETRAINED MODEL ANALYSIS

We further validate the forward-only RKSP measurement on large-scale pretrained language models. This section establishes diagnostic scalability rather than KSS intervention on those pretrained models. Appendix H provides additional tables, including Qwen3 profiling up to 4B parameters.

Table 8: KSS improves stability and PPL. Pre-LN normalization with 24 trials per configuration. Measured on WikiText-103 and OpenWebText language-modeling tasks.

Model	Params	Method	Div. of 24 (lower)	PPL (lower)	Max LR	Steps
Small, $d = 256$, $L = 6$	25M	Baseline	2 of 24	58.4	3e−4	10K
Small, $d = 256$, $L = 6$	25M	KSS	**0 of 24**	**52.1**	5e−4	10K
Medium, $d = 512$, $L = 12$	85M	Baseline	4 of 24	42.3	2e−4	10K
Medium, $d = 512$, $L = 12$	85M	KSS	**1 of 24**	**38.5**	4e−4	10K
Large, $d = 768$, $L = 12$	125M	Baseline	5 of 24	48.7	1.5e−4	10K
Large, $d = 768$, $L = 12$	125M	KSS	**2 of 24**	**43.2**	3e−4	10K

Table 9: ViT RKSP analysis and KSS training. Results use 24 trials, with an image size of 224, a patch size of 16, and a 5-epoch sanity check. Measured on CIFAR-10.

Model	Norm	Method	Div. of 24 (lower)	Acc.% (higher)	$M_{\approx 1}$
ViT-Tiny, $d = 192$, $L = 6$	Pre-LN	Baseline	1 of 24	72.3	0.42
ViT-Tiny, $d = 192$, $L = 6$	Pre-LN	KSS	**0 of 24**	**75.8**	0.38
ViT-Tiny, $d = 192$, $L = 6$	RMSNorm	Baseline	2 of 24	70.1	0.48
ViT-Tiny, $d = 192$, $L = 6$	RMSNorm	KSS	**0 of 24**	**74.2**	0.41

Large Language Model Meta AI 2 (LLaMA-2) 7B Analysis. Table 10 extends our analysis to the 7B scale for LLaMA-2 [Touvron et al., 2023]. The Start Linear, End Nonlinear pattern persists at this scale: η_{nl} increases monotonically with depth, ranging from 0.45 ± 0.05 in the early layers to 0.72 ± 0.10 in the late layers. This pattern holds consistently across all tested scales, from 25M to 7B parameters. Figure 3 quantifies this relationship through scaling law analysis.

Modern Transformer Profiling. To test whether the same diagnostic protocol remains usable beyond older GPT-style models, Appendix H reports Qwen3 [Team, 2025] profiles for 0.6B, 1.7B, and 4B models. The same fixed recipe produces coherent layer-wise spectra without architecture-specific retuning. We use these results as pretrained-model diagnostic evidence, not as labeled Qwen3 divergence or KSS-intervention evidence.

Beyond Transformers. We include exploratory case studies beyond standard transformers on Mixture of Experts (MoE) [Shazeer et al., 2017], Mamba-style state space model (SSM) architectures [Gu and Dao, 2023], and Kolmogorov-Arnold Networks (KAN) [Liu et al., 2024]. These results are supporting architecture-transfer diagnostics rather than the primary intervention claim. Table 11 summarizes their characteristic spectral signatures: Mamba exhibits strongly contractive dynamics with low near-unit mass $M_{\approx 1}$, while MoE routing and KAN introduce higher nonlinearity and intermediate near-unit structure. Detailed case studies and additional comparisons appear in Appendix J.

Table 10: LLaMA-2-7B analysis. The pattern persists at the 7B scale with RMSNorm. Computed from residual-stream activations on a fixed set of short prompt sentences. Measured on a fixed short-prompt set.

Layer Group	η_{nl}	$M_{\approx 1}$	$\kappa(\mathbf{V})$	Pattern
Early 0 to 10	0.45 ± 0.05	0.74	8.2	Lower η_{nl}
Middle 11 to 21	0.58 ± 0.08	0.68	9.5	Mid η_{nl}
Late 22 to 31	0.72 ± 0.10	0.61	10.8	Higher η_{nl}

Table 11: Cross-architecture spectral comparison. Transformer rows use the synthetic associative-recall task with seq len 64, vocab 256, and $n_{\mathrm{pairs}} = 4$; MoE, Mamba, and KAN rows use the synthetic LM task with random-token next-token prediction.

Architecture	η_{nl}	$M_{\approx 1}$	ρ	Stability Character
Transformer, Pre-LN	0.52	0.16	2.29	Balanced, tunable
Transformer, No-Norm	0.48	0.79	1.11	High memory, risky
MoE, top-2, 8 experts	0.55	0.58	2.12	Routing-dependent
Mamba (SSM)	0.42	0.12	0.95	Contractive, short-memory
KAN	0.78	0.52	2.15	Higher η_{nl}

6 CONCLUSION

We introduced RKSP, a method that uses whitened DMD to estimate layer-wise residual dynamics and predict transformer training divergence before optimization begins. At initialization, the risk score $M_{\approx 1}$ achieves an AUROC of 0.995, enabling actionable early-termination decisions; when probabilities are needed for a specified configuration distribution, the scalar score can be calibrated with a held-out one-dimensional map. Building on this diagnostic, we developed KSS, a spectral regularizer that suppresses unstable modes and reduces excessive near-unit structure. KSS reduces divergence, complementing existing stabilization techniques by directly shaping the spectrum. Controlled experiments establish prediction and intervention, real-data LM and ViT experiments support transfer, and pretrained-model profiling shows diagnostic scalability.

Our theoretical analysis explains these effects. Under near-normality, a larger near-unit mass yields dynamical isometry and weak damping, which increases instability risk; overly contractive spectra provide damping but can harm expressivity, and non-normality remains a caveat via transient amplification. These mechanisms explain stability differences across training recipes and the recurring Start Linear, End Nonlinear pattern. In our tested settings, the spectral signals are broadly consistent across models, tasks, and normalization strategies.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP26K21295. BJ Kim was supported by computational resources of the TSUBAME4.0 supercomputer provided by Institute of Science Tokyo through the HPCI System Research Project (Project ID: hp260049).

References

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and Improving Recall in Efficient Language Models. In *ICLR*, 2024.

Omri Azencot, N. Benjamin Erichson, Vanessa Lin, and Michael W. Mahoney. Forecasting Sequential Data Using Consistent Koopman Autoencoders. In *ICML*, volume 119, pages 475–485, 2020.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *CoRR*, abs/1607.06450, 2016.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Gary Cottrell, and Julian J. McAuley. ReZero is all you need: fast convergence at large depth. In *UAI*, volume 161, pages 1352–1361, 2021.

Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-Performance Large-Scale Image Recognition Without Normalization. In *ICML*, volume 139, pages 1059–1071, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *NeurIPS*, pages 6572–6583, 2018.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic Discovery of Optimization Algorithms. 2023.

Matthew J. Colbrook, Lorna J. Ayton, and Máté Szoke. Residual Dynamic Mode Decomposition: Robust and verified Koopmanism. *CoRR*, abs/2205.09779, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

N. Benjamin Erichson, Lionel Mathelin, J. Nathan Kutz, and Steven L. Brunton. Randomized Dynamic Mode Decomposition. *SIAM J. Appl. Dyn. Syst.*, 18(4):1867–1891, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware Minimization for Efficiently Improving Generalization. In *ICLR*, 2021.

Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry Hungry Hippos: Towards Language Modeling with State Space Models. In *ICLR*, 2023.

Aaron Gokaslan and Vanya Cohen. OpenWebText Corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *CoRR*, abs/2312.00752, 2023.

Eldad Haber and Lars Ruthotto. Stable Architectures for Deep Neural Networks. *CoRR*, abs/1705.03341, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016.

Nicholas J. Higham. *Functions of matrices - theory and computation*. 2008.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *NeurIPS*, pages 8580–8589, 2018.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Bernard O Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Milan Korda and Igor Mezic. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Autom.*, 93:149–160, 2018.

Heinz-Otto Kreiss. Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren. *BIT Numerical Mathematics*, 2(3): 153–181, 1962.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, and Joshua L. Proctor. *Dynamic mode decomposition - data-driven modeling of complex systems*. 2016.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-Arnold Networks. *CoRR*, abs/2404.19756, 2024.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *CoRR*, abs/1712.09707, 2017.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *ICLR*, 2017.

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

Igor Mezić. Analysis of fluid flows via spectral properties of the Koopman operator. *Annual review of fluid mechanics*, 45(1):357–378, 2013.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. In *ICLR*, 2018.

Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *ICML (3)*, volume 28, pages 1310–1318, 2013.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *NIPS*, pages 4785–4795, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *NIPS*, pages 3360–3368, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Clarence W Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S Henningson. Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, 641:115–127, 2009.

Tim Salimans and Diederik P. Kingma. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *NIPS*, page 901, 2016.

Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656: 5–28, 2010.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep Information Propagation. In *ICLR*, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*, 2017.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning Koopman Invariant Subspaces for Dynamic Mode Decomposition. In *NIPS*, pages 1130–1140, 2017.

Qwen Team. Qwen3 Technical Report. *CoRR*, abs/2505.09388, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288, 2023.

Lloyd N Trefethen and Mark Embree. Spectra and pseudospectra: the behavior of nonnormal matrices and operators. 2020.

Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

Jonathan H Tu. *Dynamic mode decomposition: Theory and applications*. PhD thesis, Princeton University, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NIPS*, pages 5998–6008, 2017.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. DeepNet: Scaling Transformers to 1,000 Layers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(10):6761–6774, 2024.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. *J. Nonlinear Sci.*, 25(6):1307–1346, 2015.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On Layer Normalization in the Transformer Architecture. In *ICML*, volume 119, pages 10524–10533, 2020.

Wei Xiong, Xiaomeng Huang, Ziyang Zhang, Ruixuan Deng, Pei Sun, and Yang Tian. Koopman neural operator as a mesh-free solver of non-linear partial differential equations. *J. Comput. Phys.*, 513:113194, 2024.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *CoRR*, abs/2203.03466, 2022.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. 2020.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *NeurIPS*, pages 12360–12371, 2019.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup Initialization: Residual Learning Without Normalization. In *ICLR*, 2019.

Residual Koopman Spectral Profiling for Predicting and Preventing Transformer Training Instability (Supplementary Material)

Bum Jun Kim[*1] **Shohei Taniguchi**[1] **Makoto Kawano**[1] **Yusuke Iwasawa**[1] **Yutaka Matsuo**[1]

[1]Graduate School of Engineering, The University of Tokyo, Japan

APPENDIX TABLE OF CONTENTS

A LIST OF NOTATION

Table 12: List of notations used in the main paper and supplementary material.

Symbol	Meaning
Core sizes and indices	
L	The number of layers.
ℓ	Layer index.
d	Hidden dimension.
N	The number of snapshots.
r	Randomized DMD rank, the number of eigenvalues used in KSS.
Dynamics and operators	
$\mathbf{h}_\ell$	Residual stream at layer ℓ.
$F_\ell(\cdot)$	Residual mapping $F_\ell(\mathbf{h}_\ell) = \mathbf{h}_\ell + f_\ell(\mathbf{h}_\ell; \theta_\ell)$.
$\mathcal{K}$	Koopman operator.
$\hat{\mathbf{A}}_\ell$	DMD estimate of the layer-ℓ Koopman operator.
$\mathbf{V}, \mathbf{\Lambda}$	Eigenvectors and eigenvalues of $\hat{\mathbf{A}}_\ell$.
λ_j	Eigenvalue.
$\rho(\mathbf{A})$	Spectral radius.
$\kappa(\mathbf{V})$	Eigenvector condition number, a measure of non-normality.
$\mathcal{K}(\mathbf{A})$	Kreiss constant.
Snapshots and whitening	
$\mathbf{X}_\ell, \mathbf{Y}_\ell$	Layer-ℓ snapshot matrices.
$\tilde{\mathbf{X}}_\ell, \tilde{\mathbf{Y}}_\ell$	Whitened snapshots.
$\hat{\mathbf{\Sigma}}_X$	Regularized sample covariance used for whitening.
$(\cdot)^\dagger$	Moore–Penrose pseudoinverse.
Spectral diagnostics	
$M_{>1}$	Unstable spectral mass.
$M_{\approx 1}$	Near-unit spectral mass.
$M_{<1}$	Contractive spectral mass.
η_{nl}	Nonlinearity ratio, a fit-error measure.
$\epsilon_u, \epsilon_n, \delta_c$	Thresholds defining spectral-mass bins.
KSS regularization	
$\mathcal{L}^\ell_{\mathrm{KSS}}$	KSS loss for layer ℓ.
α	KSS regularization weight.
τ_u, τ_l	Upper and lower band thresholds in KSS.
γ	Target near-unit mass level.
m^{soft}_ℓ	Soft near-unit mass estimate.
Probability and norms	
D	Divergence indicator.
$P(D = 1 \mid \mathcal{S})$	Predicted divergence probability.
$\mathbb{E}[\cdot]$	Expectation.
$\mathrm{Tr}(\cdot)$	Trace.
$\|\cdot\|_2, \|\cdot\|_F$	Operator and Frobenius norms.
$\|\cdot\|$	Absolute value.
$\mathbb{R}, \mathbb{C}$	Real and complex number fields.

B ADDITIONAL THEORETICAL RESULTS

The following results extend the theoretical analysis presented in Section 4.

B.1 SUPPLEMENT TO THEOREM 1

Proposition 3 (Bauer–Fike: Non-normality Caveat). *Let $\mathbf{A} \in \mathbb{C}^{d\times d}$ be diagonalizable with eigendecomposition $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$. For any perturbation $\mathbf{E}$ with $\|\mathbf{E}\|_2 \le \delta$ and any eigenvalue $\tilde{\lambda} \in \operatorname{spec}(\mathbf{A}+\mathbf{E})$, we have*

$$\min_k |\tilde{\lambda} - \lambda_k| \le \kappa(\mathbf{V}) \cdot \delta. \tag{20}$$

Proof. Let $\tilde{\lambda} \in \operatorname{spec}(\mathbf{A}+\mathbf{E})$ with eigenvector $\mathbf{x} \neq \mathbf{0}$, that is, $(\mathbf{A}+\mathbf{E})\mathbf{x} = \tilde{\lambda}\mathbf{x}$. Write $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ and set $\mathbf{y} \triangleq \mathbf{V}^{-1}\mathbf{x} \neq \mathbf{0}$. Left-multiplying by $\mathbf{V}^{-1}$ gives

$$(\mathbf{\Lambda} - \tilde{\lambda}\mathbf{I})\mathbf{y} = -\mathbf{V}^{-1}\mathbf{E}\mathbf{V}\mathbf{y}.$$

Taking Euclidean norms,

$$\left\|(\mathbf{\Lambda} - \tilde{\lambda}\mathbf{I})\mathbf{y}\right\|_2 \le \left\|\mathbf{V}^{-1}\right\|_2 \|\mathbf{E}\|_2 \|\mathbf{V}\|_2 \|\mathbf{y}\|_2 = \kappa(\mathbf{V})\delta \|\mathbf{y}\|_2 .$$

Because $\mathbf{\Lambda} - \tilde{\lambda}\mathbf{I}$ is diagonal with diagonal entries $(\lambda_j - \tilde{\lambda})$,

$$\left\|(\mathbf{\Lambda} - \tilde{\lambda}\mathbf{I})\mathbf{y}\right\|_2 \ge \min_k |\lambda_k - \tilde{\lambda}| \cdot \|\mathbf{y}\|_2 .$$

Canceling $\|\mathbf{y}\|_2$ yields Eq. 20. □

B.2 DMD CONVERGENCE ANALYSIS

We first establish finite-sample convergence guarantees for DMD estimation in the presence of nonlinearity.

Assumption 1 (Data Distribution). *Let $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d$ be a random pair drawn from the joint distribution induced by a layer transition. We assume centered covariates: $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}$ with $\sigma_{\min}(\mathbf{\Sigma}) \ge \sigma_0 > 0$. We also assume sub-Gaussian tails: $\|\mathbf{x}\|_{\psi_2} \le K$ and $\|\mathbf{y}\|_{\psi_2} \le K$ for some $K > 0$. Define the cross-covariance $\mathbf{C}_{yx} \triangleq \mathbb{E}[\mathbf{y}\mathbf{x}^\top]$. Let $\epsilon \ge 0$ be the whitening regularizer in Eq. 7, and set $\mathbf{\Sigma}_\epsilon \triangleq \mathbf{\Sigma} + \epsilon\mathbf{I}$. Then $\sigma_{\min}(\mathbf{\Sigma}_\epsilon) \ge \sigma_0$. Define*

$$\mathbf{M}_\epsilon \triangleq \mathbf{\Sigma}_\epsilon^{-1/2}\mathbf{C}_{yx}\mathbf{\Sigma}_\epsilon^{-1/2}, \tag{21}$$

$$\mathbf{G}_\epsilon \triangleq \mathbf{\Sigma}_\epsilon^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}_\epsilon^{-1/2}, \tag{22}$$

and the regularized whitened population least-squares operator

$$\mathbf{A}_{\mathrm{w},\epsilon}^{\mathrm{LS}} \triangleq \mathbf{M}_\epsilon\mathbf{G}_\epsilon^{-1}. \tag{23}$$

The nonlinearity ratio η_{nl} is a normalized linear-fit error that we use as a practical diagnostic for linear-approximation reliability.

Theorem 4 (Whitened DMD Finite-Sample Convergence). *Under Assumption 1, let $\sigma_\epsilon \triangleq \sigma_{\min}(\mathbf{\Sigma}_\epsilon)$. Then $\sigma_\epsilon \ge \sigma_0$. Let $\delta_{\mathrm{fail}} \in (0,1)$. Consider the events*

$$\left\|\hat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_\epsilon\right\|_2 \le \frac{1}{2}\sigma_\epsilon, \tag{24}$$

and

$$\left\|\hat{\mathbf{G}} - \mathbf{G}_\epsilon\right\|_2 \le \frac{1}{2}\sigma_{\min}(\mathbf{G}_\epsilon), \tag{25}$$

where $\hat{\mathbf{G}} \triangleq \hat{\mathbf{\Sigma}}_X^{-1/2}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Sigma}}_X^{-1/2}$. Then $\hat{\mathbf{G}}$ is invertible, and $\left\|\hat{\mathbf{G}}^{-1}\right\|_2 \le 2\left\|\mathbf{G}_\epsilon^{-1}\right\|_2$. Assuming $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ is invertible, for example, when $N \ge d$ and $\operatorname{rank}(\tilde{\mathbf{X}}) = d$, there exist absolute constants $C, c > 0$ such that if $N \ge c(d + \log(2/\delta_{\mathrm{fail}}))$, then on the event Eq. 24 and Eq. 25:

$$\left\|\hat{\mathbf{A}} - \mathbf{A}_{\mathrm{w},\epsilon}^{\mathrm{LS}}\right\|_2 \le C\left\|\mathbf{G}_\epsilon^{-1}\right\|_2 \frac{K^2}{\sigma_\epsilon}\Delta_N + C\left\|\mathbf{G}_\epsilon^{-1}\right\|_2^2 \frac{K^2\left\|\mathbf{C}_{yx}\right\|_2}{\sigma_\epsilon^2}\Delta_N, \tag{26}$$

where $\Delta_N \triangleq \sqrt{\frac{d+\log(2/\delta_{\text{fail}})}{N}} + \frac{d+\log(2/\delta_{\text{fail}})}{N}$.

Proof. We bound the estimation error relative to the whitened population least-squares operator $\mathbf{A}_{\text{w},\epsilon}^{\text{LS}}$. Assume $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ is invertible, so that $\tilde{\mathbf{X}}^\dagger = \tilde{\mathbf{X}}^\top(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)^{-1}$ and $\hat{\mathbf{A}} = \hat{\mathbf{M}}\hat{\mathbf{G}}^{-1}$ with $\hat{\mathbf{M}} \triangleq \hat{\mathbf{\Sigma}}_X^{-1/2}\hat{\mathbf{C}}_{yx}\hat{\mathbf{\Sigma}}_X^{-1/2}$ and $\hat{\mathbf{G}} \triangleq \hat{\mathbf{\Sigma}}_X^{-1/2}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Sigma}}_X^{-1/2}$, where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$, $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^N \mathbf{y}_i$, $\hat{\mathbf{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^N(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$, $\hat{\mathbf{C}}_{yx} = \frac{1}{N-1}\sum_{i=1}^N(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$, and $\hat{\mathbf{\Sigma}}_X = \hat{\mathbf{\Sigma}} + \epsilon\mathbf{I}$.

First, we estimate the covariance. For N independently and identically distributed samples with $\|\mathbf{x}\|_{\psi_2} \le K$, standard covariance concentration for the centered sample covariance yields [Tropp, 2012]

$$\left\|\hat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_\epsilon\right\|_2 \le CK^2\Delta_N, \tag{27}$$

with probability $\ge 1 - \delta_{\text{fail}}/2$ for $N \gtrsim d + \log(1/\delta_{\text{fail}})$.

Second, we bound the whitening perturbation. Standard perturbation theory for matrix square roots gives [Higham, 2008]:

$$\left\|\hat{\mathbf{\Sigma}}_X^{-1/2} - \mathbf{\Sigma}_\epsilon^{-1/2}\right\|_2 \le \frac{2}{\sigma_\epsilon^{3/2}}\left\|\hat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_\epsilon\right\|_2, \tag{28}$$

for $\left\|\hat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_\epsilon\right\|_2 \le \frac{1}{2}\sigma_\epsilon$.

Now, we combine the cross-covariance and covariance estimation errors. Let $\hat{\mathbf{C}}_{yx} = \frac{1}{N-1}\sum_{i=1}^N(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$. A similar sub-exponential matrix concentration bound gives [Tropp, 2012] $\left\|\hat{\mathbf{C}}_{yx} - \mathbf{C}_{yx}\right\|_2 \le CK^2\Delta_N$ with probability $\ge 1 - \delta_{\text{fail}}/2$. Decomposing

$$\hat{\mathbf{A}} - \mathbf{A}_{\text{w},\epsilon}^{\text{LS}} = \hat{\mathbf{M}}\hat{\mathbf{G}}^{-1} - \mathbf{M}_\epsilon\mathbf{G}_\epsilon^{-1} = (\hat{\mathbf{M}} - \mathbf{M}_\epsilon)\mathbf{G}_\epsilon^{-1} + \hat{\mathbf{M}}\left(\hat{\mathbf{G}}^{-1} - \mathbf{G}_\epsilon^{-1}\right),$$

and using a standard matrix inverse perturbation bound,

$$\left\|\hat{\mathbf{G}}^{-1} - \mathbf{G}_\epsilon^{-1}\right\|_2 \le \left\|\hat{\mathbf{G}}^{-1}\right\|_2\left\|\hat{\mathbf{G}} - \mathbf{G}_\epsilon\right\|_2\left\|\mathbf{G}_\epsilon^{-1}\right\|_2 \le 2\left\|\mathbf{G}_\epsilon^{-1}\right\|_2^2\left\|\hat{\mathbf{G}} - \mathbf{G}_\epsilon\right\|_2$$

on Eq. 25 yields a cross-covariance term scaling as $\left\|\mathbf{G}_\epsilon^{-1}\right\|_2\sigma_\epsilon^{-1}$ and a whitening term plus a covariance term scaling as $\left\|\mathbf{G}_\epsilon^{-1}\right\|_2^2\|\mathbf{C}_{yx}\|_2\,\sigma_\epsilon^{-2}$, giving Eq. 26.

Combining the bounds yields Eq. 26 under the event Eq. 24 and Eq. 25. □

Remark 1 (On $\mathbf{G}_\epsilon^{-1}$ for $\mathbf{\Sigma}_\epsilon = \mathbf{\Sigma} + \epsilon\mathbf{I}$). Because $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_\epsilon$ commute, $\mathbf{G}_\epsilon$ has eigenvalues $\lambda_i/(\lambda_i + \epsilon)$, hence

$$\left\|\mathbf{G}_\epsilon^{-1}\right\|_2 = \frac{\sigma_{\min}(\mathbf{\Sigma}) + \epsilon}{\sigma_{\min}(\mathbf{\Sigma})}.$$

If $\sigma_{\min}(\mathbf{\Sigma})$ is treated as a fixed constant bounded away from 0, the factors of $\left\|\mathbf{G}_\epsilon^{-1}\right\|_2$ can be absorbed into the constant C.

Remark 2 (Sample Complexity). Theorem 4 suggests that, under the stability events Eq. 24–Eq. 25, $N = \tilde{O}\left(d\left(\left\|\mathbf{G}_\epsilon^{-1}\right\|_2\frac{K^2}{\sigma_\epsilon} + \left\|\mathbf{G}_\epsilon^{-1}\right\|_2^2\frac{K^2\|\mathbf{C}_{yx}\|_2}{\sigma_\epsilon^2}\right)^2\varepsilon^{-2}\right)$ samples are sufficient for ε-accurate DMD estimation, up to logarithmic factors. For typical transformers with $d = 256$ to 768, $N \approx 2048$ provides reliable estimates.

Remark 3 (Modeling Mismatch and η_{nl}). Theorem 4 is an estimation bound for the whitened population least-squares operator $\mathbf{A}_{\text{w},\epsilon}^{\text{LS}}$. When the layer transition is nonlinear, $\mathbf{A}_{\text{w},\epsilon}^{\text{LS}}$ can be a poor proxy for other targets, such as a local Jacobian or a richer Koopman approximation, even if it is well-estimated. We use the empirical nonlinearity ratio η_{nl} defined in Eq. 5 as a practical diagnostic for when linear DMD features are less reliable.

B.3 NON-NORMALITY AND TRANSIENT GROWTH

Spectral radius bounds alone are insufficient for analyzing non-normal matrices. The Kreiss matrix theorem provides tight bounds on transient behavior [Kreiss, 1962, Trefethen and Embree, 2020].

Theorem 5 (Kreiss Constant Characterization)**.** *The Kreiss constant of* $\mathbf{A} \in \mathbb{C}^{d\times d}$ *is:*

$$\mathcal{K}(\mathbf{A}) \triangleq \sup_{|z|>1} (|z|-1) \left\|(z\mathbf{I}-\mathbf{A})^{-1}\right\|_2 \tag{29}$$

Assume $\mathbf{A}$ *is power-bounded, that is,* $\sup_{n\geq 0} \|\mathbf{A}^n\|_2 < \infty$*; equivalently,* $\mathcal{K}(\mathbf{A}) < \infty$*. Then the Kreiss matrix theorem states:*

$$\mathcal{K}(\mathbf{A}) \leq \sup_{n\geq 0} \|\mathbf{A}^n\|_2 \leq e \cdot d \cdot \mathcal{K}(\mathbf{A}) \tag{30}$$

For a diagonalizable $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$*:*

$$\mathcal{K}(\mathbf{A}) \leq \kappa(\mathbf{V}) \cdot \sup_{|z|>1} \max_j \frac{|z|-1}{|z-\lambda_j|} \tag{31}$$

Proof. We first prove the two inequalities in Eq. 30 and then Eq. 31.

Lower bound: $\mathcal{K}(\mathbf{A}) \leq \sup_{n\geq 0} \|\mathbf{A}^n\|_2$**.** Let $M \triangleq \sup_{n\geq 0} \|\mathbf{A}^n\|_2 < \infty$. Because $\mathbf{A}$ is power-bounded, $\sup_{n\geq 0} \|\mathbf{A}^n\|_2 < \infty$, so the Neumann series converges in operator norm for any $|z| > 1$. For any $|z| > 1$, the Neumann series gives the operator-norm expansion

$$(z\mathbf{I}-\mathbf{A})^{-1} = z^{-1}\sum_{n=0}^{\infty} \mathbf{A}^n z^{-n},$$

hence

$$\left\|(z\mathbf{I}-\mathbf{A})^{-1}\right\|_2 \leq \frac{1}{|z|}\sum_{n=0}^{\infty} \frac{\|\mathbf{A}^n\|_2}{|z|^n} \leq \frac{M}{|z|}\sum_{n=0}^{\infty} |z|^{-n} = \frac{M}{|z|-1}.$$

Multiplying by $(|z|-1)$ and taking the supremum over $|z| > 1$ yields $\mathcal{K}(\mathbf{A}) \leq M$.

Upper bound: $\sup_{n\geq 0} \|\mathbf{A}^n\|_2 \leq ed\mathcal{K}(\mathbf{A})$**.** This is the finite-dimensional Kreiss matrix theorem: the resolvent bound $\sup_{|z|>1}(|z|-1)\left\|(z\mathbf{I}-\mathbf{A})^{-1}\right\|_2 < \infty$ is equivalent to power-boundedness, and quantitatively implies $\sup_{n\geq 0} \|\mathbf{A}^n\|_2 \leq C_d\mathcal{K}(\mathbf{A})$ for an explicit dimension-dependent constant C_d; one standard choice is $C_d = ed$.

Diagonalizable case. If $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, then for any $z \notin \operatorname{spec}(\mathbf{A})$,

$$(z\mathbf{I}-\mathbf{A})^{-1} = \mathbf{V}(z\mathbf{I}-\mathbf{\Lambda})^{-1}\mathbf{V}^{-1}.$$

Taking norms gives

$$\left\|(z\mathbf{I}-\mathbf{A})^{-1}\right\|_2 \leq \|\mathbf{V}\|_2 \left\|(z\mathbf{I}-\mathbf{\Lambda})^{-1}\right\|_2 \left\|\mathbf{V}^{-1}\right\|_2 = \kappa(\mathbf{V}) \left\|(z\mathbf{I}-\mathbf{\Lambda})^{-1}\right\|_2.$$

Because $(z\mathbf{I}-\mathbf{\Lambda})^{-1}$ is diagonal with diagonal entries $(z-\lambda_j)^{-1}$, its spectral norm is $\max_j |z-\lambda_j|^{-1}$, hence

$$(|z|-1)\left\|(z\mathbf{I}-\mathbf{A})^{-1}\right\|_2 \leq \kappa(\mathbf{V}) \cdot \max_j \frac{|z|-1}{|z-\lambda_j|}.$$

Taking the supremum over $|z| > 1$ yields Eq. 31. □

The interpretation is as follows: a high $\mathcal{K}(\mathbf{A})$ indicates hidden instability. Even when $\rho(\mathbf{A}) \leq 1$, non-orthogonal eigenvectors produce a transient growth $\|\mathbf{A}^n\|_2 \gg 1$ for intermediate n.

C PROOF OF THEOREM 1

Proof. For $\mathbf{x}$ uniform on the unit sphere, rotational invariance implies $\mathbb{E}[\mathbf{x}\mathbf{x}^*] = \frac{1}{d}\mathbf{I}$. Therefore

$$\mathbb{E}\|\mathbf{A}\mathbf{x}\|_2^2 = \mathbb{E}[\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}] = \operatorname{Tr}\left(\mathbf{A}^*\mathbf{A}\mathbb{E}[\mathbf{x}\mathbf{x}^*]\right) = \frac{1}{d}\operatorname{Tr}(\mathbf{A}^*\mathbf{A}) = \frac{1}{d}\|\mathbf{A}\|_F^2.$$

For a normal $\mathbf{A}$, $\|\mathbf{A}\|_F^2 = \sum_{j=1}^d |\lambda_j|^2$, giving

$$\mathbb{E}\,\|\mathbf{A}\mathbf{x}\|_2^2 = \frac{1}{d}\sum_{j=1}^{d} |\lambda_j|^2.$$

If $\rho(\mathbf{A}) \le 1+\epsilon_u$, then $|\lambda_j|^2 \le (1+\epsilon_u)^2$ for all j, giving the upper bound. For the lower bound, at least a fraction $M_{\approx 1}(\mathbf{A})$ of the eigenvalues satisfy $|\lambda_j| \ge 1-\epsilon_n$, so

$$\mathbb{E}\,\|\mathbf{A}\mathbf{x}\|_2^2 \ge (1-\epsilon_n)^2 M_{\approx 1}(\mathbf{A}).$$

For the diagonalizable extension $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, note that for an isotropic $\mathbf{x}$ we always have $\mathbb{E}\,\|\mathbf{A}\mathbf{x}\|_2^2 = \frac{1}{d}\,\|\mathbf{A}\|_F^2$. Moreover, $\frac{1}{\kappa(\mathbf{V})}\,\|\mathbf{\Lambda}\|_F \le \|\mathbf{A}\|_F \le \kappa(\mathbf{V})\,\|\mathbf{\Lambda}\|_F$, so $\mathbb{E}\,\|\mathbf{A}\mathbf{x}\|_2^2$ is within factors $\kappa(\mathbf{V})^{\pm 2}$ of $\frac{1}{d}\sum_j |\lambda_j|^2$. Combining this with $\rho(\mathbf{A}) \le 1+\epsilon_u$ and the definition of $M_{\approx 1}(\mathbf{A})$ yields the stated bound. □

Proof of Corollary 2.

Proof. Assume $\mathbb{E}[\mathbf{h}_\ell \mathbf{h}_\ell^*] = \frac{1}{d}\mathbb{E}\,\|\mathbf{h}_\ell\|_2^2\,\mathbf{I}$ for each $\ell = 0, \dots, L-1$. Let $q_\ell \triangleq \frac{1}{d}\sum_{j=1}^d |\lambda_j^\ell|^2$. Then

$$\mathbb{E}\,\|\mathbf{h}_{\ell+1}\|_2^2 = \mathrm{Tr}\left(\mathbf{A}_\ell^* \mathbf{A}_\ell \mathbb{E}[\mathbf{h}_\ell \mathbf{h}_\ell^*]\right) = q_\ell \mathbb{E}\,\|\mathbf{h}_\ell\|_2^2,$$

and recursion yields $\mathbb{E}\,\|\mathbf{h}_L\|_2^2 = \mathbb{E}\,\|\mathbf{h}_0\|_2^2 \prod_{\ell=0}^{L-1} q_\ell$. By Theorem 1, $q_\ell \ge (1-\epsilon_n)^2 M_{\approx 1}(\mathbf{A}_\ell)$, proving the stated bound. □

D EXPERIMENTAL DETAILS

We ran experiments on 4× NVIDIA A100-SXM4-40GB graphics processing unit (GPU) devices. We ran experiments across six normalization strategies, including Pre-LN, Post-LN, RMSNorm, DeepNorm, SubLN, and No-Norm, with additional architecture-specific studies on MoE [Shazeer et al., 2017], Mamba [Gu and Dao, 2023], and KAN [Liu et al., 2024].

We use the following hyperparameters. Models use $d \in \{128, 256, 512, 768, 1024\}$, $n_{\text{heads}} \in \{4, 8, 16\}$, and $L \in \{4, 6, 8, 12, 16, 24\}$. For training, the mini-batch size is 16 to 32 with 5 to 20 epochs and 100 to 1000 warmup steps. We use the AdamW and sweep learning rates within standard ranges. We use seeds $\{42, 123, 456\}$ for reproducibility. Our default RKSP recipe estimates randomized DMD eigenvalues with rank $r = 32$ using $N = 2048$ snapshots, whitening $\epsilon = 10^{-5}$, ResDMD reliability threshold $\tau = 0.10$, and near-unit bin $[0.90, 1.05]$. KSS is applied every 10 to 20 steps, sampling 50% of layers per update to reduce overhead, with $T = 20$, $\tau_l = 0.90$, $\tau_u = 1.05$, $\beta = 1$, and $\gamma \in [0.3, 0.5]$ fixed. We sweep the regularization weight $\alpha \in \{0.01, 0.05, 0.10, 0.15, 0.20\}$; the resulting overhead ranges from 8% to 12% in practice.

D.1 ASSOCIATIVE-RECALL TASK

The associative-recall task refers to a synthetic key-value retrieval classification task commonly used to probe associative memory and recall in long-context sequence models [Fu et al., 2023, Arora et al., 2024]. For each sample, we generate n_{pairs} key-value pairs $\{(k_i, v_i)\}_{i=1}^{n_{\text{pairs}}}$ and construct

$$\mathbf{x} = [k_1, v_1, \dots, k_{n_{\text{pairs}}}, v_{n_{\text{pairs}}}, p_1, \dots, p_m, q],$$

where q is a query key chosen from $\{k_i\}$ and the label is the corresponding matched value $y = v_j$. The model predicts only this final target with cross-entropy on the last-position logits, implemented as `F.cross_entropy(logits[:, -1, :], y)`.

D.2 SYNTHETIC LM TASK

This synthetic token-level language modeling setup is inspired by prior work using controlled synthetic sequences to analyze recall and long-range behavior in efficient sequence models [Fu et al., 2023, Arora et al., 2024]. Each sample is a length-T token sequence over a vocabulary of size V; unless stated otherwise, we use $T = 256$ and $V = 10{,}000$. A sequence is generated by concatenating randomly sampled segments until reaching length T, then truncating. Segments come in three types: repetition segments repeat a short pattern of length 2 to 5 for 2 to 4 repeats, sequential segments are contiguous integer runs of length 5 to 15, and random segments are independently and identically distributed tokens of length 5 to 15. All tokens are sampled from $\{10, \ldots, V-1\}$ so that a small identifier range remains available for special tokens. Sequences are generated directly at the token level by these rules.

The learning objective is standard next-token prediction. Given tokens $(x_1, \ldots, x_T)$, the model predicts x_{t+1} from the prefix $(x_1, \ldots, x_t)$ and is trained with token-level cross-entropy over $t = 1, \ldots, T-1$. We report validation token accuracy and perplexity $\exp(\text{mean cross-entropy})$ on a held-out synthetic validation split. For large-scale runs, we use 20K training sequences and 2K validation sequences per trial, regenerated deterministically from the run seed.

D.3 PRETRAINED LM FIXED-PROMPT PROTOCOL

We use a forward-only profiling protocol with a deterministic text set and no fine-tuning updates, in the same spirit as prompt-based evaluation and activation-probing analyses of pretrained transformers [Brown et al., 2020, Elhage et al., 2021]. The fixed short-prompt setting is implemented as explicit prompt lists with 32 total prompts per run, either 4 prompts repeated 8 times or 8 prompts repeated 4 times, and both variants use a token length cap of 64. For each model, we run a single batched forward pass with hidden-state outputs enabled and collect residual-stream activations from the embedding and transformer layer outputs.

For each layer transition $(\ell, \ell+1)$, we flatten token positions, subsample up to $N \in \{1024, 2048\}$ token states, and apply whitened DMD. Spectral partitions use the same thresholds as the analysis code: unstable when $|\lambda| > 1.05$, near-unit when $0.90 \le |\lambda| \le 1.05$, and over-damped when $|\lambda| < 0.80$. We report early, middle, and late summaries by splitting layers into depth thirds and averaging each metric within the corresponding group.

E PROTOCOL ROBUSTNESS CHECKS

Since RKSP is intended as a fixed diagnostic rather than a tuned classifier, we ran one-factor-at-a-time robustness checks for prompt set, snapshot count, whitening regularization, ResDMD reliability threshold, and spectral-bin thresholds. The center setting uses rank-32 reduced DMD, whitening $\epsilon = 10^{-5}$, reliability threshold $\tau = 0.10$, near-unit bin $[0.90, 1.05]$, and 1024 snapshots; the snapshot sweep also includes the 2048-snapshot setting used in the main controlled experiments. Each prompt-set sweep uses a deterministic pool of 96 prompts per set, so the 2048-snapshot condition can use the full requested count.

Table 13: Robustness of the RKSP near-unit mass under protocol changes on GPT-2 124M hidden states. The aggregate risk regime is stable under prompt choice, snapshot count, reliability threshold, and moderate spectral-bin perturbations. Whitening regularization is the most sensitive numerical choice, but practical values 10^{-6} to 10^{-4} keep $M_{\approx 1}$ in the low-to-moderate range.

Design choice	Values tested	Mean $M_{\approx 1}$ range	Max $\lvert\Delta M_{\approx 1}\rvert$	Min layer-rank corr.	Reliable modes
Prompt set	science, news, code, dialogue	0.156–0.193	0.031	0.667	1.00–1.00
Snapshots	512, 1024, 1536, 2048	0.161–0.182	0.021	0.702	1.00–1.00
Whitening ϵ	10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}	0.159–0.352	0.190	0.575	1.00–1.00
Reliability τ	0.05, 0.10, 0.20, 0.50	0.161–0.161	0.000	1.000	1.00–1.00
Spectral bin	$[0.88, 1.05]$, $[0.90, 1.03]$, $[0.90, 1.05]$, $[0.90, 1.07]$, $[0.92, 1.05]$	0.125–0.198	0.036	0.917	1.00–1.00

Table 14: Comparison of μP, standard parameterization, and KSS. Results use the No-Norm setting with learning rate ranges from 0.005 to 0.01 across 24 trials. Measured on the associative-recall task.

Method	Div.% (lower)	Acc.% (higher)	$M_{\approx 1}$	Transfer	Overhead
Standard Init, AdamW	66.7	28.5	0.85	✗	—
μP, width transfer	41.7	35.8	0.72	✓	$< 1\%$
μP + higher learning rate	50.0	38.2	0.78	✓	$< 1\%$
KSS, $\alpha = 0.15$	**12.5**	**48.2**	0.58	✗	10.8%
μP + KSS	**8.3**	**51.5**	0.52	✓	11.2%

The reliability threshold has no effect in this sweep because all tested rank-32 modes pass the ResDMD filter, so the score is not a filtering artifact. Even the deliberately large whitening stress value $\epsilon = 10^{-3}$ gives $M_{\approx 1} = 0.352$, far below the high-risk No-Norm regime in Table 1, where $M_{\approx 1} = 0.80$. Thus, the qualitative risk classification does not depend on hidden architecture-specific tuning.

F COMPUTATIONAL COST

Whitened DMD requires $O(d^2N + d^3)$ operations. Randomized singular value decomposition reduces this cost to $O(dNr + r^3)$ for rank-r approximation. With typical values $d = 768$, $N = 2048$, and $r = 32$, full RKSP analysis completes in 2.5 to 3.5 seconds per layer on a single GPU.

G EXTENDED BASELINE AND OPTIMIZER COMPARISONS

G.1 EXTENDED OPTIMIZER BASELINES

We extend the baseline comparisons to include μP and the Layer-wise Adaptive Moments for Batch training (LAMB) optimizer [You et al., 2020].

μP. The μP enables hyperparameter transfer across different model widths by appropriately scaling learning rates appropriately. Table 14 summarizes the μP comparisons and the combined μP + KSS setting. We observe moderate stability gains from μP, with a 38% divergence reduction through initialization scaling. The combined μP + KSS approach achieves the lowest divergence, 8.3%, and the highest accuracy, 51.5%. The μP setting enables transfer, while KSS provides stability, and the distinct mechanisms suggest orthogonal benefits.

LAMB Optimizer. LAMB normalizes updates per layer, which changes spectral dynamics. Table 15 reports the optimizer comparison, including LAMB and Lion. LAMB outperforms AdamW: its layer-wise normalization provides implicit stability with a 31% divergence reduction. LAMB enables higher learning rates, reaching $3\times$ to $4\times$ AdamW, or up to $7\times$ with learning rate warmup. KSS still provides orthogonal benefits, and the LAMB + KSS combination achieves the best results at 8.3% divergence and 52.8% accuracy. From a spectral perspective, LAMB suppresses expansive mass and reduces near-unit mass, from 0.85 with AdamW to 0.75, suggesting that its stability gains come from damping unstable modes and increasing damping; within comparable unstable-mass regimes, a lower $M_{\approx 1}$ aligns with greater stability.

H LARGE-SCALE PRETRAINED MODEL ANALYSIS

GPT-2 Analysis. Table 16 reports layer-group spectral statistics for GPT-2 [Radford et al., 2019], and Figure 4 reveals a recurring pattern. The normalized linear-fit error η_{nl} increases with depth, rising from $[0.48, 0.52]$ in the early layers to $[0.68, 0.71]$ in late layers. Simultaneously, the near-unit mass decreases from $M_{\approx 1} \in [0.68, 0.72]$ to $M_{\approx 1} \in [0.58, 0.60]$. This depth-wise trend has a clear implication: early layers are more linearly approximable, making DMD features more reliable, whereas late layers are less so.

Qwen3 Analysis. Table 17 reports the same forward-only RKSP protocol on Qwen3 [Team, 2025] models with 0.6B, 1.7B, and 4B parameters. The protocol uses 32 deterministic short prompts, token length cap 64, no parameter updates,

Table 15: Optimizer comparison among AdamW, LAMB, and Lion. Results use the No-Norm setting, 24 trials each. Measured on the associative-recall task.

Optimizer	Div.% (lower)	Acc.% (higher)	$M_{\approx 1}$	Best LR	Overhead
AdamW	66.7	28.5	0.85	0.003	—
AdamW + grad clip	58.3	30.8	0.83	0.005	$< 1\%$
LAMB	45.8	36.2	0.75	0.01	about 5%
LAMB + warmup	37.5	40.8	0.68	0.02	about 5%
Lion	45.8	36.5	0.78	0.001	about 15%
KSS, $\alpha = 0.15$	**12.5**	**48.2**	0.58	0.008	10.8%
LAMB + KSS	**8.3**	**52.8**	0.52	0.015	about 16%

Table 16: GPT-2 layer-wise spectral analysis. Start Linear, End Nonlinear pattern, shorthand for increasing η_{nl}. Spectral statistics are computed from residual-stream activations on a fixed set of short prompt sentences. Measured on a fixed short-prompt set.

Model	Layers	ρ	$\kappa(\mathbf{V})$	η_{nl}	$M_{\approx 1}$
GPT-2 124M	Early 0 to 3	1.15	8.3	0.52	0.68
GPT-2 124M	Middle 4 to 7	1.23	10.1	0.61	0.62
GPT-2 124M	Late 8 to 11	1.31	11.2	0.71	0.58
GPT-2 355M	Early 0 to 7	1.12	7.5	0.48	0.72
GPT-2 355M	Middle 8 to 15	1.19	9.2	0.58	0.65
GPT-2 355M	Late 16 to 23	1.27	10.8	0.68	0.60

1024 sampled residual-stream snapshots per layer, rank-32 whitened DMD, whitening $\epsilon = 10^{-5}$, ResDMD reliability threshold $\tau = 0.10$, and near-unit bin $[0.90, 1.05]$. Across all three scales, η_{nl} increases with depth, so the late layers are not interpreted as purely linear; RKSP reports this reduced linear-fit fidelity alongside localized non-normality through $\kappa(\mathbf{V})$.

Table 17: Qwen3 pretrained-model profiling with the fixed forward-only RKSP protocol. The central risk feature remains in a moderate pretrained-model range, $M_{\approx 1} \in [0.201, 0.430]$, well below the high-risk No-Norm regime in Table 1. These results support diagnostic scalability without architecture-specific retuning, not Qwen3-scale KSS intervention.

Model	Layers	$M_{\approx 1}$	η_{nl}	$\kappa(\mathbf{V})$
Qwen3-0.6B	Early 0–9	0.316 ± 0.132	0.936 ± 0.052	240.31 ± 567.66
Qwen3-0.6B	Middle 10–18	0.233 ± 0.051	0.990 ± 0.005	122.85 ± 102.95
Qwen3-0.6B	Late 19–27	0.201 ± 0.093	0.999 ± 0.001	49.47 ± 25.91
Qwen3-1.7B	Early 0–9	0.284 ± 0.095	0.964 ± 0.025	153.91 ± 295.31
Qwen3-1.7B	Middle 10–18	0.250 ± 0.055	0.994 ± 0.005	86.02 ± 134.02
Qwen3-1.7B	Late 19–27	0.295 ± 0.136	1.000 ± 0.000	99.24 ± 107.93
Qwen3-4B	Early 0–11	0.250 ± 0.133	0.972 ± 0.040	144.12 ± 150.20
Qwen3-4B	Middle 12–23	0.328 ± 0.108	0.988 ± 0.005	142.98 ± 98.07
Qwen3-4B	Late 24–35	0.430 ± 0.147	0.999 ± 0.001	79.77 ± 104.58

I CALIBRATION

Beyond discrimination measured by AUROC, calibrated probability estimation requires a probability map. The unfitted map $\sigma(10(M_{\approx 1} - 0.5))$ used in Figure 5 is a monotone visualization of the RKSP ranking signal, not a trained calibrator. On the same 215 divergence-prediction runs (81 diverged, 134 converged), it has AUROC 0.995 but ECE 0.283. We therefore fit one-dimensional post hoc maps from the same scalar RKSP score using 5-fold stratified cross-validation, so every reported probability is out-of-fold. This adds no model retraining, RKSP recomputation, or spectral tuning.

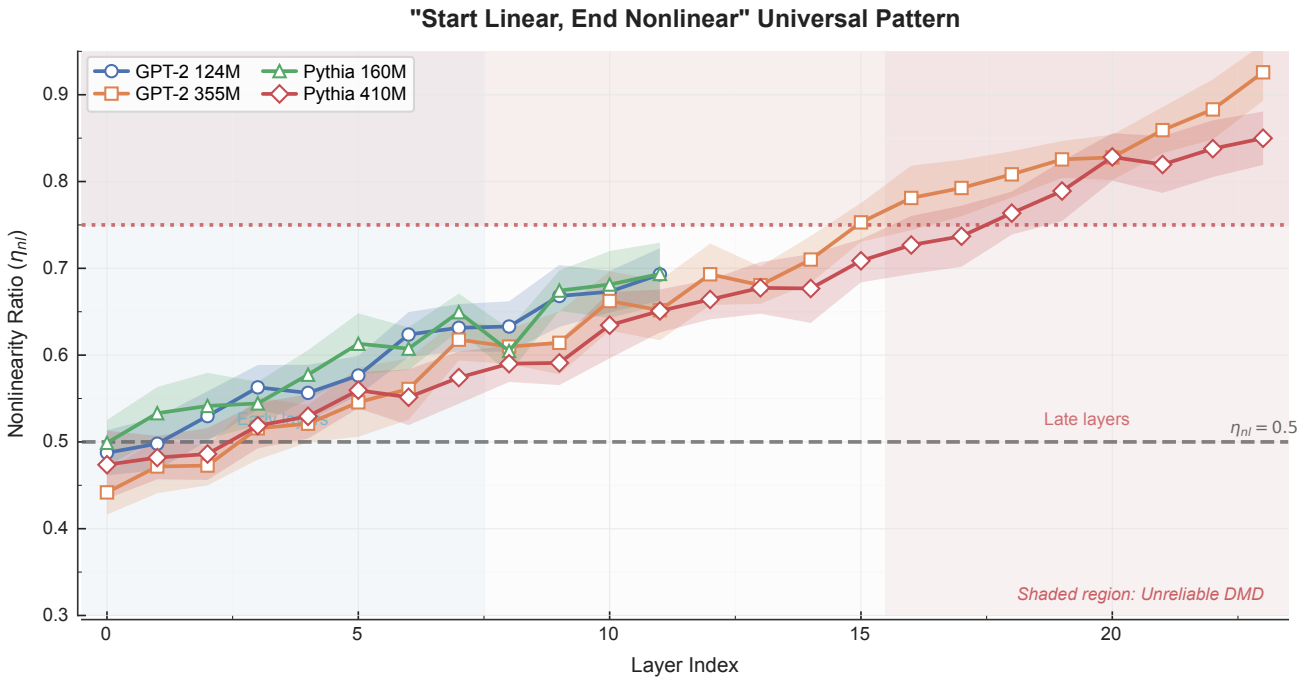

Figure 4: Start Linear, End Nonlinear pattern. Layer-wise normalized linear-fit error η_{nl} across four pretrained models. All models exhibit a monotonically increasing η_{nl} with depth, suggesting a consistent linear-approximation signature across models. Computed from residual-stream activations on a fixed set of short prompt sentences.

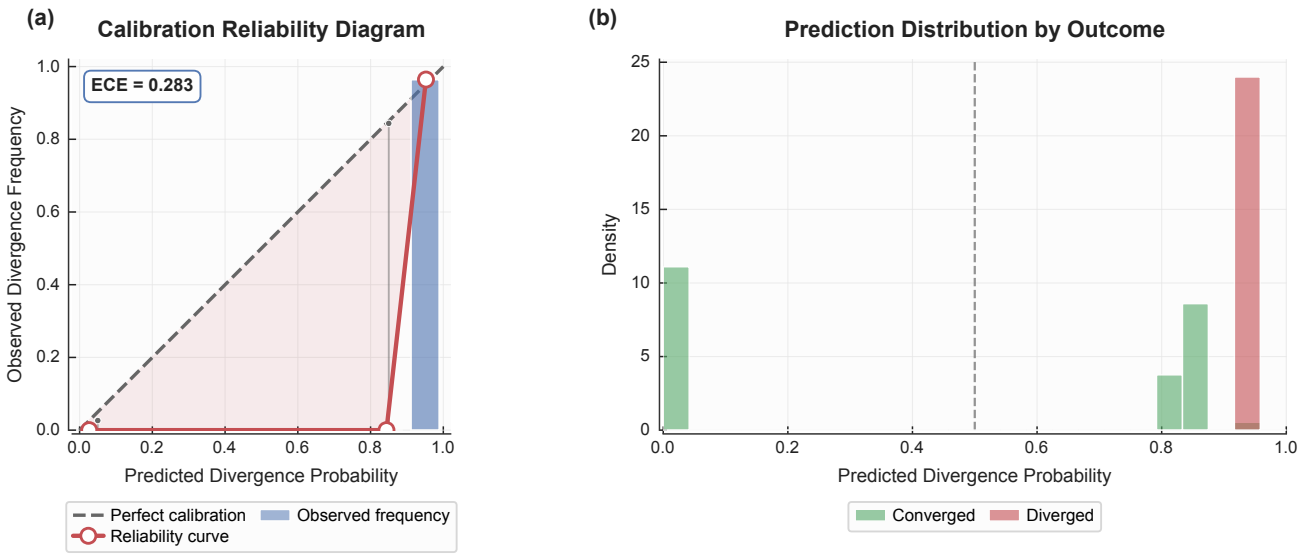

Figure 5: Reliability diagram for the unfitted monotone map $\sigma(10(M_{\approx 1} - 0.5))$. The ECE of 0.283 shows that the raw RKSP score is a strong ranking signal but is not itself a calibrated probability model. Based on associative-recall runs, calibration compares predictions to divergence outcomes from that task.

Table 18: Post hoc calibration of the scalar RKSP score on the divergence-prediction table. Probabilities for fitted maps are out-of-fold from 5-fold stratified cross-validation. Repeating cross-fitting over 30 stratified partitions gives Platt ECE 0.012 ± 0.002 and isotonic ECE 0.015 ± 0.005.

Probability mapping	Protocol	AUROC	AUPRC	ECE (lower)	ECE reduction	Brier (lower)	NLL (lower)
Sigmoid on $M_{\approx 1}$	No fitting	0.995	0.989	0.283	0.0%	0.242	0.666
Platt scaling on $M_{\approx 1}$	5-fold cross-fit	0.990	0.978	**0.012**	95.7%	**0.014**	**0.066**
Isotonic regression on $M_{\approx 1}$	5-fold cross-fit	0.990	0.972	0.016	94.4%	0.015	0.131

This separates the RKSP score from the probability map. The raw near-unit mass is the initialization-time risk score; for a given target configuration distribution, a held-out one-dimensional calibrator converts it into calibrated probabilities. If the model family, task, or optimizer distribution changes, only the calibrator must be refit, while the RKSP computation and one-forward-pass deployment remain unchanged.

J NOVEL ARCHITECTURE CASE STUDIES

As exploratory architecture-transfer case studies, we analyze three emerging architectures: MoE [Shazeer et al., 2017], SSMs including Mamba [Gu and Dao, 2023], and KAN [Liu et al., 2024].

MoE Transformers Table 19 presents a comparison of MoE routing and stability [Shazeer et al., 2017]. MoE routing induces a higher normalized linear-fit error: η_{nl} increases 15% to 20% compared to dense transformers due to discrete routing decisions. The choice of top-k affects spectral stability—higher k shifts spectral mass and changes $M_{\approx 1}$ alongside non-normality and unstable modes. The divergence reductions are consistent with suppressing unstable modes, and within comparable non-normality regimes, larger $M_{\approx 1}$ aligns with greater instability. KSS stabilizes MoE in this controlled case study, yielding a 4× divergence reduction with 6% accuracy improvement and providing supporting evidence for RKSP and KSS beyond dense transformers.

Table 19: MoE transformer with RKSP analysis. Routing instability revealed via spectral signatures. Results use $d = 256$, $L = 6$, and 24 trials. Load balancing loss $\lambda = 0.01$. Measured on the synthetic LM task with random-token next-token prediction.

Configuration	Top-k	$M_{\approx 1}$	ρ	η_{nl}	Div. of 24	Acc.%
MoE-Small, 8 experts	$k = 1$	0.72	2.85	0.68	6 of 24	32.4
MoE-Small, 8 experts	$k = 2$	0.58	2.12	0.55	2 of 24	41.7
MoE-Small, 8 experts	$k = 4$	0.45	1.78	0.48	1 of 24	38.2
MoE-Medium, 16 experts	$k = 2$	0.65	2.45	0.61	4 of 24	38.9
MoE-Medium, 16 experts	$k = 2$ + KSS	0.48	1.92	0.58	**1 of 24**	**44.5**

State Space Models: Mamba Table 20 compares SSM and transformer spectral properties for Mamba [Gu and Dao, 2023]. The theoretical explanation is straightforward: SSMs are designed with stable discrete-time dynamics via highly structured polynomial projection operator initialization. RKSP reveals this design choice explicitly in the spectral signature: Mamba exhibited $M_{<1} \approx 0.85 \gg M_{\approx 1} \approx 0.12$. This separation indicates strongly contractive dynamics with short memory and weak near-isometry; stability here comes from suppressed unstable modes in a highly contractive regime. In transformer regimes that are closer to near-normal, larger $M_{\approx 1}$ corresponds to weaker damping, longer-range signal retention, and higher instability risk.

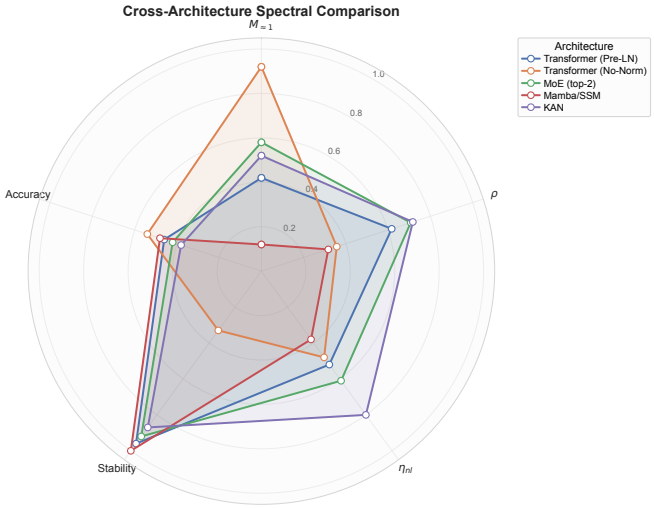

Figure 6: Cross-architecture spectral radar chart. Comparison of five architectures across five normalized metrics. Mamba exhibits strong contraction with low $M_{\approx 1}$ and short memory; stability is maintained via suppressed unstable modes, while in near-normal transformer regimes, higher $M_{\approx 1}$ aligns with more unstable, near-isometric propagation. The No-Norm transformer shows high memory capacity but poor stability. KAN exhibits high η_{nl}. Metrics are derived from Table 11.

Table 20: Mamba with RKSP analysis. Inherently stable spectral structure. Results use $L = 6$, 24 trials. Measured on the synthetic LM task with random-token next-token prediction.

Model	$M_{\approx 1}$	$M_{<1}$	ρ	Div. of 24	Acc.%
Mamba-Small, $d = 256$	0.12	0.85	0.95	0 of 24	48.2
Mamba-Medium, $d = 512$	0.15	0.82	0.97	0 of 24	52.6
Transformer, Pre-LN, comparable	0.42	0.38	1.85	1 of 24	45.8

KAN Table 21 reports KAN spectral diagnostics and KSS outcomes [Liu et al., 2024]. KAN shows high normalized linear-fit error: B-spline basis functions produce $\eta_{\text{nl}} \approx [0.78, 0.82]$, higher than the typical transformer layers with $[0.4, 0.7]$. Despite this high η_{nl}, RKSP remains informative—ResDMD filtering enables spectral analysis for 68% to 78% of modes. KSS benefits KAN with a $3\times$ to $4\times$ divergence reduction in this controlled setting, suggesting that spectral shaping may transfer beyond standard transformers.

Table 21: KAN transformer with RKSP analysis. The B-spline nonlinearity challenges linear approximation. The B-spline order is B. We use $d = 256$, $L = 6$, and 24 trials. Measured on the synthetic LM task with random-token next-token prediction.

Model	η_{nl}	$M_{\approx 1}$	ρ	Div. of 24	DMD reliability
KAN-Transformer, $B = 4$	0.78	0.52	2.15	3 of 24	Marginal, 68%
KAN-Transformer, $B = 8$	0.82	0.48	2.35	4 of 24	Low, 52%
KAN-Transformer + KSS	0.75	0.42	1.92	**1 of 24**	Improved, 78%

Cross-Architecture Summary Table 11 summarizes cross-architecture metrics, while Figure 6 provides a normalized radar-chart view of the same comparison.

K PRACTICAL NOTES

RKSP and KSS are most valuable in three scenarios. First, when mechanistic understanding matters, RKSP explains why Pre-LN outperforms Post-LN through spectral signatures. Second, when pushing training limits, KSS enables learning rates

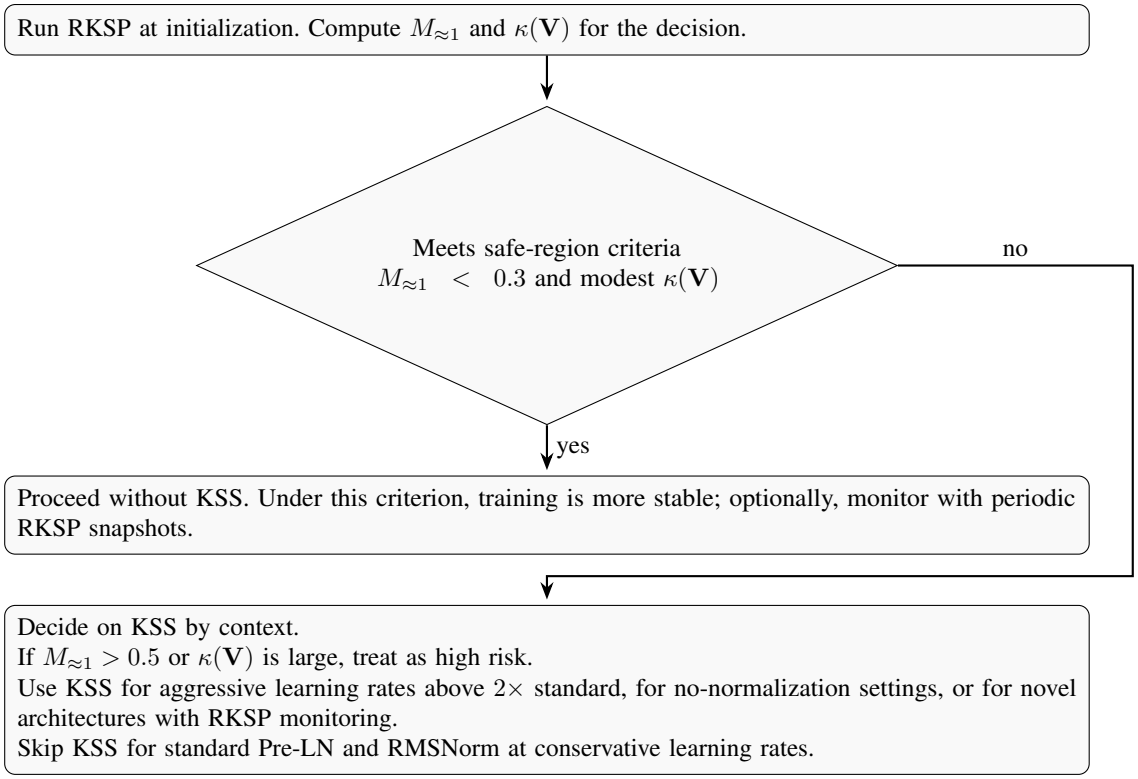

Figure 7: Decision flowchart for when to use RKSP and KSS in practice.

that are 50% to 150% higher for faster convergence. Third, when deploying novel architectures, RKSP screens stability risk before expensive training runs. Edge cases benefit most from these diagnostics—situations where standard normalization fails or where training operates near stability boundaries.

Fixup and ReZero-style identity initialization. A common stabilization trick in deep residual networks and transformers is to initialize the final projection of each residual branch to zero, for example the attention and MLP output weights, so that the network starts close to an identity map [Zhang et al., 2019, Bachlechner et al., 2021]. In our notation, this yields a residual-off regime with a vanishing layer update $\mathbf{h}_{\ell+1} - \mathbf{h}_\ell \approx \mathbf{0}$, so the snapshot pairs satisfy $\mathbf{Y}_\ell \approx \mathbf{X}_\ell$ and DMD returns $\hat{\mathbf{A}}_\ell \approx \mathbf{I}$. Consequently, $M^\ell_{\approx 1}$ can be close to 1 across layers even though training is often stable under Fixup and ReZero at initialization.

Taken alone, a near-identity spectrum might seem to imply maximal instability risk. However, our instability mechanism assumes two conditions: weak damping with large $M_{\approx 1}$ under near-normality, and non-degenerate layer-wise dynamics with appreciable updates so that perturbations and optimization noise are repeatedly injected and propagated across depth. Fixup and ReZero violate the second condition at initialization. When $\left\|\tilde{\mathbf{Y}}_\ell - \tilde{\mathbf{X}}_\ell\right\|_F$ is near zero, there is essentially no layer-wise update to analyze, and the resulting DMD spectrum is not informative about the noisy training-time regime we target.

Practically, this degeneracy is detectable from the same quantities RKSP already computes. When $\left\|\tilde{\mathbf{Y}}_\ell - \tilde{\mathbf{X}}_\ell\right\|_F \approx 0$, the normalization in the nonlinearity ratio Eq. 5 becomes ill-conditioned, so $\eta_{\mathrm{nl}}(\ell)$ should be interpreted as a DMD reliability flag rather than as a meaningful nonlinearity estimate. For Fixup and ReZero, RKSP becomes informative after a small amount of training, once the zero-initialized residual projections move away from zero and layer-wise updates become observable; at that point, RKSP can again capture whether the residual stream exhibits excessive near-isometric propagation (large $M_{\approx 1}$) that correlates with high-learning-rate divergence.

Practical deployment is straightforward. We recommend using RKSP in four scenarios: first, as a fast filter during architecture search; second, before expensive hyperparameter grid search; third, for periodic spectral monitoring during training; and fourth, for debugging checkpoints before divergence. Figure 7 provides an actionable decision process.