STATE SPACE DUALITY ATTENTION NON-COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

State space models are emerging as compelling linear-time alternatives to Transformers. Yet efficiency alone does not explain what fundamentally separates the two architectures: how do they organize and preserve information across a sequence? Through the *State Space Duality* (SSD) structured attention matrix, $\mathbf{M} = \mathbf{L} \odot \mathbf{QK}^\top$, we prove that bounded-memory SSD attention spreads information across a growing number of directions, and establish rank non-collapse for structured causal mixers. As a causal-softmax contrast, triangular masking alone does not force relative collapse, whereas a coherent early-token score sink can drive attention toward collapse. Together, these results separate relative rank collapse from absolute spectral degeneration and show that spectral concentration is governed jointly by mask structure and score geometry. Across context lengths, sink-dominated causal self-attention exhibits spectral *compression*, whereas SSD attention with decaying memory exhibits spectral *diffusion*.

1 INTRODUCTION

Transformers and State Space Models (SSMs) are two prominent approaches to sequence modeling. Their computational distinction is well known: dense self-attention mixes a sequence at quadratic cost, whereas recurrent SSM mixers can be evaluated in linear time. Computation alone, however, does not explain how these architectures distribute signal across token directions. We call concentration into a bounded number of dominant singular directions *spectral compression*. Do the structured causal operators used by linear-time models compress information in the same way, or do they spread it across many directions?

We investigate this question using the structured masked-attention view provided by *State Space Duality* (SSD) (Dao & Gu, 2024). At the level of token-to-token mixing, several scalar-transition recurrent mechanisms can be represented as $\mathbf{M} = \mathbf{L} \odot \mathbf{QK}^\top$, where $\mathbf{QK}^\top$ is a query-key score matrix, $\mathbf{L}$ is a structured causal mask, and $\odot$ denotes the Hadamard product. We focus on the causal all-one mask of linear attention (Katharopoulos et al., 2020), the 1-semiseparable mask associated with scalar Mamba transitions (Dao & Gu, 2024), and the Toeplitz decay mask used by RetNet retention (Sun et al., 2023).

The Hadamard product in structured attention can turn a low-rank score matrix into a full-rank causal operator. The relevant question is therefore spectral rather than algebraic: is the resulting mass carried by a few dominant directions or spread over a growing number of directions? To answer it precisely, we distinguish two notions of collapse. *Absolute spectral degeneration* means that an unrescaled empirical spectral law converges to δ_0. *Relative rank collapse* means that a scale-invariant rank proxy remains bounded as the sequence length grows. A matrix may exhibit the first behavior without the second. We use *spectral diffusion* when normalized spectral mass remains distributed over many directions, even if the overall scale shrinks.

Our first regime is a finite-rank proxy for trained or compressed query and key matrices. Because $\mathbf{Q} = XW_Q$ and $\mathbf{K} = XW_K$ share the same input, their token-side singular spaces may be strongly dependent. We model an extreme form of this dependence with a common isotropic token frame, while taking the projection-side singular spaces to be independent and isotropic. Under these assumptions, singular-value products of order $\ell\sqrt{d}$ yield a positive expected fraction of eigenvalues above any fixed magnitude threshold, regardless of the unit-diagonal causal mask. For singular values, bounded-memory masks yield nearly linear lower bounds on spectral rank proxies. Thus

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the effective dimension of attention can keep growing even when the query and key ranks remain uniformly bounded.

Our second regime is an idealized i.i.d. Gaussian query–key model motivated by random initialization. Its eigenvalue ranks diverge, while singular values distinguish the masks: when sequence length and projection dimension grow proportionally, causal linear attention has a nondegenerate limit, whereas the semiseparable and decay cases have an unrescaled law converging to δ_0. Nevertheless, all three scale-invariant rank proxies diverge at nearly linear rates, separating absolute degeneration from relative rank collapse.

As a complementary causal-softmax comparison, we show that triangular masking alone does not force relative rank collapse: bounded within-row logit variation yields diverging proxies, whereas a coherent early-token score sink over a diffuse background forces approximate rank-one collapse in Frobenius energy. This isolates a score-geometry mechanism on the compressor side and provides a contrasting benchmark for our main SSD non-collapse results.

1.1 Related Work

The starting point is the Transformer self-attention mechanism of Vaswani et al. (2017). A growing theoretical literature studies when attention matrices lose rank or spectral mass. Rank collapse across depth was identified by Dong et al. (2021) and further connected to signal propagation by Noci et al. (2022), and extended to SSM representations with lambda-skip prevention guarantees by Arangath Joseph et al. (2025); these works concern propagation across layers, whereas the collapse in width has also been studied in an i.i.d. Markov-matrix setting (Nait Saada et al., 2025). At initialization, Hayase et al. (2026) characterize the asymptotic spectral distribution of attention through Gaussian equivalence. Persistent concentration on initial tokens has also been documented as the attention-sink phenomenon and exploited for streaming inference (Xiao et al., 2024). Our causal-softmax result instead gives a sufficient score-level condition under which such a sink produces spectral concentration toward rank one. At the feature level, Shehper & Vaswani (2026) express unembedding LAR through weight and activation energy over singular directions, and associate concentration with generalization and diffusion with memorization or overfitting.

The quadratic sequence-length complexity of self-attention has motivated a large class of recurrent alternatives, even as implementations such as FlashAttention substantially reduce its memory traffic (Dao et al., 2022). Linear attention reorganizes causal attention into a recurrent computation (Katharopoulos et al., 2020). RetNet introduces head-dependent exponential retention (Sun et al., 2023), while selective SSMs and Mamba use input-dependent state transitions (Gu & Dao, 2024; Dao & Gu, 2024). State Space Duality and its algebraic extensions place scalar-transition SSM mixers in the same structured-matrix language as masked attention (Dao & Gu, 2024; Hu et al., 2026). The latter work studies representation and rank-explosion barriers; we instead derive sequence-length spectral non-collapse and quantitative bounds for the resulting attention operators. Toeplitz and Fourier mixers provide related alternatives (Qin et al., 2023; Lee-Thorp et al., 2022).

1.2 Structured Masked-Attention Model

Let ℓ be the sequence length, let d be the projection dimension, and let $\odot$ denote the Hadamard product. The SSD attention matrix $\mathbf{M}$ with causal mask $\mathbf{L}$ is defined as

$$\mathbf{M} = \mathbf{L} \odot \mathbf{Q}\mathbf{K}^\top, \qquad \mathbf{Q}, \mathbf{K} \in \mathbb{R}^{\ell \times d}, \quad \mathbf{L} \in \mathbb{R}^{\ell \times \ell}. \tag{1}$$

Here $\mathbf{Q}\mathbf{K}^\top$ is the query-key score matrix and $\mathbf{L}$ is lower triangular and unit diagonal. $\mathbf{L}$ depends on ℓ and $\mathbf{M}$ depends on ℓ and d. Instead of writing $\mathbf{L}_\ell$ and $\mathbf{M}_{\ell,d}$, we suppress the subscripts below.

1. **Causal linear attention** (Katharopoulos et al., 2020). The causal mask is
$$\mathbf{L}_{ij} = \mathbf{1}_{\{i \geq j\}}.$$
The corresponding $\mathbf{M}$ is the unnormalized causal score matrix. This baseline does not include the feature maps or row normalizers used by some linear-attention implementations.
2. **Mamba-2** (Dao & Gu, 2024). The 1-semiseparable mask is
$$\mathbf{L}_{ij} = \mathbf{1}_{\{i \geq j\}} \prod_{t=j+1}^{i} a_t, \qquad 0 \leq a_t \leq 1, \tag{2}$$

with the empty product equal to one. The bounded-memory condition used for the near-linear rank bounds is stated in Assumption 3. In the SSM notation, the query and key matrices $\mathbf{Q}, \mathbf{K}$ correspond to the $\mathbf{C}$- and $\mathbf{B}$-branches. In a selective model, the coefficients a_t may depend on the input and therefore may be statistically dependent on the token-side factors. Accordingly, the low-rank results apply conditionally on a realized mask only when the Haar assumptions continue to hold under that conditioning; in particular, the token frame must remain conditionally Haar and independent of the right-space factors. The i.i.d. Gaussian results instead impose explicit independence between the mask and the query–key factors in Assumption 5.

3. **Retentive Network (RetNet)** (Sun et al., 2023). The Toeplitz decay mask is

$$\mathbf{L}_{ij} = \mathbf{1}_{\{i \geq j\}} \gamma^{i-j}, \qquad 0 < \gamma < 1.$$

This is a special case of a stationary 1-semiseparable mask.

1.3 Contribution

We study the asymptotic spectral behavior of the structured attention matrix in SSD models and make the following main contributions:

- **Attention non-collapse in SSD models.** To the best of our knowledge, this is the first sequence-length spectral non-collapse theory for the structured token-mixing attention matrices arising in SSD models. We formulate it using spectral rank proxies and prove spectral diffusion under explicit low-rank and i.i.d. Gaussian query–key models. In the bounded-memory regimes, their effective dimension grows with sequence length rather than collapsing to a fixed number of directions.
- **A causal compressor–diffuser distinction.** We complement the SSD non-collapse results with a score-dependent analysis of causal softmax attention. Bounded within-row logit variation forces diverging rank proxies, whereas a coherent early-token score sink over a diffuse background forces approximate rank-one collapse in Frobenius energy. This provides a concrete conditional compression mechanism to contrast with SSD attention non-collapse: *a sufficiently strong, coherent attention sink induces energy-rank collapse.*
- **Two complementary stylized score regimes.** For the structured SSD mixers, we establish non-collapse under both a low-rank proxy for learned query–key representations and an idealized i.i.d. Gaussian query–key model. The finite-rank theory gives a mask-dependent stable-rank separation, while the Gaussian analysis shows that non-collapse persists for independent dense Gaussian factors and separates spectral shrinkage from rank collapse.

2 Theory of Rank Non-Collapse

We use stable rank (srank) and entropy-based effective ranks ($\mathrm{erank}_1, \mathrm{erank}_2$) (Roy & Vetterli, 2007) to characterize spectral concentration and rank non-collapse in attention matrices. Formal definitions, including the eigenvalue-based proxies, are given in Definition 1 in the appendix.

2.1 Low-Rank Model

We first study a finite-rank proxy motivated by trained query and key representations, with ranks bounded independently of the growing dimensions. The model isolates the interaction between low-rank scores and a structured causal mask. We write the singular value decompositions as

$$\mathbf{Q} = \mathbf{U}_Q \mathbf{\Sigma}_Q \mathbf{V}_Q^\top = \sum_{a=1}^{r_Q} \sigma_a^Q \mathbf{u}_a^Q (\mathbf{v}_a^Q)^\top, \quad \mathbf{K} = \mathbf{U}_K \mathbf{\Sigma}_K \mathbf{V}_K^\top = \sum_{b=1}^{r_K} \sigma_b^K \mathbf{u}_b^K (\mathbf{v}_b^K)^\top, \tag{3}$$

Here $1 \leq \mathrm{rank}(\mathbf{Q}) \leq r_Q$ and $1 \leq \mathrm{rank}(\mathbf{K}) \leq r_K$, where r_Q, r_K are fixed independently of ℓ, d. Singular values are ordered nonincreasingly and padded with zeros; the corresponding orthonormal singular vectors are completed to r_Q and r_K columns.

For linear projections $\mathbf{Q} = XW_Q$ and $\mathbf{K} = XW_K$, the shared input X directly couples the token-side geometry. By contrast, independent random initialization of W_Q and W_K motivates an isotropic

projection-side proxy. Training need not preserve this independence, so the following assumption should be understood as a tractable conditional proxy whose fit can be assessed empirically.

Assumption 1 (Independent Haar right spaces)**.** *Conditioned on* $(\mathbf{U}_Q, \mathbf{U}_K, \mathbf{\Sigma}_Q, \mathbf{\Sigma}_K)$ *(and the ranks and dimensions),* $\mathbf{V}_Q \sim \mathrm{Haar}(\mathrm{St}(d, r_Q))$, $\mathbf{V}_K \sim \mathrm{Haar}(\mathrm{St}(d, r_K))$ *are independent.*

Both $\mathbf{Q}$ and $\mathbf{K}$ are linear functions of the *same* input matrix X, so their column spaces are contained in the common ambient space $\mathrm{col}(X)$. This containment alone does not imply that their left singular vectors coincide. To isolate the maximally aligned form of token-side dependence, we impose the following stronger common-frame idealization. Throughout this low-rank model, the nonzero singular values are treated as deterministic parameters; equivalently, the statements may be read conditionally on them whenever the frame laws below continue to hold.

Assumption 2 (Common Haar token space)**.** *Let* $r := \max(r_Q, r_K)$*. The left singular vectors used by* $\mathbf{Q}$ *and* $\mathbf{K}$ *are columns of a common matrix* $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \sim \mathrm{Haar}(\mathrm{St}(\ell, r))$*, independent of the right singular spaces. The frame dimensions* r_Q, r_K*, and hence* r*, remain fixed in all asymptotic statements below. Thus* $\mathbf{u}_a^Q = \mathbf{u}_a, \mathbf{u}_b^K = \mathbf{u}_b$*.*

$\mathrm{Haar}(\mathrm{St}(n, r))$ denotes the uniform Stiefel law, equivalently the law of the first r columns of an $n \times n$ Haar orthogonal matrix. The common Haar frame removes any preferred token-space orientation while retaining maximal alignment between the left singular vectors of $\mathbf{Q}$ and $\mathbf{K}$. This is an isotropic modeling choice analogous to random-rotation assumptions in score-geometry and vector-compression analyses (Gao & Long, 2024; Gao et al., 2025; Zandieh et al., 2026).

The near-linear singular-value rank bounds use the following additional mask assumption, invoked explicitly where needed.

Assumption 3 (Causal Mask)**.** *We assume* $\sup_\ell \|\mathbf{L}\|_{\mathrm{op}} < \infty$*.*

- *For the RetNet decay mask, this condition always holds.*
- *For the Mamba-2* 1*-semiseparable mask, it is sufficient to assume that there exists* $a < 1$ *such that* $a_t \le a$ *for all* ℓ *and* t*. This can be weakened to uniform contraction over fixed-length blocks; the exact bounded-memory characterization is given in Proposition 10.*

2.1.1 Eigenvalue Rank Non-Collapse

It is easier to characterize the eigenvalues of a triangular matrix than its singular values. Since the masks considered here are lower triangular with unit diagonal, the eigenvalues of $\mathbf{M}$ are the diagonal entries of $\mathbf{QK}^\top$. We first characterize their laws and rank proxies and then compare these with the corresponding results for singular values.

Theorem 1 (Eigenvalue spectral phenomena)**.** *Assume that Assumptions 1 and 2 hold, and let* $\mathbf{L}$ *be unit-diagonal and lower triangular. For each* (ℓ, d)*, let* $\theta_{ab} := \frac{\sigma_a^Q \sigma_b^K}{\ell\sqrt{d}}$, $s_{\ell,d}^2 := \sum_{a=1}^{r_Q}\sum_{b=1}^{r_K}(1 + 2\mathbf{1}_{\{a=b\}})\theta_{ab}^2$*. Suppose that* $s_- := \liminf_{(\ell,d)\to(\infty,\infty)} s_{\ell,d} > 0$*. Let* $\overline{\Phi}(t) := \mathbb{P}\{Z > t\}$ *for* $Z \sim N(0,1)$*. Then, as* $(\ell, d) \to (\infty, \infty)$*:*

(i) Empirical-law nondegeneracy and unbounded support. For every $\varepsilon > 0$*,*

$$\liminf_{(\ell,d)\to(\infty,\infty)} \frac{\mathbb{E}\#\{i : |\lambda_i(\mathbf{M})| > \varepsilon\}}{\ell} \ge \frac{3}{70}\overline{\Phi}\left(\frac{\sqrt{2}\,\varepsilon}{s_-}\right) > 0. \tag{4}$$

If $s_- = \infty$*, interpret* $\varepsilon/s_- = 0$*.*

(ii) Eigenvalue spectral rank non-collapse.

$$\mathrm{srank}_\lambda(\mathbf{M}) = \Theta_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right), \qquad \mathrm{erank}_\lambda(\mathbf{M}) = \Theta_{\mathbb{P}}(\ell). \tag{5}$$

Thus the empirical eigenvalue law does not converge to δ_0 in probability. Appendix C.2 further characterizes the spectrum when the coefficient array $\boldsymbol{\theta}_{\ell,d} = (\theta_{ab})_{a,b}$ converges. The empirical law then converges in distribution to a nondegenerate random probability measure with almost surely

unbounded support, arising from Gaussian quadratic forms with a random coefficient matrix shared across tokens. Moreover, the total eigenvalue magnitude is of order ℓ in probability, whereas the largest magnitude is only of order $\log \ell$. This separation explains the $\ell/\log \ell$ scaling of srank_λ, while erank_λ remains proportional to ℓ.

The joint limit $(\ell, d) \to (\infty, \infty)$ in Theorem 1 is used to obtain the Gaussian limit of the projection-side Haar overlaps. By contrast, the scale-invariant singular-value rank bounds do not require a growing projection dimension and are stated for fixed d.

The natural nontrivial scale for a contributing singular-value pair is $\sigma_a^Q \sigma_b^K \asymp \ell\sqrt{d}$ (Theorem 9). At this scale, the expected fraction of eigenvalues with magnitude above any fixed threshold remains positive, while the eigenvalue spectral rank proxies grow with the sequence length ℓ.

2.1.2 Mask-Dependent Singular-Value Rank Behavior

Eigenvalues only see the diagonal of a lower triangular mask, while singular values see the full causal operator. The scale-invariant rank proxies remove the overall query–key amplitude, so their behavior can be controlled at arbitrary fixed query and key rank bounds without imposing an amplitude normalization. The mask gain $\kappa_\ell = \|\mathbf{L}\|_{\mathrm{op}}$ then determines the resulting rank spread.

Theorem 2 (Singular-value spectral rank non-collapse)**.** *Fix positive integers r_Q, r_K, d, and set $r := \max(r_Q, r_K) \le d$. Assume that Assumptions 1 and 2 hold for this family. As $\ell \to \infty$,*

(i) For a general unit-diagonal causal mask,

$$\mathrm{srank}(\mathbf{M}),\ \mathrm{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\kappa_\ell^2 (\log \ell)^2}\right), \qquad \mathrm{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\kappa_\ell \log \ell}\right). \quad (6)$$

(ii) Under Assumption 3, in particular for RetNet and Mamba-2,

$$\mathrm{srank}(\mathbf{M}),\ \mathrm{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{(\log \ell)^2}\right), \qquad \mathrm{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right). \quad (7)$$

(iii) For the causal linear mask,

$$\mathrm{srank}(\mathbf{M}) = \Theta_{\mathbb{P}}(1), \qquad \mathrm{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}(\log \ell). \quad (8)$$

All implicit constants may depend on the fixed r_Q, r_K and d.

For Mamba-2 and RetNet under the theorem's bounded-memory and Haar assumptions, all three rank proxies diverge even though the query and key ranks remain bounded, with lower bounds that are linear in ℓ up to logarithmic factors. Causal linear attention instead gives bounded stable rank alongside a diverging erank_1: a dominant energy direction can coexist with a broad tail of weak singular values.

Theorem 3 (Singular-value empirical-law nondegeneracy)**.** *In the setting of Theorem 2, define the normalized amplitude $\gamma_\ell := \|\mathbf{Q}\|_F \|\mathbf{K}\|_F / \ell\sqrt{d}$, $\gamma_- := \liminf_{\ell\to\infty} \gamma_\ell$. Suppose that $\gamma_- > 0$ and that there is a constant $C_{\mathrm{inv}} < \infty$ such that $\sup_\ell \sup_{\varnothing \neq S \subset \{1,\dots,\ell\}} \|\{(\mathbf{L})_S\}^{-1}\|_{\mathrm{op}} \le C_{\mathrm{inv}}$. Then there are deterministic constants $\varepsilon_0, c_{\mathrm{mass}}, p_{\mathrm{nd}} > 0$ such that*

$$\liminf_{\ell\to\infty} \mathbb{P}\left\{\frac{1}{\ell}\#\{j : s_j(\mathbf{M}) > \varepsilon_0\} \ge c_{\mathrm{mass}}\right\} \ge p_{\mathrm{nd}}. \quad (9)$$

By Proposition 11, the principal-inverse condition holds for the causal linear mask and for every nonexpansive semiseparable mask in equation 2, including the decay and bounded-memory families.

The uniform principal-inverse condition holds for all three mask families considered here: causal linear attention, Mamba-2, and RetNet. The rank-one statement in Theorem 10 can give sharper conditional counts and explicit rank-one constants.

CAUSAL SELF-ATTENTION

Causal self-attention (Vaswani et al., 2017) can exhibit energy-rank collapse under a sufficiently strong, coherent attention sink. Our experiments observe both sink concentration and low energy-sensitive rank in deeper GPT-2 layers.

Theorem 4 (Informal: Self-Attention). *Let* $\mathbf{A} = \mathrm{softmax}_{\mathrm{row}}^{\mathrm{causal}}(\mathbf{S})$ *be the self-attention matrix at sequence length* ℓ*, with the score matrix* $\mathbf{S} = \mathbf{Q}\mathbf{K}^\top/\sqrt{d}$*.*

(i) ***Non-collapse with bounded score oscillation.*** *If the accessible score oscillation within every causal row is uniformly bounded in* ℓ*, then*

$$\mathrm{srank}(\mathbf{A}) = \Theta(\log \ell), \qquad \mathrm{erank}_1(\mathbf{A}) = \Omega(\log \ell), \qquad \mathrm{erank}_2(\mathbf{A}) = \Omega(\log \ell).$$

Thus causal masking alone does not force relative rank collapse.

(ii) ***Collapse under an attention sink.*** *If the first score column forms a coherent causal sink—for example,* $S_{i1} - \max_{2 \le j \le i} S_{ij} \ge c \log i - O(1)$ *for some* $c > 1/2$*, with a diffuse non-sink background—then*

$$\mathrm{srank}(\mathbf{A}) \longrightarrow 1, \qquad \mathrm{erank}_2(\mathbf{A}) \longrightarrow 1, \qquad \min_{\mathrm{rank}(\mathbf{R}) \le 1} \frac{\|\mathbf{A} - \mathbf{R}\|_F}{\|\mathbf{A}\|_F} \longrightarrow 0.$$

This is approximate rank-one collapse in Frobenius energy: finite causal softmax matrices remain algebraically full rank, and collapse of erank_1 *requires stronger singular-value-tail control.*

The precise statements are given in Theorems 11 and 12, Appendix D, with the score-level sink condition in Corollary 2. An SVD-oriented sufficient condition is given in Corollary 3, while Remark 10 explains why singular values alone cannot determine causal rank collapse.

2.2 I.I.D. GAUSSIAN QUERY–KEY MODEL

We now return to the structured SSD model. The finite-rank regime above is a useful proxy for learned score geometry; at random initialization, however, the query and key matrices need not have uniformly bounded rank. Following the orthogonal-input Gaussian models used in Nait Saada et al. (2025); Hayase et al. (2026), we study the simplest dense-factor idealization: independent query and key matrices with i.i.d. Gaussian entries at the usual projection scale.

Assumption 4 (Independent isotropic Gaussian query–key factors). *Fix scale parameters* $\sigma_Q, \sigma_K > 0$*. The rows* $\mathbf{q}_i^\top$ *of* $\mathbf{Q}$ *and* $\mathbf{k}_i^\top$ *of* $\mathbf{K}$ *are independent over* i*, and* $\mathbf{q}_i \sim N\left(0, \frac{\sigma_Q^2}{d}\mathbf{I}_d\right), \mathbf{k}_i \sim N\left(0, \frac{\sigma_K^2}{d}\mathbf{I}_d\right)$*, with* $\mathbf{Q}$ *independent of* $\mathbf{K}$*. The mask* $\mathbf{L}$ *is lower triangular and unit-diagonal.*

Remark 1 (Orthogonal-input interpretation and scope). For linear projections $\mathbf{Q} = XW_Q$ and $\mathbf{K} = XW_K$, if X is orthogonal and W_Q, W_K are appropriately scaled independent Ginibre matrices, then left orthogonal invariance gives $XW_Q \overset{d}{=} W_Q$ and $XW_K \overset{d}{=} W_K$. Assumption 4 is therefore exact in this idealized construction and provides a tractable reference for isolating the effect of the causal mask. It is not a literal model of Mamba-2 or RetNet initialization: shared inputs, normalization, and causal convolutions induce dependence across factors and tokens.

In the i.i.d. Gaussian model, each eigenvalue is a sum of independent Gaussian products, which makes the derivation of the spectral-rank proxies straightforward (see Section B.1).

Theorem 5 (Gaussian-model eigenvalue rank proxies). *Under Assumption 4, the eigenvalues* $\lambda_i(\mathbf{M}) = \langle \mathbf{q}_i, \mathbf{k}_i \rangle$ *are independent and, for each fixed token,* $\frac{\sqrt{d}}{\sigma_Q \sigma_K}\lambda_i(\mathbf{M}) \Rightarrow N(0,1)$ *when* $d \to \infty$*.*

For each fixed d*, as* $\ell \to \infty$*,*

$$\mathrm{srank}_\lambda(\mathbf{M}) = \Theta_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right), \qquad \mathrm{erank}_\lambda(\mathbf{M}) = \Theta_{\mathbb{P}}(\ell). \tag{10}$$

Unlike eigenvalues, singular values depend on the off-diagonal entries of the causal mask. We use a uniformly decaying deterministic proxy and an i.i.d. logistic-normal proxy for semiseparable transitions, together with a fixed exponential-decay proxy.

Assumption 5 (Semiseparable and exponential-decay coefficients)**.** *The mask coefficients are independent of* $\mathbf{Q}$ *and* $\mathbf{K}$*. In the* 1*-semiseparable case, impose one of the following two alternatives:*

(a) the coefficients $(a_t)_t$ *are logistic-normal,* $a_t = \frac{1}{1+\exp(-Z_t)}, Z_t \overset{\text{i.i.d.}}{\sim} N(0,1)$;

(b) the coefficients are deterministic and uniformly decaying: there is a fixed $a_{\max} < 1$*, independent of* ℓ *and* d*, such that* $0 \le a_t \le a_{\max}$ *for every* t*.*

In the Toeplitz-decay case, $L_{ij} = \mathbf{1}_{\{i \ge j\}} \gamma^{i-j}$*, where* $\gamma \in (0,1)$ *is fixed as* $\ell, d \to \infty$*.*

The deterministic alternative has uniformly bounded cumulative memory. The logistic-normal alternative instead has realized row and column memory sums of order $O(\log \ell)$ with probability tending to one, while $\mathbb{E}\|\mathbf{L}\|_F^2 = O(\ell)$. These estimates, together with the more general fixed-block extension, are proved in Appendix B.

The following criterion isolates when the unrescaled squared-singular-value limiting spectral distribution (LSD) is degenerate. It depends on the average squared mask energy, rather than on a particular aspect ratio.

Theorem 6 (Degenerate Gaussian-model LSD under vanishing mask energy)**.** *Under Assumption 4, fix positive* σ_Q, σ_K*. Let* $\mathbf{L}$ *be deterministic or independent of* $(\mathbf{Q}, \mathbf{K})$*, and let* $\ell \to \infty$ *while* $d = d_\ell$ *may vary. Suppose* $\|\mathbf{L}\|_F^2/(d\ell) \xrightarrow{\mathbb{P}} 0$*. Let* $\nu_\ell^{(2)} := \frac{1}{\ell}\sum_{j=1}^{\ell} \delta_{s_j(\mathbf{M})^2}$ *be the empirical measure of squared singular values of* $\mathbf{M}$*. Then*

$$\nu_\ell^{(2)} \Rightarrow \delta_0$$

in probability. Under Assumption 5, this applies to both 1*-semiseparable alternatives and the fixed-*γ *Toeplitz-decay mask as* $\ell, d \to \infty$*, with no restriction on* ℓ/d*, since* $\mathbb{E}\|\mathbf{L}\|_F^2 = O(\ell)$*.*

There is a complementary boundary case for the causal-linear mask. When $\ell/d \to \tau \in (0, \infty)$, its empirical squared-singular-value distribution converges to a nondegenerate compactly supported law, whose moments are given in Appendix B. Thus the δ_0 limit for semiseparable and decay masks is a mask-energy effect, rather than a generic consequence of Gaussian query–key factors.

Theorem 7 (Gaussian-model singular-value rank bounds)**.** *Under Assumption 4, fix positive* σ_Q, σ_K *and assume that* $\ell, d \to \infty$ *and* $\ell/d \to \tau \in (0, \infty)$*. Any random mask below is assumed independent of* $\mathbf{Q}$ *and* $\mathbf{K}$*. Then the following rank bounds hold.*

(i) For a general unit-diagonal causal mask, with $\kappa_\ell := \|\mathbf{L}\|_{\text{op}}$*,*

$$\operatorname{srank}(\mathbf{M}),\ \operatorname{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\kappa_\ell^2 \log \ell}\right), \qquad \operatorname{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\kappa_\ell \sqrt{\log \ell}}\right). \quad (11)$$

For random κ_ℓ*, interpret these bounds in the equivalent product formulation.*

(ii) Under Assumption 5, for deterministic 1*-semiseparable and Toeplitz-decay masks,*

$$\operatorname{srank}(\mathbf{M}),\ \operatorname{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right), \qquad \operatorname{erank}_1(\mathbf{M}) = \Theta_{\mathbb{P}}(\ell). \quad (12)$$

For random $(a_t)_t$*,* $\operatorname{srank}(\mathbf{M}), \operatorname{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}(\ell/(\log \ell)^2), \operatorname{erank}_1(\mathbf{M}) = \Theta_{\mathbb{P}}(\ell)$*.*

(iii) For the causal linear mask, for every fixed $\varepsilon > 0$*,*

$$\operatorname{srank}(\mathbf{M}) = \Omega_{\mathbb{P}}(\ell^{1-\varepsilon}), \qquad \operatorname{erank}_1(\mathbf{M}) = \Theta_{\mathbb{P}}(\ell), \qquad \operatorname{erank}_2(\mathbf{M}) = \Theta_{\mathbb{P}}(\ell). \quad (13)$$

Remark 2 (Degenerate LSD versus non-collapsing rank). The convergence $\nu_\ell^{(2)} \Rightarrow \delta_0$ is an absolute-scale statement, whereas stable and effective ranks are scale invariant. For example, $\ell^{-1/2}\mathbf{I}_\ell$ has squared-singular-value law $\delta_{1/\ell} \Rightarrow \delta_0$, but all three rank proxies equal ℓ. Theorems 6 and 7 realize this distinction: many singular values may shrink while their normalized mass remains spread over a diverging number of directions. The bounds establish non-collapse, not a cross-mask rank ordering.

3 EXPERIMENTS

We test whether trained causal token mixers spread singular-value mass across an increasing number of token directions as context grows. By reconstructing the attention mixing matrices of GPT-2 (Radford et al., 2019), Based-360M (Arora et al., 2024), Mamba-Codestral-7B, and RetNet-1.3B on nested prefixes from 32 document-disjoint WikiText-103 validation articles, we report three scale-invariant ranks: stable rank and erank_2 emphasize concentration of spectral energy, whereas erank_1 also counts a broad tail of weak singular directions. Exact datasets, checkpoints, layers, uncertainty estimates, and additional mechanism diagnostics are described task by task in Appendix A.

Rank spreading in trained mixers. Figure 1 is our main empirical result and reveals a checkpoint- and depth-dependent separation. Mamba-2 and RetNet show growth in all three rank proxies at every selected depth, consistent respectively with the gain-dependent and bounded-decay non-collapse regimes of Theorem 2 (i)–(ii). GPT-2 instead moves from an early layer where all ranks grow to middle and final layers with stable rank near 1 and nearly constant erank_2 as the first-column sink strengthens, matching the non-collapse and sink-induced energy-collapse regimes of Theorem 4. The final raw Based operator similarly has order-one stable rank but growing erank_1, as predicted for the causal-linear mask by Theorem 2(iii); row normalization preserves the same pattern. The Mamba-2 and RetNet trends also replicate on Wikipedia. We interpret these as finite-context correspondences, not verification of the asymptotic assumptions; layerwise results and hypothesis audits are in Appendix A, Tasks 2–5 and 7.

Isolating the mask contribution. The raw cross-model comparison changes both the mask and the learned score factors. We therefore hold the document, prefix, layer, component, and learned factors fixed, and replace only the realized Mamba-2 forgetting mask or RetNet decay mask by the undiscounted causal all-one mask. Removing forgetting or decay lowers all three rank proxies for both checkpoints, with every document-bootstrap interval excluding zero; the same direction repeats on the separate Wikipedia sample. Thus, under the realized learned score geometry, replacing the native forgetting or decay mask by an all-one causal mask produces a more dominant singular direction. This controlled intervention supports a mask-dependent mechanism at the operator level, while remaining a conditional comparison rather than a full model intervention. Details appear in Appendix A, Tasks 6–7.

Causal-softmax contrast. For GPT-2, the first-column sink approximation is nearly as accurate as the optimal rank-one approximation, and neutralizing the early-column score advantage broadens the energy-sensitive spectrum. These observations are compatible with the score-sink mechanism proposed by our causal-softmax theory. The effect depends on the boundary token, and the stronger asymptotic sink condition is not observed over the tested lengths. We therefore interpret the experiment as finite-context evidence compatible with the mechanism, not as a verification of its asymptotic condition; see Appendix A, Task 3.

Overall, the experiments support the proposed distinction, as context length grows, between spectral spreading in structured SSD mixers and energy concentration in sink-dominated GPT-2 attention.

4 CONCLUSION

We have identified attention non-collapse as a characteristic spectral phenomenon of the structured causal mixers in State Space Duality models. Their structured attention $\mathbf{M} = \mathbf{L} \odot \mathbf{Q}\mathbf{K}^\top$ reveals how a causal mask can spread even low-rank query–key signal across a growing number of token directions. Our results yield a mechanism-level spectral compressor–diffuser distinction between sequence mixers. For causal softmax, the outcome is score dependent: bounded rowwise score oscillation prevents relative collapse, while a coherent early-token sink can compress Frobenius energy toward rank one. The structured SSD and retention masks in our finite-rank theory instead preserve broad spectral support. These complementary roles make existing hybrid architectures that combine SSM-style mixers with Transformer self-attention in large language models (LLMs) a natural setting in which to study composed signal propagation (NVIDIA et al., 2025; 2026). Signal propagation through hybrid SSM–attention stacks remains an open problem.

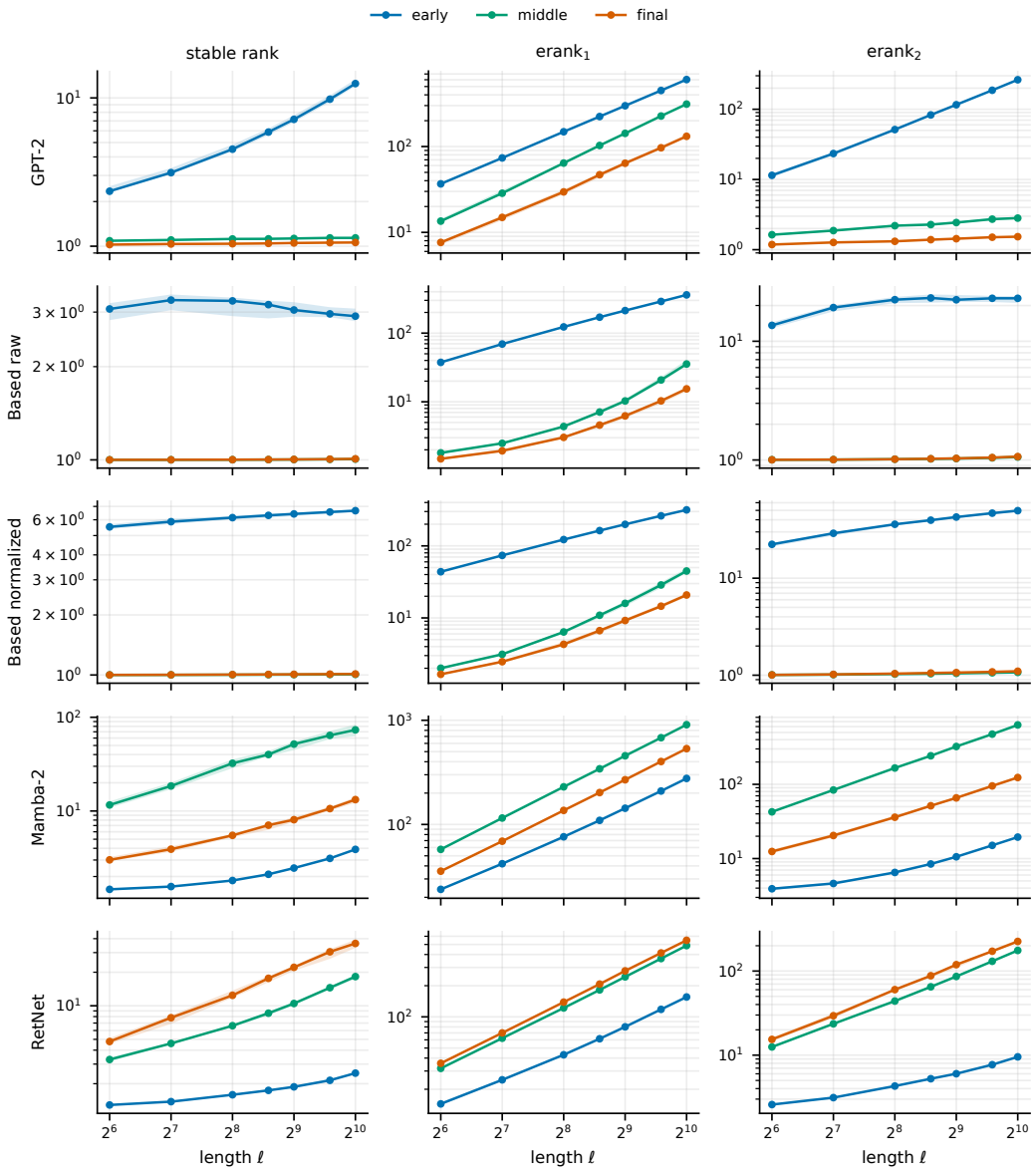


Figure 1: Rank of the trained token-mixing operators as context length grows. Stable rank and erank_2 are most sensitive to a dominant energy direction; erank_1 also counts weak singular directions. Mamba-2 and RetNet grow on all three measures, whereas the final GPT-2 and Based operators retain an almost constant energy-sensitive rank. Lines summarize 32 WikiText-103 documents and bands are 95% document-bootstrap intervals.

An open direction is to extend this spectral perspective to gated delta-rule linear attention (Yang et al., 2024; 2025; Qiu et al., 2025; Kimi et al., 2025; 2026), which appears in recent LLMs. Their key-dependent matrix transitions replace a fixed scalar causal mask with products of content-dependent operators, enabling selective retention and overwriting. A central question is how these transition products determine spectral diffusion, memory capacity, and the preservation or erasure of long-range information.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, refine hypotheses, design experiments, implement methods, interpret results and provide feedback on research methodology. We have not used generative AI tools to generate synthetic datasets, assist with translation, clean or reformat datasets, or support qualitative and thematic data analysis. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code for our experiments is released at `https://anonymous.4open.science/r/mamba-B492`. We use the `Salesforce/wikitext` dataset (Merity et al., 2017), from Hugging Face `https://huggingface.co/datasets/Salesforce/wikitext`. The cross-corpus replication uses `wikimedia/wikipedia` (revision `20231101.en`), from `https://huggingface.co/datasets/wikimedia/wikipedia`.

REFERENCES

Federico Arangath Joseph, Jerome Sieber, Melanie Zeilinger, and Carmen Amo Alonso. Lambda-skip connections: the architectural component that prevents rank collapse. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 11871–11896, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1f7b3b0dd7710af02aac0db5be4cfc8d-Paper-Conference.pdf.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Re. Simple linear attention language models balance the recall-throughput tradeoff. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1763–1840. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/arora24a.html.

Benoît Collins and Piotr Śniady. Integration with respect to the haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006. ISSN 1432-0916. doi: 10.1007/s00220-006-1554-3. URL https://doi.org/10.1007/s00220-006-1554-3.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with io-awareness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/67d57c32e20fd0a7a302cb81d36e40d5-Paper-Conference.pdf.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

Jianyang Gao and Cheng Long. RaBitQ: Quantizing High-Dimensional Vectors with a Theoretical Error Bound for Approximate Nearest Neighbor Search. In *Proceedings of the ACM on Management of Data*, volume 2, New York, NY, USA, May 2024. Association for Computing Machinery. doi: 10.1145/3654970. URL https://doi.org/10.1145/3654970.

Jianyang Gao, Yutong Gou, Yuexuan Xu, Yongyi Yang, Cheng Long, and Raymond Chi-Wing Wong. Practical and asymptotically optimal quantization of high-dimensional vectors in euclidean space for approximate nearest neighbor search. In *Proceedings of the ACM on Management of Data*, volume 3, New York, NY, USA, June 2025. Association for Computing Machinery. doi: 10.1145/3725413. URL https://doi.org/10.1145/3725413.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Tomohiro Hayase, Benoit Collins, and Ryo Karakida. Gaussian equivalence for self-attention: Asymptotic spectral analysis of attention matrix. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=ABbVGvqaj3.

Jerry Yao-Chieh Hu, Xiwen Zhang, Ali ElSheikh, Weimin Wu, and Han Liu. On structured state-space duality. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DKathyl3XN.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Kimi, : Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, Wentao Li, Enzhe Lu, Weizhou Liu, Yanru Chen, Weixin Xu, Longhui Yu, Yejie Wang, Yu Fan, Longguang Zhong, Enming Yuan, Dehao Zhang, Yizhi Zhang, T. Y. Liu, Haiming Wang, Shengjun Fang, Weiran He, Shaowei Liu, Yiwei Li, Jianlin Su, Jiezhong Qiu, Bo Pang, Junjie Yan, Zhejun Jiang, Weixiao Huang, Bohong Yin, Jiacheng You, Chu Wei, Zhengtao Wang, Chao Hong, Yutian Chen, Guanduo Chen, Yucheng Wang, Huabin Zheng, Feng Wang, Yibo Liu, Mengnan Dong, Zheng Zhang, Siyuan Pan, Wenhao Wu, Yuhao Wu, Longyu Guan, Jiawen Tao, Guohong Fu, Xinran Xu, Yuzhi Wang, Guokun Lai, Yuxin Wu, Xinyu Zhou, Zhilin Yang, and Yulun Du. Kimi linear: An expressive, efficient attention architecture, 2025. URL https://arxiv.org/abs/2510.26692.

Kimi, : Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu,

Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. FNet: Mixing tokens with Fourier transforms. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.319. URL https://aclanthology.org/2022.naacl-main.319/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45561–45587. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/nait-saada25a.html.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27198–27211. Curran Associates, Inc.,

2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ae0cba715b60c4052359b3d52a2cff7f-Paper-Conference.pdf.

NVIDIA, : Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, Ameya Sunil Mahabaleshwarkar, Andrew Tao, Anna Shors, Ashwath Aithal, Ashwin Poojary, Ayush Dattagupta, Balaram Buddharaju, Bobby Chen, Boris Ginsburg, Boxin Wang, Brandon Norick, Brian Butterfield, Bryan Catanzaro, Carlo del Mundo, Chengyu Dong, Christine Harvey, Christopher Parisien, Dan Su, Daniel Korzekwa, Danny Yin, Daria Gitman, David Mosallanezhad, Deepak Narayanan, Denys Fridman, Dima Rekesh, Ding Ma, Dmytro Pykhtar, Dong Ahn, Duncan Riach, Dusan Stosic, Eileen Long, Elad Segal, Ellie Evans, Eric Chung, Erick Galinkin, Evelina Bakhturina, Ewa Dobrowolska, Fei Jia, Fuxiao Liu, Gargi Prasad, Gerald Shen, Guilin Liu, Guo Chen, Haifeng Qian, Helen Ngo, Hongbin Liu, Hui Li, Igor Gitman, Ilia Karmanov, Ivan Moshkov, Izik Golan, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jarno Seppanen, Jason Lu, Jason Sewall, Jiaqi Zeng, Jiaxuan You, Jimmy Zhang, Jing Zhang, Jining Huang, Jinze Xue, Jocelyn Huang, Joey Conway, John Kamalu, Jon Barker, Jonathan Cohen, Joseph Jennings, Jupinder Parmar, Karan Sapra, Kari Briski, Kateryna Chumachenko, Katherine Luna, Keshav Santhanam, Kezhi Kong, Kirthi Sivamani, Krzysztof Pawelec, Kumar Anik, Kunlun Li, Lawrence McAfee, Leon Derczynski, Lindsey Pavao, Luis Vega, Lukas Voegtle, Maciej Bala, Maer Rodrigues de Melo, Makesh Narsimhan Sreedhar, Marcin Chochowski, Markus Kliegl, Marta Stepniewska-Dziubinska, Matthieu Le, Matvei Novikov, Mehrzad Samadi, Michael Andersch, Michael Evans, Miguel Martinez, Mike Chrzanowski, Mike Ranzinger, Mikolaj Blaz, Misha Smelyanskiy, Mohamed Fawzy, Mohammad Shoeybi, Mostofa Patwary, Nayeon Lee, Nima Tajbakhsh, Ning Xu, Oleg Rybakov, Oleksii Kuchaiev, Olivier Delalleau, Osvald Nitski, Parth Chadha, Pasha Shamis, Paulius Micikevicius, Pavlo Molchanov, Peter Dykas, Philipp Fischer, Pierre-Yves Aquilanti, Piotr Bialecki, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Rabeeh Karimi, Rahul Kandu, Ran El-Yaniv, Raviraj Joshi, Roger Waleffe, Ruoxi Zhang, Sabrina Kavanaugh, Sahil Jain, Samuel Kriman, Sangkug Lym, Sanjeev Satheesh, Saurav Muralidharan, Sean Narenthiran, Selvaraj Anandaraj, Seonmyeong Bak, Sergey Kashirsky, Seungju Han, Shantanu Acharya, Shaona Ghosh, Sharath Turuvekere Sreenivas, Sharon Clay, Shelby Thomas, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shyamala Prayaga, Siddhartha Jain, Sirshak Das, Slawek Kierat, Somshubra Majumdar, Song Han, Soumye Singhal, Sriharsha Niverty, Stefania Alborghetti, Suseella Panguluri, Swetha Bhendigeri, Syeda Nahida Akter, Szymon Migacz, Tal Shiri, Terry Kong, Timo Roman, Tomer Ronen, Trisha Saar, Tugrul Konuk, Tuomas Rintamaki, Tyler Poon, Ushnish De, Vahid Noroozi, Varun Singh, Vijay Korthikanti, Vitaly Kurin, Wasi Uddin Ahmad, Wei Du, Wei Ping, Wenliang Dai, Wonmin Byeon, Xiaowei Ren, Yao Xu, Yejin Choi, Yian Zhang, Ying Lin, Yoshi Suhara, Zhiding Yu, Zhiqi Li, Zhiyu Li, Zhongbo Zhu, Zhuolin Yang, and Zijia Chen. Nemotron-H: A family of accurate and efficient hybrid mamba-transformer models, 2025. URL https://arxiv.org/abs/2504.03624.

NVIDIA, : Aaron Blakeman, Aaron Thomas, Aastha Jhunjhunwala, Abhibha Gupta, Abhinav Khattar, Adam Rajfer, Adi Renduchintala, Adil Asif, Aditya Vavre, Adriana Flores Miranda, Ahmad Bilal, Aileen Zaman, Ajay Hotchandani, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Alex Gronskiy, Alex Kondratenko, Alex Steiner, Alex Ye, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alice Gatti, Alisa Liu, Alok Kumar, Amar Phanishayee, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Anahita Bhiwandiwalla, Ananth Subramaniam, Andrea Santilli, Andrew Fulks, Andrew McHarg, Andrew Tao, Andrii Skliar, Anjulie Agrusa, Ankur Srivastava, Ankur Verma, Anna Shors, Anna Warno, Antoni-Joan Solergibert I Llaquet, Arham Mehta, Arkadiusz Nowaczynski, Arti Jain, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asit Mishra, Asma Kuriparambil Thekkumpate, Atefeh Sohrabizadeh, Avinash Kaur, Avinash Vem, Ayush Dattagupta, Barath Subramaniam Anandan, Bardiya Sadeghi, Ben Lanir, Benedikt Schifferer, Besmira Nushi, Bilal Kartal, Bill Thiede, Bita Darvish Rouhani, Bo Deng, Bob Schatz, Boris Ginsburg, Boxin Wang, Brad Nemire, Brandon Norick, Brian Dang, Brian Westphal, Brian Yu, Brucek Khailany, Bryan Catanzaro, Carlo del Mundo, Caryln Aarish, Chankyu Lee, Chantal Hwang, Charbel Sakr, Charles Wang, Charlie Truong, Chen Cui, Cheng Cheng, Cheng-Ping Hsieh, Chenghao Zhang, Chenhui Deng, Chintan Patel, Chris Alexiuk, Christian Cosgrove, Christian Munley, Christine Harvey, Christopher Parisien, Chunyang Shen, Coco Li, Collin Neale, Cynthia Gao, Cyril Meurillon, Dan Gil, Dan Su, Dan Zhao, Dane Corneil, Daniel Afrimi, Daniel Egert, Daniel Korzekwa, Daniel Lo, Daniel Machlab, Daniel Serebrenik, Daniil Sorokin, Daria Gitman, Daria Levy, Darko Stosic, David Mosallanezhad, David Yu, Davit

Karamyan, Deena Donia, Deep Debroy, Deepak Narayanan, Devin O'Kelly, Dheeraj Peri, Dhruv Nathawani, Di, Wu, Dima Rekesh, Divyanshu Kakwani, Donald Plummer, Dong Anh, Dongfeng Yu, Dongfu Jiang, Donnie Kim, Dorrin Poorkay, Duncan Riach, Dusan Stosic, Dustin VanStee, Eavan Meng, Edgar Minasyan, Edward Lin, Eileen Margaret Peters Long, Elad Sarafin, Elad Segal, Elena Lantz, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper, Eric Pham-Hung, Eric Tramel, Eric Yang, Erick Galinkin, Erik Pounds, Erika Goncalves Goncalves, Evan Briones, Evan Wu, Evelina Bakhturina, Evgeny Tsykunov, Ewa Dobrowolska, Faisal Ladhak, Farzan Memarian, Fay Wang, Fei Jia, Felipe Soares, Felipe Vieira Frujeri, Feng Chen, Fengguang Lin, Ferenc Galko, Frank Sun, Frankie Siino, Frida Hou, Gal Hubara Agam, Gal Kaplun, Gantavya Bhatt, Gargi Prasad, Garvit Kulshreshtha, George Armstrong, Gerald Shen, Giulio Borghesi, Gordana Neskovic, Gorkem Batmaz, Grace Lam, Greg Mason, Greg Pauloski, Grigor Nalbandyan, Grzegorz Chlebus, Grzegorz Karch, Guan-Ting Liu, Guoming Zhang, Guyue Huang, Haggai Maron, Haifeng Qian, Haim Elisha, Haoxing Ren, Haran Kumar Shiv Kumar, Haribhau Hud, Harris Nover, Harrison Saturley Hall, Hayate Iso, Helen Ngo, Herbert Hum, Herman Sahota, Hexin Wang, Himanshu Soni, Hovhannes Tamoyan, Hua Li, Huanhuan Chen, Hui Li, Hui Wang, Huy Nguyen, Ian Chiles, Ido Galil, Ido Shahaf, Igor Gitman, Igor Shovkun, Ilya Loshchilov, Ingo Guehring, Itamar Schen, Itay Levy, Itay Neeman, Ivan Moshkov, Izik Golan, Izzy Putterman, Jaemin Choi, Jakub Slowikowski, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jiacheng Xu, Jiafan Zhu, Jialin Song, Jian Zhang, Jiantao Jiao, Jiaqi Zeng, Jie Lou, Jim King, Jimmy Zhang, Jingquan Wang, Jinhang Choi, Jinju Chu, Joey Conway, Joey Guman, Johan Jatko, Johannes Rausch, John Kamalu, John Roberts, Johnny Greco, Johnny Mensel, Jonah Alben, Jonas Yang, Jonathan Cohen, Jonathan Raiman, Joseph Jennings, Joshua Mabry, Joshua Pierce, Joyjit Daw, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kajal Jain, Kan Zhu, Kari Briski, Katherine Cheung, Katherine Luna, Keith Willowhawk, Keith Wyss, Keshav Santhanam, Kevin Shih, Kezhi Kong, Khanh Nguyen, Khushi Bhardwaj, Kirthi Shankar Sivamani, Konstantinos Krommydas, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Kyle Keprios, Kylie Day, Lawrence McAfee, Leo Du, Leon Derczynski, Li Ding, Linda Liu, Lingjie Wu, Lior Kadoch, Lizzie Wei, Luis Vega, Luke Robison, Lun Su, Maarten Van Segbroeck, Maciej Jakub Mikulski, Maer Rodrigues de Melo, Magda Sypula, Mahan Fathi, Makesh Narsimhan Sreedhar, Makesh Tarun Chandran, Manoj Kilaru, Maor Ashkenazi, Marc Cuevas, Marc Romeijn, Marcin Chochowski, Mark Cai, Mark Mozolewski, Markus Kliegl, Marta Stepniewska-Dziubinska, Martyna Patelka, Mattei Machczynski, Matvei Novikov, Mauricio Ferrato, Maximilian Golub, Mehrzad Samadi, Melissa Corpuz, Mengru Wang, Mengxi Wu, Meredith Price, Meriem Boubdir, Micah Schaffer, Michael Andersch, Michael Boone, Michael Gschwind, Michael Lightstone, Michael Loh, Michal Bien, Michal Zawalski, Michelle Gill, Miguel Martinez, Mikail Khona, Mike Chrzanowski, Mike Houston, Mingyuan Ma, Minseok Lee, Mohamed Fawzy, Mohammad Dabbah, Mohammad Shoeybi, Mostofa Patwary, Nabin Mulepati, Najeeb Nabwani, Namit Dhameja, Narimane Hennouni, Natalie Hereth, Nathaniel Pinckney, Nave Algarici, Nave Assaf, Netanel Haber, Nicholas Knight, Nick Reamaroon, Nickson Quak, Nidhi Bhatia, Nikhil Desai, Nikolai Ludwig, Nima Tajbakhsh, Ning Xu, Nir Ailon, Nirmal Juluru, Nitin Nitin, Ofri Masad, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivia Viessmann, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Omri Puny, Oren Tropp, Pablo Ribalta, Pallab Bhattacharya, Panos Lampropoulos, Parth Mannan, Pasha Shamis, Patrick Legresley, Paul Gibbons, Pavlo Molchanov, Pawel Morkisz, Peter Dykas, Peter Jin, Pierre-Yves Aquilanti, Pinky Xu, Piotr Januszewski, Piotr Laskiewicz, Pooya Jannaty, Prakash Gurumurthy, Pranav Prashant Thombre, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Puhui Meng, Qiyu Wan, Rabeeh Karimi Mahabadi, Rachel Oberman, Rachit Garg, Radha Sri-Tharan, Rahul Kandu, Rakshit Sanadhya, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Ray Macalisang, Rayen Tian, Reka Kovacs, Renjie Pi, Rick Izzo, Rima Shahbazyan, Rishabh Garg, Rishi Puri, Rita Fernandes Neves, Ritchie Zhao, Ritika Borkar, Ritu Gala, Riyad Islam, Robert Clark, Robert Hesse, Robert Kirby, Roger Waleffe, Rohit Watve, Roi Koren, Ron Banner, Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Ryan Stewart, Ryota Egashira, Sadegh Mahdavi, Saee Paliwal, Sagar Singh, Sahil Modi, Salika Dave, Samantha Shinagawa, Samuel Kriman, Sandip Bhaskar, Sangkug Lym, Sanjay Kariyappa, Sanjeev Satheesh, Saran Vikas Murari, Satish Pasumarthi, Saurabh Mishra, Saurav Muralidharan, Scott Hara, Sean Narentharen, Selvaraj Anandaraj, Seonjin Na, Seonmeyong Bak, Seonmyeong Bak, Sepehr Sameni, Seph Mard, Serge Panev, Seth Henneman, Seth Poulos, Shahar Mor, Shantanu Acharya, Shaona Ghosh, Sharath Turuvekere Sreenivas, Sharon Mendelson, Shaun Kotek, Shawn Wang, Shay Aharon, Shaya Gharghabi, Sheng-Chieh Lin, Shi Chen, Shiqing Fan, Shirish Baskaran, Shreya Gopa, Shrimai Prabhumoye, Shubham Pa-

chori, Shubham Toshniwal, Shuoyang Ding, Shwetha Krishnamurthy, Siddharth Singh, Simeng Sun, Sirshak Das, Sivakumar Arayandi Thottakara, Smita Ithape, Somshubra Majumdar, Soumye Singhal, Sri Harsha Singudasu, Sridhar Bhuvanapalli, Srimukh Veccham, Stas Sergienko, Stefania Alborghetti, Stephen Ge, Su Rong, Sugam Dipak Devare, Sukrit Rao, Sumeet Kumar Barua, Sungsoo Ha, Sunny Gai, Suriya Gunasekar, Suseella Panguluri, Suyog Gupta, Sviataslau Hinzburh, Sweta Priyadarshi, Syeda Nahida Akter, Talor Abramovich, Tan Bui, Tanay Varshney, Tatevik Ter-Hovhannisyan, Teodor-Dumitru Ene, Terry Kong, Thanh Do, Tianhe Zhang, Tiffany Moore, Tijmen Blankevoort, Tim Moon, Tiyasa Mitra, Tom Balough, Tomasz Grzegorzek, Tomasz Hliwiak, Tomer Asida, Tomer Bar Natan, Tomer Keren, Tomer Ronen, Tony Salim, Tony Wang, Traian Rebedea, Tugrul Konuk, Twinkle Vashishth, Udi Karpas, Ushnish De, Vahid Noorozi, Venkat Srinivasan, Venmugil Elango, Vibhor Agrawal, Victor Cui, Vijay Korthikanti, Vikas Mehta, Vinay Rao, Virginia Wu, Vitaly Kurin, Vitaly Lavrukhin, Vladimir Anisimov, Vu Pham, Wanli Jiang, Wasi Uddin Ahmad, Wataru Ishihara, Wei Du, Wei Ping, Weiheng Chai, Wenliang Dai, Wesley Helmholz, Will Jennings, Will Zhu, Wojciech Prazuch, Xiaowei Ren, Xiwen Yu, Yan Breek, Yang Chen, Yang Yu, Yangyi Chen, Yaniv Galron, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Ying Lin, Yonatan Geifman, Yonggan Fu, Youngeun Kwon, Yu Yao, Yugi Guvvla, Yuki Huang, Yunsheng Liu, Zach Moshe, Zachary Newell, Zhilin Wang, Zhiyu Li, Zhongbo Zhu, Zhuolin Yang, Zihan Liu, Zijie Yan, and Zsolt-Alon Wertheimer. Nemotron 3 ultra: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2026. URL https://arxiv.org/abs/2606.15007.

Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. Toeplitz neural network for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IxmWsm4xrua.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 100092–100118. Curran Associates, Inc., 2025. doi: 10.52202/085713-3345. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/904e89bb4e632e75fb47f093b620b257-Paper-Conference.pdf.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models Are Unsupervised Multitask Learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007. URL https://www.eurasip.org/Proceedings/Eusipco/Eusipco2007/Papers/a5p-h05.pdf.

Ali Shehper and Ashish Vaswani. A training-time diagnostic for generalization via the log-alignment ratio, 2026. URL https://arxiv.org/abs/2605.28975.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models, 2023. URL https://arxiv.org/abs/2307.08621.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and

Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5e5fd18f863cbe6d8ae392a93fd271c9-Paper-Conference.pdf.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yang24ab.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4904fad153f6434a7bcf04465d4be2cc-Paper-Conference.pdf.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. Turboquant: Online vector quantization with near-optimal distortion rate. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 56418–56439, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5c802ef38ab6e366c2ea06eee554c088-Paper-Conference.pdf.

A Experimental Details

A.1 Data, Checkpoints, and Common Protocol

Datasets. Our primary corpus is the `wikitext-103-raw-v1` validation split of the Hugging Face `Salesforce/wikitext` dataset (https://huggingface.co/datasets/Salesforce/wikitext) (Merity et al., 2017). WikiText-103 contains more than 100 million tokens from verified Good and Featured English Wikipedia articles and preserves article structure, case, punctuation, and numbers. We reconstruct full validation articles from its line-based records, discard headings, and sample without replacement 32 articles that contain at least 1,056 tokens under every tokenizer used in the primary comparison. No example crosses an article boundary. Each selected article supplies exact nested prefixes of lengths $\ell \in \{64, 128, 256, 384, 512, 768, 1024\}$, so a length comparison changes only how much of the same document is visible. For a cross-corpus replication, we use the English `20231101.en` configuration of `wikimedia/wikipedia` (https://huggingface.co/datasets/wikimedia/wikipedia), whose records are cleaned full articles with identifiers, titles, URLs, and text. We sample 32 articles that are long enough under both the Mamba-2 and RetNet tokenizers and repeat only the two principal experiments at their final selected layers.

Checkpoints. We analyze four public causal language-model checkpoints. GPT-2 small is a 12-layer, 12-head autoregressive Transformer with approximately 124 million parameters (https://huggingface.co/openai-community/gpt2). Based-360M is an approximately 360-million-parameter hybrid causal model whose global mixing layers use second-order Taylor-feature linear attention (https://huggingface.co/hazyresearch/based-360m) (Arora et al., 2024). Mamba-Codestral-7B-v0.1 is a 64-layer Mamba-2 code model with 7.3 billion parameters and 128 scalar SSD heads (https://huggingface.co/mistralai/Mamba-Codestral-7B-v0.1). RetNet-1.3B-100B is a 24-layer, 8-head causal retention model with 1.35 billion parameters (https://huggingface.co/fla-hub/retnet-1.3B-100B). We pin the four model revisions to `607a30d`, `4a63ec6`, `4f086c0`, and `7fddefc`, respectively, and pin WikiText to `b08601e`. These models are comparison cases for different causal mixing rules; differences in scale, tokenizer, and training data prevent a controlled causal conclusion from their raw cross-model curves.

Analyzed operators. For GPT-2 layers $\{0, 5, 11\}$, we reconstruct each causal-softmax matrix $\mathbf{A}_h = \mathrm{softmax}(\mathbf{Q}_h\mathbf{K}_h^\top/\sqrt{d_h} + \mathbf{C})$, where $\mathbf{C}$ is the causal logit mask. For the global Based

layers $\{1, 11, 21\}$, we reconstruct both the raw and row-normalized Taylor-feature linear-attention operators. For Mamba-2 layers $\{0, 31, 63\}$, we reconstruct $\mathbf{M}_h = \mathbf{L}_h \odot (\mathbf{C}_h \mathbf{B}_h^\top)$ using the realized input-dependent transitions. The Mamba-2 spectral measurements concern $\mathbf{M}_h$ itself; the pre-gating recurrent core acts on the value branch through $\mathbf{M}_h \operatorname{diag}(\Delta_h) + D_h \mathbf{I}$, where Δ_h contains the source-token discretization steps and D_h is the direct feedthrough coefficient. These additional factors are included in the reconstruction check but excluded from the reported spectra. For RetNet layers $\{0, 11, 23\}$, we multiply each rotary query–key score matrix by its head-specific exponential decay mask. We use all 12 GPT-2 heads, all 16 Based heads, all 8 RetNet heads, and eight Mamba-2 heads fixed before the sweep at indices $\{0, 16, 32, 48, 64, 80, 96, 112\}$.

Rank measurements and uncertainty. For every document, prefix, selected layer, and component, we compute the full singular spectrum and the three quantities in Definition 1. Stable rank and erank_2 weight squared singular values and therefore react strongly when one direction carries most of the energy; erank_1 gives more influence to a broad low-energy tail. The document is the independent statistical unit. We first take the median over heads or channels within a document, then report the median over 32 documents and 95% percentile intervals from 2,000 document-bootstrap resamples (seed `20270802`). To summarize growth, we fit the slope of log rank against log length over $\ell \geq 256$ and refit it inside every bootstrap resample. These slopes summarize the finite interval 256–1024; they are not estimates of an asymptotic theorem exponent.

A.2 Task 1: Validate the Reconstructed Mixing Operators

At lengths 16, 32, and 64, we compare matrix-based value mixing with a reference computation using the same captured inputs and report relative Frobenius output error. The GPT-2 reference is the checkpoint's native attention path. Based uses a separate Taylor-feature implementation with a cumulative-key denominator; Mamba-2 uses an explicit scalar recurrence built from the extracted transitions, factors, and discretization steps; RetNet uses the library's `naive_retention` reference in BF16 and FP32. The comparison includes the output projection for GPT-2 and Based, excludes it for Mamba-2 and RetNet, and excludes residual additions and subsequent normalizations for all models. For Mamba-2, both sides include source-token discretization and direct feedthrough, as specified above. Table 1 shows agreement to floating-point precision for GPT-2, Based, and Mamba-2. RetNet differs from the BF16 reference by at most 1.79×10^{-3}, while its error against the FP32 reference is 8.71×10^{-8}, consistent with a precision effect. These checks establish algebraic consistency with the specified references; they do not independently validate the native optimized kernels for Based, Mamba-2, or RetNet, or test a spectral theorem. In particular, the GPT-2 value of 0 should be read as "below the reported numerical resolution," not as an exact mathematical identity: the validation computes the same eager-attention path from the same captured Q, K, V, causal mask, softmax, value mixing, and c_{proj} in the checkpoint dtype. Since the model is in evaluation mode, attention and residual dropout are identities. Thus this result verifies agreement with the native implementation rather than showing that GPT-2 is trivial or that every possible attention implementation would reconstruct exactly.

The short validation lengths are sufficient to check the local algebra, but they do not by themselves validate numerical behavior at the longer lengths used below. In particular, for chunked or recurrent implementations they do not constitute a separate test of multi-chunk state propagation. The long-context spectral analyses should therefore be understood as relying on the implementation audit together with the explicit operator constructions used at those lengths.

Operator	Median error	Maximum error	FP32-reference maximum
GPT-2	0	0	0
Based	2.21×10^{-7}	3.04×10^{-7}	3.04×10^{-7}
Mamba-2	7.10×10^{-8}	1.91×10^{-7}	1.91×10^{-7}
RetNet	1.66×10^{-3}	1.79×10^{-3}	8.71×10^{-8}

Table 1: Reconstruction error against the reference computations specified in Task 1. A displayed 0 denotes a value below numerical resolution in the stored summary. The FP32 RetNet comparison separates reconstruction error from the numerical difference of the BF16 reference.

A.3 Task 2: Measure Rank Growth with Context Length

Target and method. This is the direct trained-checkpoint test of the paper's central empirical claim: whether the effective number of token directions grows as more context is included. For every native operator in the common protocol, we compare the three rank measures across nested prefixes and fit final-layer log–log slopes over lengths 256–1024. Theorem 2 proves diverging lower bounds for idealized fixed-rank factors when mask gain is controlled. The experiment asks only whether trained finite operators move in the same qualitative direction; it does not assume or establish the theorem's fixed-rank or Haar hypotheses.

Result and conclusion. Figure 1 shows increasing stable rank, erank_1, and erank_2 for Mamba-2 and RetNet at every selected depth. The exact final-layer slopes and document-bootstrap intervals appear in Table 2; all six lower bounds are positive. For Mamba-2, this agrees qualitatively with the general gain-dependent bounds in part (i) of Theorem 2. We do not use the stronger bounded-memory specialization as a checkpoint-specific explanation because the measured mask gain increases over the tested range (Task 5). RetNet's fixed exponential-decay mask has bounded gain, making its near-proportional middle- and final-layer growth the closest trained-model counterpart of part (ii).

GPT-2 changes regime with depth: all ranks grow in the early layer, whereas stable rank and erank_2 are nearly flat in the middle and final layers while erank_1 continues to grow. Task 3 relates this energy–tail separation to the causal-softmax sink mechanism. Both Based variants show the same separation in their final layers. For the raw causal-linear operator, the order-one stable rank and growing erank_1 agree with part (iii) of Theorem 2; row normalization is outside the theorem as stated but preserves the empirical pattern. These comparisons concern finite-context directions and order separations, not asymptotic exponents, complete verification of the random-geometry assumptions, or model accuracy.

Operator	srank	erank_1	erank_2
GPT-2	$0.014\ [0.009, 0.019]$	$1.069\ [1.033, 1.129]$	$0.111\ [0.087, 0.135]$
Based raw	$0.004\ [0.003, 0.006]$	$1.171\ [1.102, 1.218]$	$0.036\ [0.027, 0.043]$
Based normalized	$0.005\ [0.004, 0.006]$	$1.134\ [1.084, 1.184]$	$0.040\ [0.036, 0.045]$
Mamba-2	$0.624\ [0.568, 0.685]$	$0.986\ [0.974, 0.995]$	$0.889\ [0.850, 0.927]$
RetNet	$0.778\ [0.685, 0.857]$	$0.993\ [0.983, 1.003]$	$0.955\ [0.940, 0.993]$

Table 2: Final-layer log–log slopes over $\ell \geq 256$, with 95% document-bootstrap intervals. A value near zero means that rank is nearly constant over the tested range; a value near one means approximately proportional growth.

A.4 Task 3: Test the Proposed Causal-Softmax Sink Mechanism

Target and method. Theorem 12 gives a sufficient condition for a coherent early-token sink to concentrate causal-softmax attention. For the first causal column, write $\mathbf{A} = \mathbf{a}\mathbf{e}_1^\top + \mathbf{E}$, where $a_i = A_{i1}$ is the attention assigned to the first token and $\mathbf{E}$ contains all remaining columns. We measure the sink energy $\mathcal{S}_\ell = \sum_i a_i^2$, residual energy $\mathcal{D}_\ell = \|\mathbf{E}\|_F^2$, their ratio, the stronger quantity $(\mathcal{D}_\ell/\mathcal{S}_\ell)\log \ell$ needed for the erank_2 conclusion, and the localization of the leading singular vectors. We also construct two paired diagnostics from the same GPT-2 logits. The first replaces each first-column logit by the median accessible nonsink logit in its row before recomputing softmax; the second moves the removed score advantage to a later causal position. We run both native prefixes and prefixes with an explicit <|endoftext|> boundary token because token identity may change an apparent first-position sink.

Result and conclusion. GPT-2 becomes substantially more sink dominated with depth: the final-minus-early difference in $\log(\mathcal{D}_\ell/\mathcal{S}_\ell)$ is between -5.5 and -4.3 over the tested lengths. Accordingly, the early layer spreads on all three ranks, whereas the middle and final layers have stable rank near 1, nearly constant erank_2, and a growing erank_1 tail. This transition is compatible with the two regimes of Theorem 4: a non-collapsing bounded-score regime and sink-induced rank-one concentration in Frobenius energy, which need not eliminate the weak singular-value tail.

The observed first-column approximation is almost as accurate as the optimal rank-one SVD approximation: their median error ratio is 1.03. Neutralizing the first-column advantage increases stable rank and erank_2 throughout the tested range (Figure 3), further supporting the proposed spectral mechanism. However, $(\mathcal{D}_\ell/\mathcal{S}_\ell)\log \ell$ does not decrease with length, and the effect changes when the explicit boundary token is inserted. The trained GPT-2 observations are therefore compatible with the theorem's mechanism but do not verify either the bounded-score premise or the sink's asymptotic condition, nor do they establish the sink as the causal explanation of the checkpoint behavior.

A.5 Task 4: Check Whether Trained Factors Match the Finite-Rank Proxy

Target and method. The finite-rank theory assumes compressed query/key factors, independent Haar projection-side spaces (Assumption 1), and a common isotropic token frame (Assumption 2). These assumptions are tractable idealizations, not automatic consequences of the algebraic head dimension. We assess their plausibility by measuring how many singular directions are needed to retain 95% of each factor's energy, the canonical correlation and Procrustes alignment of the leading token-side spaces, and the fidelity obtained after truncating both factors to ranks $r \in \{1, 2, 4, 8, 16, 32\}$. A stringent supportive result requires some $r \le \min(16, d_h/4)$ to retain at least 95% of both factor energies and to reproduce the mixer and all three rank proxies within 10%.

Result and conclusion. The outcome is mixed (Figure 4). Based has the smallest factor energy dimension and the lowest rank-16 mixer error, but that error reaches approximately 10% at the longest tested context. The other three checkpoints have rank-16 mixer errors above 10% throughout the tested range. These diagnostics do not establish the combined compression-and-fidelity criterion for any checkpoint across all tested lengths, nor do they verify the Haar assumptions. We therefore retain the finite-rank model as a deliberately stylized mechanism that exposes how masks can reshape low-rank scores, without claiming that the trained checkpoints learn the theorem's factor geometry.

A.6 Task 5: Check Bounded Cumulative Memory in Trained Mamba

Target and method. Part (ii) of Theorem 2 specializes the general rank bound to masks with uniformly bounded gain $\kappa_\ell = \|\mathbf{L}\|_{\mathrm{op}}$. For a scalar semiseparable Mamba mask, Proposition 10 connects this condition to uniform contraction of fixed-length transition blocks and to bounded maximum row and column memory, denoted R_ℓ and C_ℓ. For every realized transition sequence in the trained Mamba checkpoint, we therefore compute the 95th percentile of transition products over fixed block lengths, R_ℓ, C_ℓ, and κ_ℓ. Recurrences compute the memory sums without materializing every dense mask, and power iteration for κ_ℓ is checked against exact small-matrix singular values. This task concerns only the stronger bounded-memory specialization of the singular-value theorem; it does not test the eigenvalue results.

Result and conclusion. Blocks of 4, 8, and 16 transitions contract over the tested range, but local contraction is not enough to demonstrate a length-uniform bound at these finite lengths. The cumulative row memory, column memory, and measured mask gain all increase with context, with fitted slopes between 0.20 and 0.24 and bootstrap intervals bounded away from zero (Figure 5). We therefore do not describe this trained checkpoint as empirically bounded-memory. The general gain-dependent part of Theorem 2 still applies under its theoretical assumptions, but the stronger bounded-memory rate is not used as a checkpoint-specific explanation.

A.7 Task 6: Replace Only the Mask While Holding Learned Factors Fixed

Target and method. The cross-model comparison in Figure 1 changes both the mask and the learned score factors. To isolate the contribution of forgetting or decay, let $\mathbf{S} = \mathbf{C}\mathbf{B}^\top$ for a realized Mamba-2 head and let $\mathbf{S}$ be the rotary query–key score matrix for RetNet. We compare

$$\mathbf{M}_{\mathrm{native}} = \mathbf{L}_{\mathrm{native}} \odot \mathbf{S}, \qquad \mathbf{M}_{\mathrm{all}} = \mathbf{J} \odot \mathbf{S},$$

where $\mathbf{J}$ is the lower-triangular all-one mask. Thus the document, prefix, layer, component, and every learned factor are identical in the pair; only the native decay is removed. At the final selected layer and length 1024, we take the within-document median of $\mathrm{rankproxy}(\mathbf{M}_{\mathrm{all}}) - \mathrm{rankproxy}(\mathbf{M}_{\mathrm{native}})$ and bootstrap documents. This is the precise meaning of "holding learned factors fixed" in Figure 6.

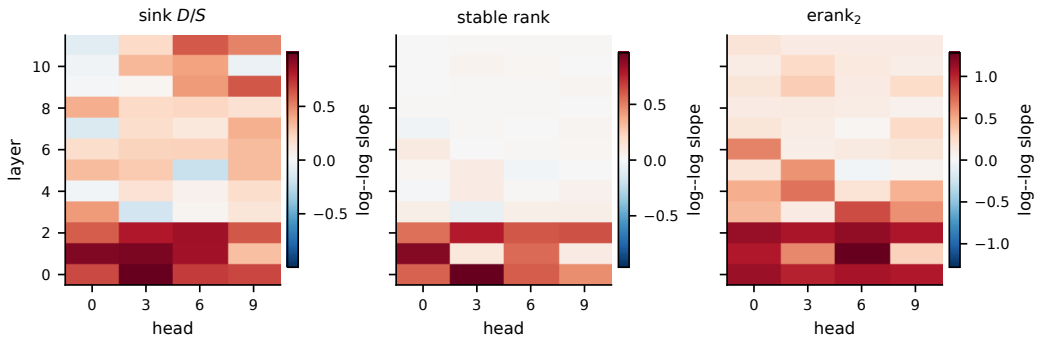


Figure 2: GPT-2 sink diagnostics by layer and selected head. Each cell is a log–log slope over the tested context lengths. The left panel tracks the residual-to-first-column energy ratio; the other panels track the two energy-sensitive ranks. Heterogeneity across depth and heads is one reason not to turn the aggregate association into a universal checkpoint claim.

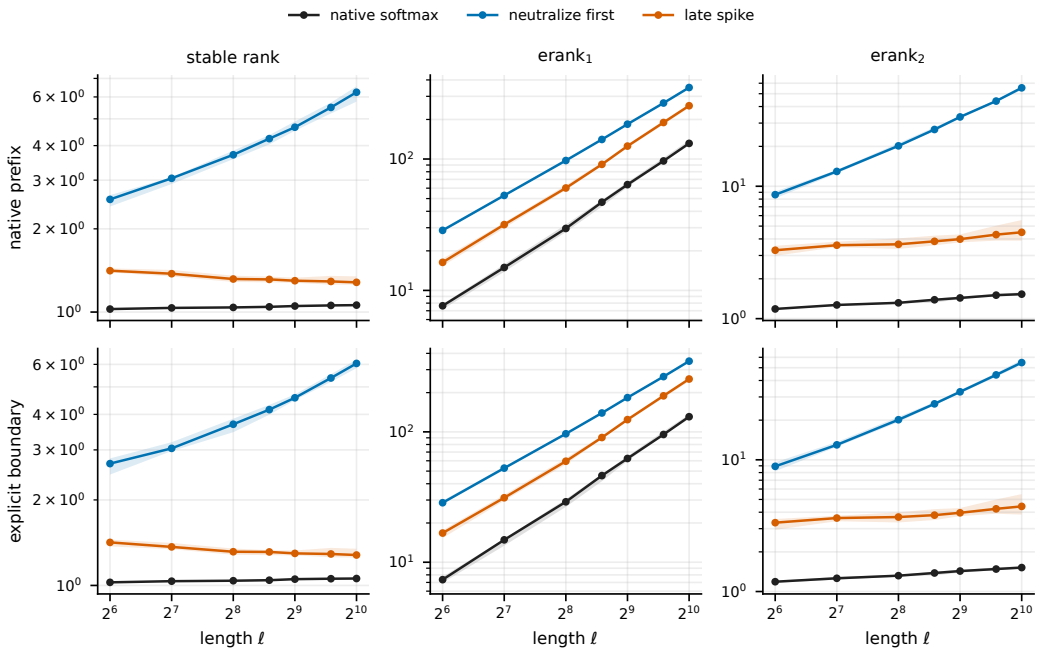


Figure 3: Conditional GPT-2 score diagnostics at the final layer. The blue curves remove the first-column score advantage before recomputing softmax; the orange curves move that advantage to a later accessible position. Rows compare native prefixes with prefixes containing an explicit boundary token. These matrices diagnose score geometry and are not complete model forward passes.

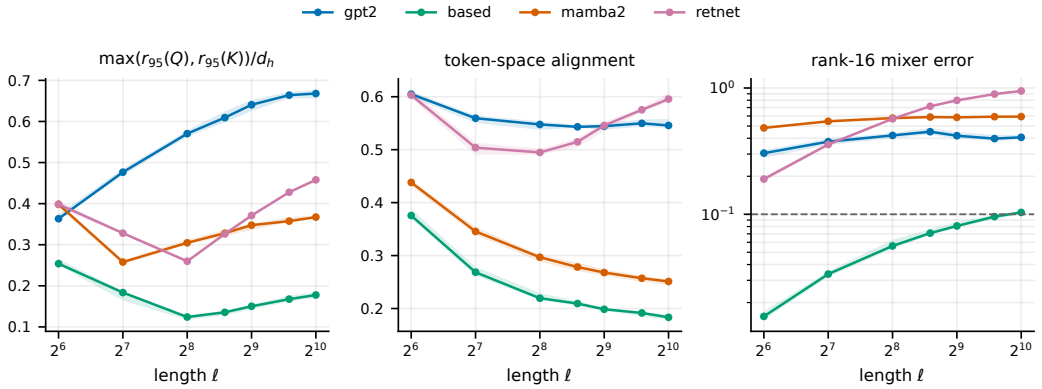


Figure 4: Final-layer audit of the finite-rank model assumptions. Left: fraction of head dimension needed to retain 95% of both factor energies. Middle: alignment of the leading token-side spaces. Right: relative mixer error after retaining rank 16; the dashed line marks a 10% mixer-error threshold. Mixer fidelity alone does not establish the combined compression-and-rank criterion.

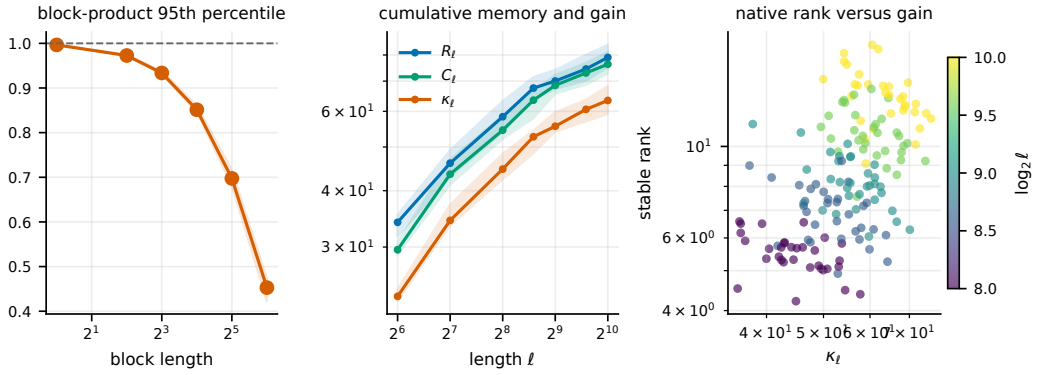


Figure 5: Memory diagnostics for trained Mamba-2. Left: long transition blocks contract more strongly. Middle: maximum row memory R_ℓ, maximum column memory C_ℓ, and mask gain κ_ℓ nevertheless increase over the observed lengths. Right: document-level stable rank and gain, shown descriptively rather than as a causal relation.

Result and conclusion. For Mamba-2 and RetNet, the all-one replacement lowers stable rank, erank_1, and erank_2; all six 95% intervals lie below zero. Removing forgetting or decay therefore creates a more dominant singular direction for the factors learned by both checkpoints. This provides controlled support for the qualitative mask mechanism behind the gain-dependent theory. It is not a direct test of the theorem's rates, because Task 4 did not verify all fixed-rank/Haar assumptions, and it is not an end-to-end intervention: gates, value paths, output projections, residual connections, and downstream task behavior are not recomputed.

The all-one replacement also has an intrinsic spectral effect: the lower- triangular integration mask $\mathbf{J}$ has its own dominant singular mode. Moreover, the learned score factors were optimized together with the native forgetting or decay mask, so the paired contrast can include representation co-adaptation to that mask. The observed rank decrease should therefore be interpreted as the response of the realized factors to this specific mask replacement, rather than as evidence that mask structure alone determines the spectrum independently of the factors.

A.8 Task 7: Replicate the Main Findings on English Wikipedia

Target, method, and result. To test whether the two principal observations depend on the WikiText sample, we repeat Tasks 2 and 6 for final-layer Mamba-2 and RetNet on 32 seeded English articles from `wikimedia/wikipedia` (configuration `20231101.en`). We use the same seven nested lengths, the same component-within-document aggregation, and a 2,000-resample document bootstrap. All three rank measures again increase with context for both models, and every all-one-minus-native mask effect is again negative (Figure 7). The RetNet stable-rank growth rate is smaller than on WikiText, so the replicated claim is the direction of finite-range growth, not a corpus-invariant exponent. The replication provides additional finite-range, cross-corpus support for the two operator-level observations, but does not address downstream language modeling or additional model families.

A.9 Scope of the Empirical Evidence

The experiments concern the spectra of reconstructed core token-mixing operators over lengths at most 1024. They do not measure task accuracy, long-context retrieval, or whether a high-rank direction carries useful information. Raw cross-model comparisons remain confounded by model size, training corpus and budget, tokenizer, and projection dimension. The mask replacement removes the learned-factor confound within each checkpoint but is a conditional matrix construction rather than a full forward pass. Finally, none of these trained-checkpoint experiments tests the idealized i.i.d. Gaussian query–key model of Section 2.2; it is not identified with the distribution induced by literal initialization of an architecture.

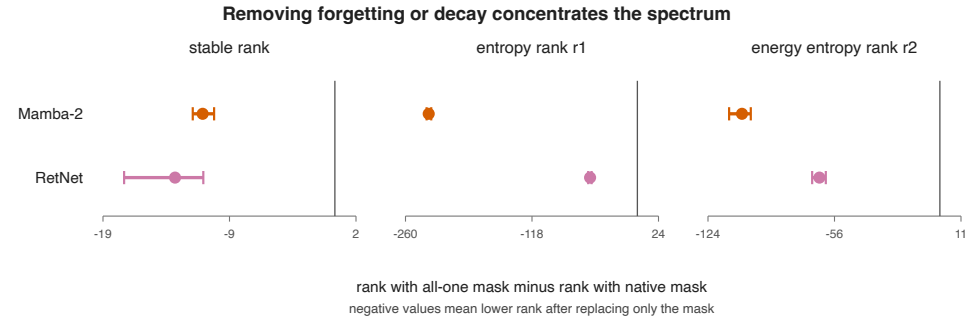


Figure 6: Matched-factor mask intervention at the final selected layer and length 1024. For each document and component, the learned factors are held fixed and only the native forgetting or decay mask is replaced by an undiscounted all-one causal mask. The horizontal axis is *rank with all-one mask minus rank with native mask*; negative values mean that removing forgetting or decay concentrates the spectrum. Points are paired document effects and bars are 95% document-bootstrap intervals.

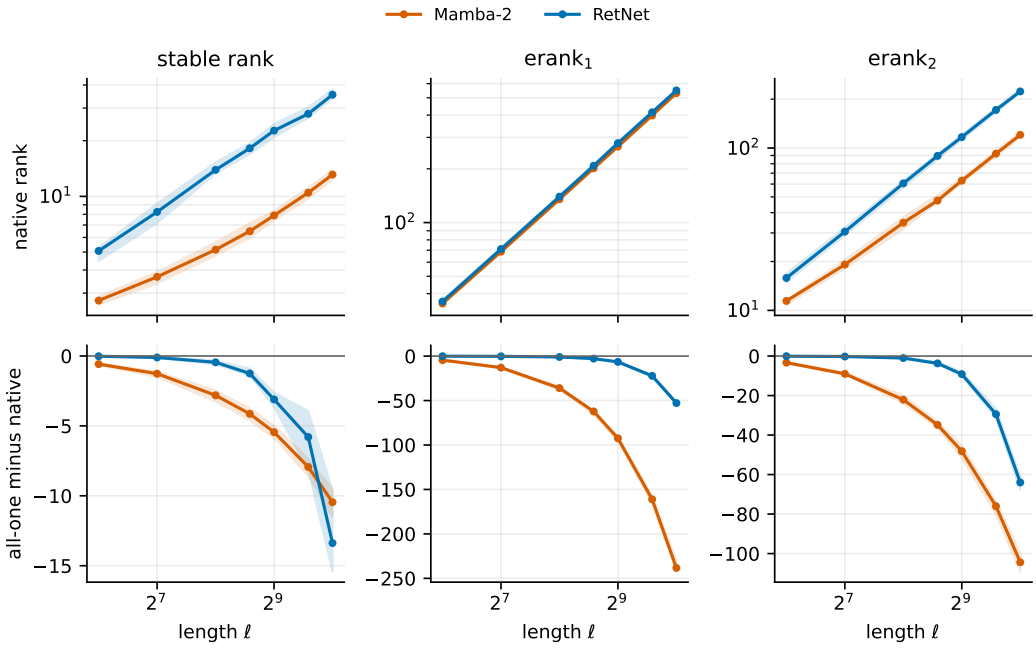


Figure 7: Cross-corpus replication on 32 English Wikipedia articles. Top: final-layer native ranks as context grows. Bottom: rank with the all-one mask minus rank with the native mask. The six top curves increase and the six bottom contrasts are negative, reproducing both principal WikiText findings. Bands are 95% document-bootstrap intervals.

PROOF ROADMAP

This appendix organizes the proofs around four reusable mechanisms. The finite-rank and Gaussian arguments share the same spectral rank proxies, but use different sources of randomness and therefore have different limiting arguments.

Finite-rank eigenvalues. The lower-triangular structure first reduces the eigenvalues to diagonal entries. The exact reduction and the common-Haar quadratic-form limit are given in Propositions 3 and 4, with the joint limit in Theorem 9. Haar-row concentration and extreme-value estimates then give the non-collapsing empirical law, effective-count bounds, and rank asymptotics in Propositions 6, 7, and 8.

Finite-rank singular values and mask memory. The low-rank matrix is reduced to a finite sum of diagonally weighted masks in Proposition 9. Haar overlap and token-row lemmas prevent cancellation among the finitely many components. Diagonal entries provide Frobenius and nuclear mass, while the mask gain $\kappa_\ell = \|\mathbf{L}\|_{\mathrm{op}}$ controls the largest singular direction. The equivalence between uniform block forgetting, bounded cumulative memory, and bounded gain is Proposition 10; principal inverse bounds and causal summation norms are supplied by Proposition 11 and Lemma 9. These ingredients yield the rank-one empirical-law and mask-gain theorem 10, the memory-dependent corollary 1, the higher-rank bounds in Theorem 2, and the principal-inverse higher-rank empirical nondegeneracy in Theorem 3. The latter includes the undiscounted causal mask; its sharper rank rates are handled separately by the weighted Volterra proposition 12.

Gaussian query–key factors. Under the independent Gaussian model, each diagonal eigenvalue is a sum of independent Gaussian products; Propositions 1 and 2 convert its law into eigenvalue rank proxies. For singular values, diagonal mass and mask-gain bounds give the general lower bounds in Theorem 7. The causal-linear limiting law is obtained from Wick pairings and the causal order-polytope volumes $V^{\mathrm{caus}}_{\pi,\rho}$ in Theorem 8; the semiseparable and decay cases instead use a vanishing unrescaled first spectral moment to prove the degenerate LSD in Theorem 6.

Causal softmax. The row-softmax denominator requires a separate argument from the entrywise masked model. Bounded within-row score oscillation is compared with the causal averaging kernel in Theorem 11. For a coherent early-token sink, the attention matrix is split into a first column and a nonsink residual; the score margin makes the residual negligible in Frobenius norm. This gives the sink-collapse theorem 12 and its score-margin and aligned-SVD corollaries in Appendix D.

Definition 1 (Rank proxies)**.** For a probability vector $w = (w_1, \dots, w_m)$, define its Shannon entropy by

$$H(w) := -\sum_{j=1}^{m} w_j \log w_j,$$

using natural logarithms and the convention $0 \log 0 := 0$. For a nonzero matrix $\mathbf{X} \in \mathbb{R}^{\ell \times \ell}$, let $s_1(\mathbf{X}) \geq \cdots \geq s_\ell(\mathbf{X}) \geq 0$ be its singular values and define the probability vectors $q(\mathbf{X})$ and $\widetilde{q}(\mathbf{X})$ by

$$q_j(\mathbf{X}) := \frac{s_j(\mathbf{X})}{\sum_{k=1}^{\ell} s_k(\mathbf{X})}, \qquad \widetilde{q}_j(\mathbf{X}) := \frac{s_j(\mathbf{X})^2}{\sum_{k=1}^{\ell} s_k(\mathbf{X})^2}.$$

The stable rank is

$$\mathrm{srank}(\mathbf{X}) := \frac{\|\mathbf{X}\|_F^2}{\|\mathbf{X}\|_{\mathrm{op}}^2} = \frac{\sum_{j=1}^{\ell} s_j(\mathbf{X})^2}{s_1(\mathbf{X})^2}.$$

The effective rank of Roy & Vetterli (2007) and its squared-singular-value analogue are

$$\mathrm{erank}_1(\mathbf{X}) := e^{H(q(\mathbf{X}))}, \qquad \mathrm{erank}_2(\mathbf{X}) := e^{H(\widetilde{q}(\mathbf{X}))}.$$

For the eigenvalue-based proxies, let $\lambda_1(\mathbf{X}), \dots, \lambda_\ell(\mathbf{X})$ denote the eigenvalues, counted with algebraic multiplicity. When $\sum_j |\lambda_j(\mathbf{X})| > 0$, define

$$p_j(\mathbf{X}) := \frac{|\lambda_j(\mathbf{X})|}{\sum_{k=1}^{\ell} |\lambda_k(\mathbf{X})|},$$

and set

$$\mathrm{srank}_\lambda(\mathbf{X}) := \frac{\sum_{j=1}^{\ell} |\lambda_j(\mathbf{X})|}{\max_{1\le j\le \ell} |\lambda_j(\mathbf{X})|}, \qquad \mathrm{erank}_\lambda(\mathbf{X}) := e^{H(p(\mathbf{X}))}.$$

For the lower-triangular attention matrix $\mathbf{M}$, the eigenvalues are $\lambda_i(\mathbf{M}) = M_{ii}$. For $\mathbf{X} = 0$, set all five proxies to zero. If $\mathbf{X} \neq 0$ but all its eigenvalues vanish, set only the two eigenvalue-based proxies to zero.

Notation We write $X_\ell = O_{\mathbb{P}}(a_\ell)$ if X_ℓ/a_ℓ is tight, $X_\ell = \Omega_{\mathbb{P}}(a_\ell)$ if X_ℓ/a_ℓ is bounded below away from zero in probability, and $X_\ell = \Theta_{\mathbb{P}}(a_\ell)$ if both hold. Whenever both dimensions vary, $(\ell, d) \to (\infty, \infty)$ denotes the unrestricted joint limit $\min\{\ell, d\} \to \infty$; in particular, neither dimension is assumed to be a function of the other. The same probabilistic order notation is used for double-indexed families along this joint limit. Deterministic constants implicit in $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ may depend on fixed mask parameters but not on varying ℓ or d; when d is explicitly fixed, it is an allowed fixed parameter. Subscripts such as $O_R(\cdot)$ and $\Theta_{r,R}(\cdot)$ indicate the allowed parameter dependence.

B I.I.D. Gaussian Model Details

B.1 Eigenvalue Distribution

The proof of Theorem 5 uses the following two propositions. Proposition 1 identifies the exact law of the diagonal Gaussian inner products, and Proposition 2 converts it into stable-rank and entropy-effective-rank estimates.

Roadmap First identify the diagonal entries using triangularity, then apply a law of large numbers to their magnitudes. The maximum controls the stable-rank denominator, while the empirical averages of w and $w \log w$ determine the entropy effective rank. For convenience, the Gaussian limit stated inline in the main text is recorded here with an appendix-only label:

$$\frac{\sqrt{d}}{\sigma_Q \sigma_K} \lambda_i(\mathbf{M}) \Rightarrow N(0,1) \qquad (d \to \infty). \tag{14}$$

For fixed d, the accompanying rank conclusions are

$$\mathrm{srank}_\lambda(\mathbf{M}) = \Theta_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right), \qquad \mathrm{erank}_\lambda(\mathbf{M}) = \Theta_{\mathbb{P}}(\ell) \qquad (\ell \to \infty).$$

Proposition 1 (Diagonal spectrum under independent Gaussian factors)**.** *Under Assumption 4, the eigenvalues of* $\mathbf{M}$ *are independent and*

$$\lambda_i(\mathbf{M}) = \langle \mathbf{q}_i, \mathbf{k}_i \rangle \overset{d}{=} \frac{\sigma_Q \sigma_K}{d} \sum_{a=1}^{d} Z_{ia} W_{ia}, \tag{15}$$

where Z_{ia}, W_{ia} *are independent standard Gaussians. Equivalently,*

$$\mathbb{E} e^{t\lambda_i(\mathbf{M})} = \left(1 - \frac{\sigma_Q^2 \sigma_K^2 t^2}{d^2}\right)^{-d/2}, \qquad |t| < \frac{d}{\sigma_Q \sigma_K}. \tag{16}$$

In particular

$$\mathbb{E}\lambda_i(\mathbf{M}) = 0, \qquad \mathrm{Var}\{\lambda_i(\mathbf{M})\} = \frac{\sigma_Q^2 \sigma_K^2}{d},$$

and the Gaussian limit equation 14 holds.

Proof. Since $\mathbf{L}$ is lower triangular and $L_{ii} = 1$, the eigenvalues of $\mathbf{M} = \mathbf{L} \odot \mathbf{Q}\mathbf{K}^\top$ are

$$\lambda_i(\mathbf{M}) = M_{ii} = L_{ii} \langle \mathbf{q}_i, \mathbf{k}_i \rangle = \langle \mathbf{q}_i, \mathbf{k}_i \rangle.$$

The pairs $(\mathbf{q}_i, \mathbf{k}_i)$ are independent across i, so the eigenvalues are independent. Write $\mathbf{q}_i = (\sigma_Q/\sqrt{d})(Z_{i1}, \ldots, Z_{id})$ and $\mathbf{k}_i = (\sigma_K/\sqrt{d})(W_{i1}, \ldots, W_{id})$. This gives equation 15. For independent standard Gaussians Z, W, conditioning on Z gives

$$\mathbb{E}\{e^{sZW} \mid Z\} = e^{s^2 Z^2/2}.$$

Averaging over Z yields

$$\mathbb{E}e^{sZW} = (1-s^2)^{-1/2}, \qquad |s| < 1.$$

Multiplying d independent copies gives equation 16. The mean and variance follow either by differentiating the moment generating function at zero or directly from $\mathbb{E}ZW = 0$ and $\mathrm{Var}(ZW) = 1$. Finally,

$$\frac{\sqrt{d}}{\sigma_Q \sigma_K} \lambda_i(\mathbf{M}) = \frac{1}{\sqrt{d}} \sum_{a=1}^{d} Z_{ia} W_{ia},$$

so the central limit theorem gives equation 14. □

Proposition 2 (Eigenvalue rank proxies under the Gaussian model)**.** *Let* srank_λ *and* erank_λ *be defined as in Definition 1. Set*

$$X_d := \frac{1}{\sqrt{d}} \sum_{a=1}^{d} Z_a W_a, \qquad \mu_d := \mathbb{E}|X_d|, \qquad \eta_d := \mathbb{E}\{|X_d| \log |X_d|\}.$$

For each fixed d*, as* $\ell \to \infty$*,*

$$\mathrm{srank}_\lambda(\mathbf{M}) = \Theta_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right), \qquad \frac{\mathrm{erank}_\lambda(\mathbf{M})}{\ell} \xrightarrow{\mathbb{P}} c_d := \mu_d \exp\left\{-\frac{\eta_d}{\mu_d}\right\}. \tag{17}$$

Proof. By Proposition 1,

$$\lambda_i(\mathbf{M}) \overset{d}{=} \frac{\sigma_Q \sigma_K}{\sqrt{d}} X_{i,d},$$

where $X_{1,d}, \ldots, X_{\ell,d}$ are independent copies of X_d. The scalar factor $\sigma_Q \sigma_K / \sqrt{d}$ cancels from both srank_λ and erank_λ, so it is enough to analyze the absolute weights $w_i = |X_{i,d}|$.

For fixed d, the moment generating function $\mathbb{E}e^{tX_d} = (1 - t^2/d)^{-d/2}$ is finite for $|t| < \sqrt{d}$ and singular at $|t| = \sqrt{d}$. Chernoff's bound therefore gives constants $b_d, B_d > 0$ such that

$$\mathbb{P}\{|X_d| > x\} \le B_d e^{-b_d x}$$

for all sufficiently large x. For the matching lower bound, write $S_{d-1} = \sum_{a=2}^{d} Z_a W_a$ (with $S_0 = 0$), and choose a fixed $D_d < \infty$ such that $\mathbb{P}\{|S_{d-1}| \le D_d\} > 0$. Independence, together with the standard Gaussian Mills bound, gives, for all sufficiently large u,

$$\mathbb{P}\{ZW > u\} \ge \mathbb{P}\{Z > \sqrt{u},\ W > \sqrt{u}\} = \overline{\Phi}(\sqrt{u})^2 \ge c u^{-1} e^{-u}.$$

Consequently,

$$\begin{aligned}\mathbb{P}\{X_d > x\} &\ge \mathbb{P}\{Z_1 W_1 > \sqrt{d}\, x + D_d,\ |S_{d-1}| \le D_d\} \\ &\ge a_d x^{-1} e^{-A_d x}\end{aligned}$$

for constants $a_d, A_d > 0$. A union bound with the upper tail, and independence with the lower tail, now show that for constants $0 < \underline{c}_d < \overline{C}_d < \infty$,

$$\mathbb{P}\{\underline{c}_d \log \ell \le \max_i w_i \le \overline{C}_d \log \ell\} \to 1.$$

The law of large numbers gives

$$\frac{1}{\ell} \sum_i w_i \to \mu_d \qquad \text{and} \qquad \frac{1}{\ell} \sum_i w_i \log w_i \to \eta_d$$

in probability; the exponential tail guarantees the needed integrability. Therefore

$$\mathrm{srank}_\lambda(\mathbf{M}) = \frac{\sum_i w_i}{\max_i w_i} = \Theta_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right).$$

For the entropy effective-rank proxy, use

$$-\sum_i p_i \log p_i = \log\left(\sum_i w_i\right) - \frac{\sum_i w_i \log w_i}{\sum_i w_i}, \qquad p_i = \frac{w_i}{\sum_j w_j}.$$

Substituting the two laws of large numbers gives

$$\frac{\operatorname{erank}_\lambda(\mathbf{M})}{\ell} \xrightarrow{\mathbb{P}} \mu_d \exp\left\{-\frac{\eta_d}{\mu_d}\right\},$$

which implies the $\Theta_{\mathbb{P}}(\ell)$ statement in Proposition 2.

□

Proof of Theorem 5. Proposition 1 gives the diagonal representation and the Gaussian limit equation 14. The fixed-d stable-rank estimate and the linear effective-rank estimate follow from Proposition 2 because equation 17 has a positive finite constant c_d. □

Scale interpretation. The scalar factor $\sigma_Q\sigma_K/\sqrt{d}$ cancels from both eigenvalue rank proxies, although each fixed eigenvalue is $O_{\mathbb{P}}(d^{-1/2})$ in absolute magnitude as $d \to \infty$. At fixed d, the eigenvalues remain distributed over many token directions as $\ell \to \infty$: the entropy effective rank is linear in ℓ, while the stable-rank proxy loses only a logarithmic extreme-value factor.

Remark 3 (A heteroskedastic Gaussian extension). In SSD/Mamba-2, the query-key factors in the structured-attention view correspond to the C and B branches. In common implementations these branches are produced from the same input by position-shared projections and a depthwise causal convolution, then split into B and C. Literal model initialization therefore need not yield factors that are i.i.d. across tokens. One limited extension of Assumption 4 is

$$\tau_i^Q, \tau_i^K > 0, \qquad \mathbf{q}_i \mid \tau_i^Q \sim N\left(0, \frac{(\tau_i^Q)^2}{d}\mathbf{I}_d\right), \qquad \mathbf{k}_i \mid \tau_i^K \sim N\left(0, \frac{(\tau_i^K)^2}{d}\mathbf{I}_d\right),$$

with $\mathbf{q}_i$ and $\mathbf{k}_i$ conditionally independent but with token-dependent scales inherited from the input, convolution, gate, and normalization. To retain the fixed-d rank conclusions above, additionally assume that, conditional on the entire scale sequence, the row pairs $(\mathbf{q}_i, \mathbf{k}_i)$ are independent across i and are generated by Gaussian innovations independent of the scales. Then

$$\lambda_i(\mathbf{M}) \mid \tau_i^Q, \tau_i^K \overset{d}{=} \frac{\tau_i^Q\tau_i^K}{d}\sum_{a=1}^{d} Z_{ia}W_{ia}.$$

If normalization makes $\tau_i^Q\tau_i^K$ nearly constant, this reduces to Proposition 1. If the products $\tau_i^Q\tau_i^K$ are heterogeneous but uniformly bounded away from zero and infinity and their empirical law converges in probability to a deterministic limit, the conditional independence above lets the fixed-d product-normal calculation yield a stable-rank proxy of order $\ell/\log\ell$ and an effective-rank proxy linear in ℓ, with constants adjusted by the scale mixture. Varying-d extreme-value constants would require separate moderate-deviation control. This extension still assumes conditional cross-token independence and therefore does not capture the full shared-input or convolutional dependence of a literal Mamba block.

The mask-energy condition stated inline in the main text is recorded here with an appendix-only label:

$$\frac{\|\mathbf{L}\|_F^2}{d\ell} \xrightarrow{\mathbb{P}} 0. \tag{18}$$

Remark 4 (Width scope). The causal-linear LSD and singular-value rank results take $\ell, d \to \infty$ with $\ell/d \to \tau \in (0, \infty)$. By contrast, the degenerate-LSD criterion only requires the mask-energy condition equation 18. When $\mathbf{L}$ is deterministic or independent of the Gaussian factors,

$$\mathbb{E}\left[\frac{1}{\ell}\|\mathbf{M}\|_F^2 \,\middle|\, \mathbf{L}\right] = \frac{\sigma_Q^2\sigma_K^2}{d\ell}\|\mathbf{L}\|_F^2.$$

For a unit-diagonal bounded-memory mask with $\|\mathbf{L}\|_F^2 = O(\ell)$, this conditional first moment is $\Theta(1/d)$. Thus the proof of raw spectral degeneration uses $d \to \infty$ and is not a fixed-width prediction; no relation between ℓ and d is needed for that conclusion.

Mechanism behind relative spread. Under the Gaussian model, the diagonal entries alone contribute squared Frobenius mass of order ℓ/d and nuclear mass of order $\ell/\sqrt{d}$, while the semiseparable and decay masks keep the operator norm at most logarithmically above the single-entry scale. Consequently, the unrescaled spectrum can shrink to zero while its normalized mass remains distributed over a nearly linear number of directions, as summarized by the main-text remark following Theorem 7.

Further directions. A separate open problem is to connect this i.i.d. reference regime to the dependence induced by literal architecture initialization and to the finite-rank regime used to model learned representations. Tracking spectra across training and depth could test whether non-collapse is preserved, created, or eroded and whether its trajectory predicts context use, optimization stability, or downstream performance. Extending the theory to selective transitions, normalization, gating, and multihead interactions remains equally important.

B.2 Singular-Value Distribution

For the causal linear mask, we characterize the nondegenerate squared-singular-value LSD through its moments. It is the boundary case against the degenerate unrescaled laws of the semiseparable and decay masks above.

Roadmap We expand normalized trace moments into Wick pairings. The causal inequalities then reduce the leading terms to fixed order-polytopes; a four-colour pairing bound shows that all remaining pairings are lower order. The resulting moments identify the causal limiting law, while a Frobenius second-moment bound handles the vanishing-energy masks.

Lemma 1 (Four-colour pairing bound)**.** *Let $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ be pair partitions of a finite set S of even cardinality N. Form the edge-coloured graph on S which has one edge of colour r for every pair in α_r, with colours read cyclically. If this graph is connected, then*

$$\sum_{r=0}^{3} \left|\alpha_r \vee \alpha_{r+1}\right| \le N + 2, \qquad \alpha_4 := \alpha_0.$$

In particular, if $|\alpha_3 \vee \alpha_0| = c$, then the sum of the other three terms is at most $N + 2 - c$.

Proof. Give the four half-edges at every vertex the cyclic order $0, 1, 2, 3$. For every adjacent pair of colours $(r, r+1)$, attach a face along each alternating cycle of colours r and $r+1$. This produces a connected closed surface (possibly nonorientable and with multiple edges): the link of each vertex is the four-cycle of adjacent colour pairs, and every edge lies on the two faces using its two adjacent colours. The number of vertices is N, the number of edges is $4(N/2) = 2N$, and the number of faces is

$$\sum_{r=0}^{3} \left|\alpha_r \vee \alpha_{r+1}\right|,$$

because an alternating bicoloured cycle is exactly a block of the corresponding join. Euler's inequality $V - E + F \le 2$ for a connected closed surface gives the claim. □

Lemma 2 (Causal order-polytope count)**.** *Fix finite sets $\mathcal{A}, \mathcal{B}$ and a relation $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{B}$. Write $[\ell] := \{1, \dots, \ell\}$ and set*

$$P_{\mathcal{R}} := \left\{(x, y) \in [0,1]^{\mathcal{A} \sqcup \mathcal{B}} : y_B \le x_A \text{ for every } (A, B) \in \mathcal{R}\right\}.$$

Then the number of $(u, v) \in [\ell]^{\mathcal{A} \sqcup \mathcal{B}}$ satisfying $v_B \le u_A$ for every $(A, B) \in \mathcal{R}$ is

$$\ell^{|\mathcal{A}|+|\mathcal{B}|} \operatorname{vol}(P_{\mathcal{R}}) + O_{\mathcal{R}}\left(\ell^{|\mathcal{A}|+|\mathcal{B}|-1}\right).$$

Proof. Put $m = |\mathcal{A}| + |\mathcal{B}|$. This is a fixed full-dimensional polytope: for example, all y-coordinates can be chosen strictly below all x-coordinates. Partition $[0,1]^m$ into cubes of side $1/\ell$. Cubes not meeting its boundary have constant membership in $P_{\mathcal{R}}$, so only boundary cubes and the coordinate-boundary convention can affect the grid count. Its boundary lies in a fixed finite union of coordinate

and comparison hyperplanes, each of which meets $O_{\mathcal{R}}(\ell^{m-1})$ cubes. The remaining boundary convention for $[\ell]$ also changes only $O_{\mathcal{R}}(\ell^{m-1})$ grid points. Multiplying the resulting $O_{\mathcal{R}}(\ell^{-1})$ discrepancy by ℓ^m proves the formula. □

Theorem 8 (LSD of causal linear attention)**.** *Under Assumption 4, fix positive σ_Q, σ_K and let $L_{ij} = \mathbf{1}_{\{j \le i\}}$. Assume that $\ell, d \to \infty$ with $\ell/d \to \tau \in (0, \infty)$. Let*

$$\nu_\ell^{(2)} := \frac{1}{\ell} \sum_{j=1}^{\ell} \delta_{s_j(\mathbf{M})^2}$$

be the empirical measure of squared singular values of $\mathbf{M}$. For every fixed integer $k \ge 1$,

$$\frac{1}{\ell} \operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big) \longrightarrow (\sigma_Q \sigma_K)^{2k} m_k(\tau)$$

in probability, where

$$m_k(\tau) = \sum_{\substack{\pi, \rho \in \mathcal{P}_2(2k) \\ |\eta \vee \pi| + |\pi \vee \rho| + |\rho \vee \kappa| = 2k+1}} \tau^{2k - |\pi \vee \rho|} V_{\pi,\rho}^{\mathrm{caus}}.$$

Here $\mathcal{P}_2(2k)$ denotes the set of pair partitions of $\{1, \dots, 2k\}$,

$$\kappa = \big\{\{2r-1, 2r\} : r = 1, \dots, k\big\},$$

and

$$\eta = \big\{\{2r, 2r+1\} : r = 1, \dots, k\big\}, \qquad 2k+1 \equiv 1.$$

For two partitions π, ρ, $\pi \vee \rho$ denotes their join.

The quantity $V_{\pi,\rho}^{\mathrm{caus}} \in [0,1]$ is the causal order-polytope volume defined as follows. Assign one variable $x_A \in [0,1]$ to each block A of $\eta \vee \pi$, and one variable $y_B \in [0,1]$ to each block B of $\kappa \vee \rho$. For each $p = 1, 2, \dots, 2k$, let A_p denote the block of $\eta \vee \pi$ containing p, and let B_p denote the block of $\kappa \vee \rho$ containing p. Define

$$V_{\pi,\rho}^{\mathrm{caus}} = \int_{[0,1]^{|\eta \vee \pi| + |\kappa \vee \rho|}} \prod_{r=1}^{k} \mathbf{1}_{\{y_{B_{2r-1}} \le x_{A_{2r-1}}\}} \mathbf{1}_{\{y_{B_{2r}} \le x_{A_{2r}}\}} \prod_A dx_A \prod_B dy_B.$$

Consequently, $\nu_\ell^{(2)}$ converges weakly in probability to a deterministic probability measure $\nu_\tau^{(2)}$ on $[0, \infty)$, where the fixed scales σ_Q, σ_K are suppressed from the notation, satisfying

$$\int x^k \, d\nu_\tau^{(2)}(x) = (\sigma_Q \sigma_K)^{2k} m_k(\tau), \qquad k = 1, 2, \dots.$$

Moreover, $\nu_\tau^{(2)}$ has bounded support:

$$\operatorname{supp}(\nu_\tau^{(2)}) \subseteq \left[0, \sigma_Q^2 \sigma_K^2 (1 + \sqrt{\tau})^4\right].$$

Its first two moments are determined by

$$m_1(\tau) = \frac{\tau}{2}, \qquad m_2(\tau) = \frac{2\tau^2}{3} + \frac{\tau^3}{6}.$$

In particular, the limiting law is not a point mass, since

$$\operatorname{Var}_{\nu_\tau^{(2)}}(x) = (\sigma_Q \sigma_K)^4 \left(\frac{5\tau^2}{12} + \frac{\tau^3}{6}\right) > 0.$$

Proof of Theorem 8. Under Assumption 4, for the causal mask,

$$(\mathbf{M})_{ij} = \mathbf{1}_{\{j \le i\}} \sum_{a=1}^{d} q_{ia} k_{ja},$$

and

$$\mathbb{E} q_{ia} q_{i'a'} = \frac{\sigma_Q^2}{d} \mathbf{1}_{\{i=i'\}} \mathbf{1}_{\{a=a'\}}, \qquad \mathbb{E} k_{ja} k_{j'a'} = \frac{\sigma_K^2}{d} \mathbf{1}_{\{j=j'\}} \mathbf{1}_{\{a=a'\}}.$$

With the cyclic convention $i_{k+1} = i_1$,

$$\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big) = \frac{1}{\ell}\sum_{i_1,\dots,i_k=1}^{\ell}\sum_{j_1,\dots,j_k=1}^{\ell}\sum_{a_1,b_1,\dots,a_k,b_k=1}^{d}\prod_{r=1}^{k}\mathbf{1}_{\{j_r\le i_r\}}\mathbf{1}_{\{j_r\le i_{r+1}\}} \times q_{i_r a_r}k_{j_r a_r}q_{i_{r+1}b_r}k_{j_r b_r}.$$

Take expectations and apply Wick's formula separately to the q_{ia}-variables and the k_{ja}-variables. Let the q-pairing be π and the k-pairing be ρ. The covariance factors contribute $(\sigma_Q\sigma_K)^{2k}d^{-2k}$. The feature indices are identified by both pairings, hence they have $|\pi\vee\rho|$ free classes and contribute $d^{|\pi\vee\rho|}$. The i-indices are identified by the cyclic trace identifications η and by π, hence they have $|\eta\vee\pi|$ free classes. The j-indices are identified by κ and by ρ, hence they have $|\rho\vee\kappa|$ free classes. After these equality constraints, the causal indicators are exactly the inequalities in the definition of $V^{\mathrm{caus}}_{\pi,\rho}$. Applying Lemma 2 with the free i- and j-index classes as $\mathcal{A}$ and $\mathcal{B}$, respectively, gives

$$\ell^{|\eta\vee\pi|+|\rho\vee\kappa|}V^{\mathrm{caus}}_{\pi,\rho} + O\left(\ell^{|\eta\vee\pi|+|\rho\vee\kappa|-1}\right)$$

admissible choices of the sequence indices. Therefore, for fixed k,

$$\mathbb{E}\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big) = (\sigma_Q\sigma_K)^{2k}\sum_{\pi,\rho\in\mathcal{P}_2(2k)}\left(\frac{\ell}{d}\right)^{2k-|\pi\vee\rho|} \times \ell^{|\eta\vee\pi|+|\pi\vee\rho|+|\rho\vee\kappa|-2k-1}\left(V^{\mathrm{caus}}_{\pi,\rho}+O_k(\ell^{-1})\right).$$

We now check which terms can survive. Put the four pairings η,π,ρ,κ on $\{1,\dots,2k\}$. Since the union graph of κ and η is a single cycle, $|\kappa\vee\eta| = 1$ and their four-colour graph is connected. Applying Lemma 1 and removing the final $|\kappa\vee\eta| = 1$ face count gives

$$|\eta\vee\pi| + |\pi\vee\rho| + |\rho\vee\kappa| \le 2k+1.$$

Thus every exponent of ℓ in the expectation is nonpositive, and only terms with equality can contribute at order one. The $O_k(\ell^{-1})$ terms in the preceding finite sum are therefore $o(1)$. Since $\ell/d\to\tau$,

$$\mathbb{E}\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big) \longrightarrow (\sigma_Q\sigma_K)^{2k}\sum_{\substack{\pi,\rho\in\mathcal{P}_2(2k)\\ |\eta\vee\pi|+|\pi\vee\rho|+|\rho\vee\kappa|=2k+1}}\tau^{2k-|\pi\vee\rho|}V^{\mathrm{caus}}_{\pi,\rho} = (\sigma_Q\sigma_K)^{2k}m_k(\tau).$$

For $k=1$, the only pair of pairings has one free feature class and causal volume $1/2$, giving $m_1(\tau)=\tau/2$. For $k=2$, the surviving pairs (π,ρ) are (κ,κ), (η,κ), and (η,η). Their causal volumes are $1/3$, $1/6$, and $1/3$, respectively: the first and last require one variable to lie above or below two others, while the middle requires both y-variables to lie below both x-variables. Their respective powers of τ are 2, 3, and 2, which gives the stated formula for $m_2(\tau)$.

It remains to show concentration. Expanding the second moment gives the same Wick calculation on two trace cycles, with $4k$ positions for each Gaussian family and covariance factor $(\sigma_Q\sigma_K)^{4k}d^{-4k}$. Let $\widehat{\eta},\widehat{\kappa}\in\mathcal{P}_2(4k)$ be the disjoint unions of the two copies of the trace pairings η,κ. Thus $|\widehat{\kappa}\vee\widehat{\eta}| = 2$. Say that (π,ρ) connects the two trace cycles when its four-colour graph has a component meeting both copies of $\{1,\dots,2k\}$. Since $\widehat{\eta}\vee\widehat{\kappa}$ already connects every position within each copy, this is equivalent to connectedness of the four-colour graph. For such a pair, Lemma 1 gives

$$|\widehat{\eta}\vee\pi| + |\pi\vee\rho| + |\rho\vee\widehat{\kappa}| \le 4k.$$

Put $f = |\pi\vee\rho|$ and $g = |\widehat{\eta}\vee\pi| + |\rho\vee\widehat{\kappa}|$. The causal inequalities can only reduce the number of sequence-index choices, so a connected contribution, including the two normalized traces, is

$$O\left(d^{f-4k}\ell^{g-2}\right) = O\left(\left(\frac{\ell}{d}\right)^{4k-f}\ell^{f+g-4k-2}\right) = O(\ell^{-2}).$$

There are only finitely many pairings for fixed k. If the Wick pairings do not connect the trace cycles, no π- or ρ-edge can join the two copies: otherwise the four-colour graph would connect them because $\widehat{\eta}\vee\widehat{\kappa}$ is connected within each copy. In this case π and ρ are disjoint unions of

pairings on the two copies, and all Wick constraints, causal inequalities, and free-index sums factor exactly into two one-trace factors. Conversely, every ordered pair of one-trace pairings yields such a disconnected diagram. The disconnected Wick sum is therefore exactly the square of the one-trace expectation. Consequently,

$$\mathbb{E}\left(\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big)\right)^2 = \left[\mathbb{E}\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big)\right]^2 + O(\ell^{-2}),$$

and therefore

$$\operatorname{Var}\left(\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big)\right) = O(\ell^{-2}) \longrightarrow 0.$$

The moment convergence in probability follows from Chebyshev's inequality.

The first moment convergence gives tightness in probability, since for every positive R,

$$\nu_\ell^{(2)}([R,\infty)) \le \frac{1}{R}\frac{1}{\ell}\operatorname{Tr}(\mathbf{M}\mathbf{M}^\top).$$

Set $M_0 = 1$ and $M_k = (\sigma_Q\sigma_K)^{2k} m_k(\tau)$ for $k \ge 1$. These numbers form a Stieltjes moment sequence. Indeed, for every polynomial $p(x) = \sum_r c_r x^r$, the expectation limits above imply

$$\sum_{r,s} c_r c_s M_{r+s} = \lim_{\ell\to\infty} \mathbb{E}\int p(x)^2\, d\nu_\ell^{(2)}(x) \ge 0,$$

and the same argument applied to $xp(x)^2$ gives $\sum_{r,s} c_r c_s M_{r+s+1} \ge 0$. The Stieltjes moment criterion therefore gives at least one probability measure on $[0,\infty)$ with these moments.

To see that the moments determine a compactly supported law, use the same comparison at finite (ℓ, d). In the Wick expansion of each fixed moment, every covariance coefficient is nonnegative and the causal indicators are at most one. Replacing all indicators by one therefore gives

$$0 \le \mathbb{E}\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^k\big) \le \mathbb{E}\frac{1}{\ell}\operatorname{Tr}\big((\mathbf{Q}\mathbf{K}^\top(\mathbf{Q}\mathbf{K}^\top)^\top)^k\big) \le \mathbb{E}\|\mathbf{Q}\mathbf{K}^\top\|_{\mathrm{op}}^{2k}.$$

For the unmasked matrix $\mathbf{Q}\mathbf{K}^\top$,

$$\|\mathbf{Q}\mathbf{K}^\top\|_{\mathrm{op}} \le \|\mathbf{Q}\|_{\mathrm{op}}\|\mathbf{K}\|_{\mathrm{op}}, \qquad \|\mathbf{Q}\|_{\mathrm{op}} \longrightarrow |\sigma_Q|(1+\sqrt{\tau}), \quad \|\mathbf{K}\|_{\mathrm{op}} \longrightarrow |\sigma_K|(1+\sqrt{\tau}),$$

in probability. Gaussian operator-norm tails have moments of every fixed order uniformly in (ℓ, d) when ℓ/d stays in a compact subset of $(0,\infty)$, so this convergence in probability supplies the uniform integrability needed to pass to expectations of each fixed power. Hence

$$0 \le (\sigma_Q\sigma_K)^{2k} m_k(\tau) \le \left(\sigma_Q^2\sigma_K^2(1+\sqrt{\tau})^4\right)^k.$$

Let $\nu_\tau^{(2)}$ be any representing probability measure. For any positive ε,

$$\nu_\tau^{(2)}\left(\left[\sigma_Q^2\sigma_K^2(1+\sqrt{\tau})^4 + \varepsilon, \infty\right)\right) \le \left(\frac{\sigma_Q^2\sigma_K^2(1+\sqrt{\tau})^4}{\sigma_Q^2\sigma_K^2(1+\sqrt{\tau})^4 + \varepsilon}\right)^k.$$

Letting $k \to \infty$ gives zero mass above $\sigma_Q^2\sigma_K^2(1+\sqrt{\tau})^4 + \varepsilon$, and then letting ε be arbitrary gives

$$\operatorname{supp}(\nu_\tau^{(2)}) \subseteq \left[0, \sigma_Q^2\sigma_K^2(1+\sqrt{\tau})^4\right].$$

The same argument applies to every representing measure. Hence every such measure is compactly supported in the displayed interval, which gives moment determinacy and makes $\nu_\tau^{(2)}$ unique.

It remains to pass the moments through weak subsequential limits. Starting from any subsequence, use the convergence in probability of all integer moments and a diagonal extraction to choose a further subsequence along which all of them converge almost surely. On this probability-one event, convergence of the first moments gives tightness. Moreover, for every fixed $k \ge 1$, convergence of the $(k+1)$-st moments gives a uniform bound along this subsequence and

$$\int_{[R,\infty)} x^k\, d\nu_\ell^{(2)}(x) \le \frac{1}{R}\int x^{k+1}\, d\nu_\ell^{(2)}(x).$$

Thus the k-th moments are uniformly integrable, so every weak subsequential limit has moments $(M_k)_{k\ge 0}$ and must equal $\nu_\tau^{(2)}$. Every subsequence therefore has a further subsequence converging almost surely to $\nu_\tau^{(2)}$, which is equivalent to weak convergence in probability of $\nu_\ell^{(2)}$ to $\nu_\tau^{(2)}$. □

Proof of Theorem 6. It suffices to show that

$$\frac{1}{\ell}\operatorname{Tr}(\mathbf{M}\mathbf{M}^\top) = \frac{1}{\ell}\sum_{j=1}^{\ell} s_j(\mathbf{M})^2 \xrightarrow{\mathbb{P}} 0.$$

Indeed, for every fixed $\varepsilon > 0$,

$$\nu_\ell^{(2)}([\varepsilon, \infty)) \le \frac{1}{\varepsilon\ell}\operatorname{Tr}(\mathbf{M}\mathbf{M}^\top).$$

Conditioning on the mask and using Assumption 4,

$$\mathbb{E}\left[\frac{1}{\ell}\operatorname{Tr}(\mathbf{M}\mathbf{M}^\top) \mid \mathbf{L}\right] = \frac{\sigma_Q^2\sigma_K^2}{d\ell}\|\mathbf{L}\|_F^2.$$

Set

$$X_{\ell,d} := \frac{\|\mathbf{L}\|_F^2}{d\ell}, \qquad Y_{\ell,d} := \frac{1}{\ell}\operatorname{Tr}(\mathbf{M}\mathbf{M}^\top).$$

For every $\varepsilon, \delta > 0$, conditional Markov and the preceding identity give

$$\begin{aligned}\mathbb{P}\{Y_{\ell,d} > \varepsilon\} &\le \mathbb{P}\{X_{\ell,d} > \delta\} + \mathbb{E}\big[\mathbf{1}_{\{X_{\ell,d}\le\delta\}}\mathbb{P}\{Y_{\ell,d} > \varepsilon \mid \mathbf{L}\}\big] \\ &\le \mathbb{P}\{X_{\ell,d} > \delta\} + \frac{\sigma_Q^2\sigma_K^2\delta}{\varepsilon}.\end{aligned}$$

The first term tends to zero by equation 18; taking then $\delta \downarrow 0$ proves $Y_{\ell,d} \xrightarrow{\mathbb{P}} 0$. The first display implies $\nu_\ell^{(2)} \Rightarrow \delta_0$ in probability.

It remains to verify the stated mask-family applications of equation 18. For the 1-semiseparable mask, we handle the two alternatives in Assumption 5. In the logistic-normal case, independence of the a_t's gives

$$\mathbb{E}\|\mathbf{L}\|_F^2 = \sum_{1\le j\le i\le \ell}\mathbb{E}\prod_{t=j+1}^{i} a_t^2 = \sum_{r=0}^{\ell-1}(\ell-r)(\mathbb{E}a_1^2)^r \le \frac{\ell}{1-\mathbb{E}a_1^2}.$$

Here $\mathbb{E}a_1^2 < 1$ because $a_1 = (1+\exp(-Z_1))^{-1} \in (0,1)$ almost surely. In deterministic alternative (b), take $m = 1$ and $\rho = a_{\max}$. The same argument also covers the more general block condition equation 56: every lag-r product can be split into $\lfloor r/m \rfloor$ complete blocks of length m and one remaining block of length less than m. Since all coefficients lie in $[0,1]$, this gives

$$\prod_{t=j+1}^{j+r} a_t \le \rho^{\lfloor r/m \rfloor}.$$

Consequently,

$$\|\mathbf{L}\|_F^2 = \sum_{1\le j\le i\le \ell}\prod_{t=j+1}^{i} a_t^2 \le \sum_{r=0}^{\ell-1}(\ell-r)\rho^{2\lfloor r/m\rfloor} \le \frac{m\ell}{1-\rho^2}.$$

For the Toeplitz-decay mask, the fixed constant $\gamma \in (0,1)$ gives the deterministic bound

$$\|\mathbf{L}\|_F^2 = \sum_{r=0}^{\ell-1}(\ell-r)\gamma^{2r} \le \frac{\ell}{1-\gamma^2}.$$

These estimates give $\mathbb{E}\|\mathbf{L}\|_F^2 = O(\ell)$. Since $d \to \infty$, equation 18 follows, proving the claimed instances. □

Lemma 3 (Frobenius transfer under probabilistic mask energy)**.** *Under Assumption 4, let* $\mathbf{L}$ *be deterministic or independent of* $(\mathbf{Q}, \mathbf{K})$*. If*

$$\|\mathbf{L}\|_F^2 = O_{\mathbb{P}}(\ell),$$

then

$$\|\mathbf{M}\|_F^2 = O_{\mathbb{P}}\left(\frac{\ell}{d}\right).$$

Proof. Set

$$X_{\ell,d} := \frac{\|\mathbf{L}\|_F^2}{\ell}, \qquad Y_{\ell,d} := \frac{d}{\ell}\|\mathbf{M}\|_F^2.$$

Conditioning on $\mathbf{L}$ gives

$$\mathbb{E}[Y_{\ell,d} \mid \mathbf{L}] = \sigma_Q^2\sigma_K^2 X_{\ell,d}.$$

For any $A, K > 0$, conditional Markov therefore yields

$$\mathbb{P}\{Y_{\ell,d} > K\} \le \mathbb{P}\{X_{\ell,d} > A\} + \frac{\sigma_Q^2\sigma_K^2 A}{K}.$$

Since $X_{\ell,d} = O_{\mathbb{P}}(1)$, first choose A large and then K large; this proves $Y_{\ell,d} = O_{\mathbb{P}}(1)$ and hence the claim. □

Proof of Theorem 7. General mask-gain bounds. Let $\mathbf{q}^{(a)}$ and $\mathbf{k}^{(a)}$ denote the ath columns of $\mathbf{Q}$ and $\mathbf{K}$, respectively. Then

$$\mathbf{M} = \sum_{a=1}^{d} \operatorname{diag}(\mathbf{q}^{(a)})\,\mathbf{L}\,\operatorname{diag}(\mathbf{k}^{(a)}).$$

Conditionally on $\mathbf{L}$ and $\mathbf{Q}$, write $\mathbf{k}^{(a)} = \sigma_K \mathbf{h}^{(a)}/\sqrt{d}$, where the entries h_{ja} are independent standard Gaussians, and set

$$\mathbf{F}_{aj} := \operatorname{diag}(\mathbf{q}^{(a)})\,\mathbf{L}\,\mathbf{e}_j\mathbf{e}_j^\top.$$

Thus

$$\mathbf{M} = \frac{\sigma_K}{\sqrt{d}} \sum_{a=1}^{d}\sum_{j=1}^{\ell} h_{ja}\mathbf{F}_{aj}.$$

The two conditional matrix-variance parameters satisfy

$$\begin{aligned}\left\|\sum_{a,j} \mathbf{F}_{aj}\mathbf{F}_{aj}^\top\right\|_{\mathrm{op}} &= \left\|\sum_{a=1}^{d} \operatorname{diag}(\mathbf{q}^{(a)})\mathbf{L}\mathbf{L}^\top \operatorname{diag}(\mathbf{q}^{(a)})\right\|_{\mathrm{op}} \\ &\le \kappa_\ell^2 \max_i \|\mathbf{q}_i\|_2^2, \\ \left\|\sum_{a,j} \mathbf{F}_{aj}^\top\mathbf{F}_{aj}\right\|_{\mathrm{op}} &= \max_j \sum_i L_{ij}^2\|\mathbf{q}_i\|^2 \le \kappa_\ell^2 \max_i \|\mathbf{q}_i\|^2.\end{aligned}$$

Standard Gaussian maximum and chi-square bounds, together with $\ell/d \to \tau$, give

$$\max_i \|\mathbf{q}_i\|^2 = O_{\mathbb{P}}(\sigma_Q^2).$$

More explicitly, conditional on $(\mathbf{L}, \mathbf{Q})$, define

$$v_\ell := \max\left\{\left\|\sum_{a,j} \mathbf{F}_{aj}\mathbf{F}_{aj}^\top\right\|_{\mathrm{op}}, \left\|\sum_{a,j} \mathbf{F}_{aj}^\top\mathbf{F}_{aj}\right\|_{\mathrm{op}}\right\}.$$

The Gaussian matrix-series inequality and Gaussian concentration give

$$\mathbb{P}\left\{\left\|\sum_{a,j} h_{ja}\mathbf{F}_{aj}\right\|_{\mathrm{op}} > C\sqrt{v_\ell \log(2\ell)} + t\sqrt{v_\ell} \,\middle|\, \mathbf{L}, \mathbf{Q}\right\} \le 2e^{-t^2/2}.$$

Taking $t = \sqrt{8\log(2\ell)}$, then using the displayed high-probability bounds for $\mathbf{Q}$ (and absorbing the deterministic factor $\sigma_K/\sqrt{d}$), implies

$$\frac{\sqrt{d}\,\|\mathbf{M}\|_{\mathrm{op}}}{\sigma_Q\sigma_K\kappa_\ell\sqrt{\log \ell}} = O_{\mathbb{P}}(1). \tag{19}$$

Because every mask in the theorem is unit diagonal, $M_{ii} = \langle \mathbf{q}_i, \mathbf{k}_i \rangle$. Put $X_{i,d} = \sqrt{d}\, M_{ii}/(\sigma_Q \sigma_K)$. These variables are independent over i, satisfy $\mathbb{E}X_{i,d}^2 = 1$, and have $\mathbb{E}X_{i,d}^4 = 3 + 6/d \le 9$. Since $X_{i,d} \Rightarrow N(0,1)$ as $d \to \infty$, uniform integrability from the fourth-moment bound gives $\mathbb{E}|X_{i,d}| \to \sqrt{2/\pi}$. Chebyshev's inequality for the independent triangular arrays $|X_{i,d}|$ and $X_{i,d}^2$ therefore gives

$$\frac{1}{\ell}\sum_{i=1}^{\ell} |X_{i,d}| \xrightarrow{\mathbb{P}} \sqrt{\frac{2}{\pi}}, \qquad \frac{1}{\ell}\sum_{i=1}^{\ell} X_{i,d}^2 \xrightarrow{\mathbb{P}} 1.$$

In particular, for a deterministic $c > 0$,

$$\sum_{i=1}^{\ell} |M_{ii}| \ge c\frac{\sigma_Q \sigma_K \ell}{\sqrt{d}}, \qquad \sum_{i=1}^{\ell} M_{ii}^2 \ge c\frac{\sigma_Q^2 \sigma_K^2 \ell}{d} \tag{20}$$

with probability tending to one. In particular,

$$\|\mathbf{M}\|_F^2 \ge c\frac{\sigma_Q^2 \sigma_K^2 \ell}{d}.$$

Nuclear-norm duality, using the diagonal sign matrix matching the signs of M_{ii}, also gives

$$\|\mathbf{M}\|_* \ge \sum_{i=1}^{\ell} |M_{ii}| \ge c\frac{\sigma_Q \sigma_K \ell}{\sqrt{d}}.$$

For every mask in part (ii), the preceding mask-energy estimates give $\|\mathbf{L}\|_F^2 = O_{\mathbb{P}}(\ell)$. Lemma 3 therefore gives

$$\|\mathbf{M}\|_F^2 = O_{\mathbb{P}}\left(\frac{\ell}{d}\right). \tag{21}$$

For the logistic-normal masks, $\mathbb{E}\|\mathbf{L}\|_F^2 = O(\ell)$ implies the stated $O_{\mathbb{P}}(\ell)$ bound by Markov; for the deterministic and Toeplitz masks it holds directly. By Proposition 14, the entropy–Schatten lower bound gives

$$\operatorname{erank}_1(\mathbf{M}) \ge \frac{\|\mathbf{M}\|_*^2}{\|\mathbf{M}\|_F^2} = \Omega_{\mathbb{P}}(\ell) \quad \text{for every mask in part (ii).} \tag{22}$$

Since $\operatorname{erank}_1(\mathbf{M}) \le \ell$ deterministically, this already proves the claimed $\Theta_{\mathbb{P}}(\ell)$ conclusion in part (ii). Combining these estimates with equation 19, and using Shannon entropy's min-entropy lower bound, yields

$$\frac{\kappa_\ell^2 \log \ell}{\ell} \operatorname{srank}(\mathbf{M}) = \Omega_{\mathbb{P}}(1), \qquad \frac{\kappa_\ell \sqrt{\log \ell}}{\ell} \operatorname{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}(1).$$

By Proposition 13, $\operatorname{erank}_2(\mathbf{M}) \ge \operatorname{srank}(\mathbf{M})$. This proves part (i).

Semiseparable and decay masks. For the two mask families in part (ii), let

$$R_\ell := \max_i \sum_j |L_{ij}| \vee \max_j \sum_i |L_{ij}|.$$

We will prove the sharper Schur bound

$$\|\mathbf{M}\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\frac{\sigma_Q \sigma_K \{R_\ell + \sqrt{R_\ell \log \ell}\}}{\sqrt{d}}\right). \tag{23}$$

We start with a row-sum and column-sum bound for the mask:

$$\max_i \sum_j |L_{ij}| \vee \max_j \sum_i |L_{ij}| \le C \log \ell. \tag{24}$$

For the Toeplitz-decay mask, the fixed constant $\gamma \in (0,1)$ gives

$$\max_i \sum_j |L_{ij}| \vee \max_j \sum_i |L_{ij}| \le \sum_{r=0}^{\ell-1} \gamma^r \le \frac{1}{1-\gamma} \le C \log \ell$$

for all sufficiently large ℓ, where C may depend on the fixed value of γ. For deterministic alternative (b), take $m = 1$ and $\rho = a_{\max}$. More generally, under equation 56, the same block decomposition as in the preceding proof gives

$$\max_i \sum_j |L_{ij}| \vee \max_j \sum_i |L_{ij}| \le \sum_{r=0}^{\ell-1} \rho^{\lfloor r/m \rfloor} \le \frac{m}{1-\rho} \le C \log \ell.$$

For the 1-semiseparable mask with logistic-normal coefficients, let H_ℓ be the longest consecutive run of indices t for which $a_t > 1/2$. Since the indicators $\mathbf{1}_{\{a_t > 1/2\}}$ are i.i.d. Bernoulli$(1/2)$, a union bound gives $H_\ell = O(\log \ell)$ with probability tending to one. On this event, every block of $H_\ell + 1$ consecutive coefficients contains at least one coefficient at most $1/2$. Hence, for every i and $0 \le r \le i-1$,

$$\prod_{t=i-r+1}^{i} a_t \le 2^{-\lfloor r/(H_\ell+1) \rfloor},$$

with the empty product equal to one. The same bound holds for every admissible forward product of length r. Therefore every row sum and every column sum is at most

$$\sum_{r=0}^{\ell-1} 2^{-\lfloor r/(H_\ell+1) \rfloor} \le 2(H_\ell + 1) \le C \log \ell,$$

which proves equation 24 in all cases. More precisely, the deterministic alternative, its fixed-block extension, and the Toeplitz mask have $R_\ell = O(1)$, whereas the logistic-normal mask has $R_\ell = O_{\mathbb{P}}(\log \ell)$.

The standard chi-square bound and $\ell/d \to \tau$ give $\max_i \|\mathbf{q}_i\| \le C\sigma_Q$ and $\max_j \|\mathbf{k}_j\| \le C\sigma_K$ with probability tending to one. Fix i and condition only on $\mathbf{L}$ and $\mathbf{q}_i$. The variables $\langle \mathbf{q}_i, \mathbf{k}_j \rangle$ are then independent centered Gaussians with variance

$$\frac{\sigma_K^2 \|\mathbf{q}_i\|^2}{d} \le C \frac{\sigma_Q^2 \sigma_K^2}{d}$$

whenever $\|\mathbf{q}_i\| \le C\sigma_Q$. For nonnegative weights b_j measurable with respect to $\mathbf{L}$, the function $\sum_j b_j |\langle \mathbf{q}_i, \mathbf{k}_j \rangle|$ of the independent standard Gaussian coordinates in the $\mathbf{k}_j$'s has conditional mean at most $C\sigma_Q\sigma_K(\sum_j b_j)/\sqrt{d}$ and is Lipschitz with constant at most $C\sigma_Q\sigma_K\sqrt{\sum_j b_j^2}/\sqrt{d}$ on the same row-norm event. Gaussian concentration therefore gives, whenever $\|\mathbf{q}_i\| \le C\sigma_Q$, $0 \le b_j \le 1$, and $\sum_j b_j \le R_\ell$,

$$\mathbb{P}\left\{ \sum_j b_j |\langle \mathbf{q}_i, \mathbf{k}_j \rangle| > C \frac{\sigma_Q \sigma_K \{R_\ell + \sqrt{R_\ell \log \ell}\}}{\sqrt{d}} \,\middle|\, \mathbf{L}, \mathbf{q}_i \right\} \le \ell^{-3}.$$

A union bound over i, followed by removal of the conditioning and intersection with the maximum-$\mathbf{q}_i$ norm event, gives the corresponding bound for all row sums of $|\mathbf{M}|$. For columns, condition only on $\mathbf{L}$ and $\mathbf{k}_j$ and intersect with the maximum-$\mathbf{k}_j$ norm event; the same argument and a union bound over j give the identical column-sum bound. Schur's test then proves equation 23.

For logistic-normal masks, $R_\ell = O_{\mathbb{P}}(\log \ell)$, so equation 23 gives the $\ell/(\log \ell)^2$ lower bounds for srank and erank_2. For the deterministic bounded-memory and Toeplitz masks, $R_\ell = O(1)$ gives $\|\mathbf{M}\|_{\mathrm{op}} = O_{\mathbb{P}}(\sigma_Q\sigma_K\sqrt{\log \ell/d})$, hence the stronger $\ell/\log \ell$ lower bounds. Together with equation 22, this proves part (ii).

The causal linear mask. Write $s_j = s_j(\mathbf{M})$, and fix an integer $p \ge 2$. By Theorem 8, for every fixed integer $r \ge 1$,

$$\frac{1}{\ell} \sum_{j=1}^{\ell} s_j^{2r} = \frac{1}{\ell} \operatorname{Tr}\big((\mathbf{M}\mathbf{M}^\top)^r\big) \xrightarrow{\mathbb{P}} (\sigma_Q\sigma_K)^{2r} m_r(\tau).$$

For $r = 1$, the limit is positive; indeed $m_1(\tau) = \tau/2$, so

$$(\sigma_Q\sigma_K)^2 m_1(\tau) = \frac{\sigma_Q^2 \sigma_K^2 \tau}{2} > 0.$$

The limits for $r = 1, 2, p$ are finite. Hence we may choose deterministic constants $0 < a < B < \infty$, $C < \infty$, and $C_p < \infty$ such that

$$0 < a < (\sigma_Q \sigma_K)^2 m_1(\tau) < B, \qquad C > (\sigma_Q \sigma_K)^4 m_2(\tau), \qquad C_p > (\sigma_Q \sigma_K)^{2p} m_p(\tau).$$

Convergence in probability gives

$$\mathbb{P}\left\{ a\ell \le \sum_{j=1}^{\ell} s_j^2 \le B\ell \right\} \to 1,$$

and

$$\mathbb{P}\left\{ \sum_{j=1}^{\ell} s_j^4 \le C\ell \right\} \to 1, \qquad \mathbb{P}\left\{ \sum_{j=1}^{\ell} s_j^{2p} \le C_p \ell \right\} \to 1.$$

Since this is a finite intersection of high-probability events, with probability tending to one,

$$a\ell \le \sum_{j=1}^{\ell} s_j^2 \le B\ell, \qquad \sum_{j=1}^{\ell} s_j^4 \le C\ell, \qquad \sum_{j=1}^{\ell} s_j^{2p} \le C_p \ell. \tag{25}$$

On the event equation 25,

$$\|\mathbf{M}\|_{\mathrm{op}}^2 = \max_j s_j^2 \le \left(\sum_{j=1}^{\ell} s_j^{2p} \right)^{1/p} \le C_p^{1/p} \ell^{1/p}.$$

Therefore

$$\mathrm{srank}(\mathbf{M}) = \frac{\|\mathbf{M}\|_F^2}{\|\mathbf{M}\|_{\mathrm{op}}^2} = \frac{\sum_{j=1}^{\ell} s_j^2}{\|\mathbf{M}\|_{\mathrm{op}}^2} \ge c_p \ell^{1-1/p}.$$

By Proposition 14, for any probability vector $(p_j)_j$,

$$-\sum_j p_j \log p_j \ge -\log\left(\sum_j p_j^2 \right).$$

Applying this to

$$p_j = \frac{s_j^2}{\sum_{k=1}^{\ell} s_k^2}$$

gives

$$\mathrm{erank}_2(\mathbf{M}) \ge \frac{\left(\sum_{j=1}^{\ell} s_j^2 \right)^2}{\sum_{j=1}^{\ell} s_j^4} \ge \frac{a^2 \ell^2}{C\ell} \ge c\ell.$$

It remains to prove the erank_1 bound. Interpolating the empirical L^1, L^2, and L^4 norms of the singular values gives

$$\frac{1}{\ell} \sum_{j=1}^{\ell} s_j \ge \frac{\left(\ell^{-1} \sum_{j=1}^{\ell} s_j^2 \right)^{3/2}}{\left(\ell^{-1} \sum_{j=1}^{\ell} s_j^4 \right)^{1/2}} \ge c.$$

Thus

$$\sum_{j=1}^{\ell} s_j \ge c\ell.$$

Using the same entropy lower bound, now with

$$q_j = \frac{s_j}{\sum_{k=1}^{\ell} s_k},$$

we obtain

$$\operatorname{erank}_1(\mathbf{M}) \geq \frac{\left(\sum_{j=1}^{\ell} s_j\right)^2}{\sum_{j=1}^{\ell} s_j^2} \geq \frac{c\ell^2}{B\ell} \geq c\ell.$$

This proves

$$\operatorname{srank}(\mathbf{M}) \geq c_p \ell^{1-1/p}, \qquad \operatorname{erank}_1(\mathbf{M}) \geq c\ell, \qquad \operatorname{erank}_2(\mathbf{M}) \geq c\ell$$

with probability tending to one. Finally, for any fixed $\varepsilon > 0$, choose an integer $p \geq \max\{2, 1/\varepsilon\}$. Then $\ell^{1-1/p} \geq \ell^{1-\varepsilon}$, after adjusting the constant, so

$$\operatorname{srank}(\mathbf{M}) \geq c_\varepsilon \ell^{1-\varepsilon}$$

with probability tending to one. Since both entropy effective ranks are at most ℓ deterministically, their linear lower bounds give the two $\Theta_{\mathbb{P}}(\ell)$ conclusions in part (iii). □

C Low-Rank Model Details

C.1 Haar-Frame Preliminaries

The proofs use three consequences of the common Haar frame: finite collections of token rows are asymptotically Gaussian, empirical token-row averages obey a Gaussian law of large numbers, and their largest quadratic forms grow on the $\log \ell$ scale. We record these facts once here and reuse them in both the eigenvalue and singular-value arguments.

Roadmap The Gaussian polar representation of a Haar frame reduces finite rows and empirical averages to an i.i.d. Gaussian array plus a common Gram normalization. We then prove the uniform indicator version by truncation and anti-concentration, and prove the extreme-value statement by sandwiching against Gaussian quadratic forms.

Write

$$\mathbf{H} := \mathbf{V}_Q^\top \mathbf{V}_K \in \mathbb{R}^{r_Q \times r_K} \tag{26}$$

for the right-space overlap matrix.

Lemma 4 (Haar overlap scale)**.** *Under Assumption 1, for deterministic* $\mathbf{a} \in \mathbb{R}^{r_Q}$ *and* $\mathbf{b} \in \mathbb{R}^{r_K}$*,*

$$\mathbf{a}^\top \mathbf{H} \mathbf{b} \overset{d}{=} \|\mathbf{a}\| \, \|\mathbf{b}\| \, \xi_d,$$

where ξ_d *is the inner product of two independent uniform unit vectors in* $\mathbb{R}^d$*. Moreover,* $\mathbb{P}\{\xi_d = 0\} = 0$*: for* $d = 1$*,* $\xi_d \in \{-1, 1\}$*, while for* $d \geq 2$ *it has a density. In particular,*

$$\mathbb{E}\, \mathbf{a}^\top \mathbf{H} \mathbf{b} = 0, \qquad \mathbb{E}|\mathbf{a}^\top \mathbf{H} \mathbf{b}|^2 = \frac{\|\mathbf{a}\|^2 \|\mathbf{b}\|^2}{d}.$$

Moreover, for a universal constant $c > 0$*,*

$$\mathbb{P}\left(|\mathbf{a}^\top \mathbf{H} \mathbf{b}| > t\|\mathbf{a}\|\|\mathbf{b}\|\right) \leq 2\exp(-cdt^2), \qquad 0 \leq t \leq 1.$$

If r_Q, r_K *are fixed and* $d \to \infty$*, then*

$$\sqrt{d}\, \mathbf{H} \Rightarrow \mathbf{G},$$

where $\mathbf{G} \in \mathbb{R}^{r_Q \times r_K}$ *has independent* $N(0,1)$ *entries.*

Proof. By right invariance of the Haar measure on the Stiefel manifold, for every deterministic vector $\mathbf{a} \in \mathbb{R}^{r_Q}$, the vector $\mathbf{V}_Q \mathbf{a}$ is uniform on the sphere of radius $\|\mathbf{a}\|$ in the column space sampled by $\mathbf{V}_Q$. Since the column space is itself uniformly distributed in $\mathbb{R}^d$, $\mathbf{V}_Q \mathbf{a}$ is a uniform vector on the sphere $\|\mathbf{a}\| S^{d-1}$. Similarly, $\mathbf{V}_K \mathbf{b}$ is uniform on $\|\mathbf{b}\| S^{d-1}$, and the two vectors are independent. Therefore

$$\mathbf{a}^\top \mathbf{H} \mathbf{b} = \langle \mathbf{V}_Q \mathbf{a}, \mathbf{V}_K \mathbf{b} \rangle$$

has the same distribution as $\|\mathbf{a}\| \, \|\mathbf{b}\| \, \xi_d$, where ξ_d is the first coordinate of a uniform vector on S^{d-1}. Thus $\mathbb{E}\xi_d = 0$ and $\mathbb{E}\xi_d^2 = 1/d$, giving the displayed second moment. The concentration bound is the standard spherical-cap estimate for a coordinate of a uniform vector on the sphere.

It remains to justify the joint Gaussian limit. Let $\mathbf{Y}_Q \in \mathbb{R}^{d\times r_Q}$ and $\mathbf{Y}_K \in \mathbb{R}^{d\times r_K}$ have independent standard Gaussian entries. The Gaussian polar representation of Haar Stiefel matrices gives

$$\mathbf{V}_Q \overset{d}{=} \mathbf{Y}_Q(\mathbf{Y}_Q^\top\mathbf{Y}_Q)^{-1/2}, \qquad \mathbf{V}_K \overset{d}{=} \mathbf{Y}_K(\mathbf{Y}_K^\top\mathbf{Y}_K)^{-1/2}.$$

Consequently,

$$\sqrt{d}\,\mathbf{H} \overset{d}{=} \left(\frac{\mathbf{Y}_Q^\top\mathbf{Y}_Q}{d}\right)^{-1/2}\left(\frac{\mathbf{Y}_Q^\top\mathbf{Y}_K}{\sqrt{d}}\right)\left(\frac{\mathbf{Y}_K^\top\mathbf{Y}_K}{d}\right)^{-1/2}.$$

Since r_Q, r_K are fixed, $d^{-1}\mathbf{Y}_Q^\top\mathbf{Y}_Q \to \mathbf{I}_{r_Q}$ and $d^{-1}\mathbf{Y}_K^\top\mathbf{Y}_K \to \mathbf{I}_{r_K}$ in probability. Conditional on $\mathbf{Y}_Q$, the middle matrix is Gaussian with independent columns and row covariance $d^{-1}\mathbf{Y}_Q^\top\mathbf{Y}_Q$, hence it converges jointly to a matrix with independent $N(0,1)$ entries. Slutsky's theorem proves $\sqrt{d}\,\mathbf{H} \Rightarrow \mathbf{G}$. □

Lemma 5 (Token-side Haar products)**.** *Under Assumption 2, for any fixed token i and fixed indices a, b,*

$$\mathbb{E}\,u_a(i)u_b(i) = \delta_{ab}\,\mathrm{Wg}^{O(\ell)}(1) = \frac{\delta_{ab}}{\ell}.$$

Moreover,

$$\mathbb{E}\,u_a(i)^2u_b(i)^2 = \begin{cases} \dfrac{3}{\ell(\ell+2)}, & a = b, \\ \dfrac{1}{\ell(\ell+2)}, & a \neq b. \end{cases}$$

In addition, for fixed m, r and distinct fixed tokens $i = 1, \dots, m$,

$$\Big(\sqrt{\ell}\,(u_1(i), \dots, u_r(i))\Big)_{i=1}^m \Rightarrow ((\xi_{i,1}, \dots, \xi_{i,r}))_{i=1}^m, \qquad \xi_{i,a} \overset{\text{i.i.d.}}{\sim} N(0,1),$$

and hence

$$(\ell\,u_a(i)u_b(i))_{i,a,b} \Rightarrow (\xi_{i,a}\xi_{i,b})_{i,a,b}.$$

Proof. Let $O \in O(\ell)$ be Haar. The first r columns of O have the same law as $\mathbf{U}$. We apply the orthogonal Weingarten formula of Collins & Śniady (2006). For two entries, the only pairing pairs the two column indices, so

$$\mathbb{E}O_{ia}O_{ib} = \delta_{ab}\,\mathrm{Wg}^{O(\ell)}(1) = \frac{\delta_{ab}}{\ell}.$$

For four entries, write the column-index word as (a, a, b, b). The orthogonal Weingarten formula gives

$$\mathbb{E}O_{ia}O_{ia}O_{ib}O_{ib} = \sum_{\pi,\sigma\in\mathcal{P}_2(4)} \delta_i^\pi\,\delta_{aabb}^\sigma\,\mathrm{Wg}^{O(\ell)}(\pi^{-1}\sigma),$$

where $\mathcal{P}_2(4)$ is the set of pair partitions of four points. Since all row indices are equal, $\delta_i^\pi = 1$ for every π. If $a \neq b$, only the pairing $\{\{1,2\},\{3,4\}\}$ is compatible with the column word; if $a = b$, all three pairings are compatible. Equivalently, the squared row $(O_{i1}^2, \dots, O_{i\ell}^2)$ has the Dirichlet distribution with all parameters $1/2$. Therefore

$$\mathbb{E}O_{ia}^2O_{ib}^2 = \frac{1+2\delta_{ab}}{\ell(\ell+2)}.$$

This is exactly the displayed formula.

For the finite-row Gaussian limit, use the Gaussian representation $\mathbf{U} = \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1/2}$ with $\mathbf{Y} \in \mathbb{R}^{\ell\times r}$ standard Gaussian. If $\mathbf{y}_i$ is the ith row, then

$$\sqrt{\ell}\,(u_1(i), \dots, u_r(i)) = \mathbf{y}_i\left(\frac{\mathbf{Y}^\top\mathbf{Y}}{\ell}\right)^{-1/2}.$$

For fixed m, r, the rows $\mathbf{y}_1, \dots, \mathbf{y}_m$ are independent $N(0, \mathbf{I}_r)$ and $\ell^{-1}\mathbf{Y}^\top\mathbf{Y} \to \mathbf{I}_r$ in probability. Slutsky's theorem gives the joint Gaussian limit for the first m token rows. The convergence of the products $\ell u_a(i)u_b(i)$ follows by the continuous mapping theorem. □

Lemma 6 (Empirical Haar-row law of large numbers). *Let r be fixed and let $\mathbf{U}_\ell = [\mathbf{u}_1, \dots, \mathbf{u}_r] \sim \mathrm{Haar}(\mathrm{St}(\ell, r))$. Define*

$$\mathbf{z}_i := \sqrt{\ell}\,(u_1(i), \dots, u_r(i))^\top \in \mathbb{R}^r, \qquad \mu_\ell := \frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{\mathbf{z}_i}.$$

Then $\mu_\ell \Rightarrow N(0, \mathbf{I}_r)$ in probability. In addition, for every continuous $f : \mathbb{R}^r \to \mathbb{R}$ of polynomial growth,

$$\frac{1}{\ell}\sum_{i=1}^{\ell} f(\mathbf{z}_i) \longrightarrow \mathbb{E} f(\boldsymbol{\xi}), \qquad \boldsymbol{\xi} \sim N(0, \mathbf{I}_r),$$

in probability. More generally, this convergence is uniform over compact families: if $\mathcal{K}$ is compact and $f : \mathcal{K} \times \mathbb{R}^r \to \mathbb{R}$ is jointly continuous with $|f(a, \mathbf{x})| \le C(1 + \|\mathbf{x}\|_2^p)$ for all $a \in \mathcal{K}$, then

$$\sup_{a \in \mathcal{K}} \left| \frac{1}{\ell}\sum_{i=1}^{\ell} f(a, \mathbf{z}_i) - \mathbb{E} f(a, \boldsymbol{\xi}) \right| \longrightarrow 0 \quad \textit{in probability}.$$

In particular, for every $R, \varepsilon > 0$,

$$\sup_{\substack{\mathbf{C} = \mathbf{C}^\top \\ \|\mathbf{C}\|_{\mathrm{op}} \le R}} \left| \frac{1}{\ell}\sum_{i=1}^{\ell} \mathbf{1}_{\{|\mathbf{z}_i^\top \mathbf{C} \mathbf{z}_i| > \varepsilon\}} - \mathbb{P}\{|\boldsymbol{\xi}^\top \mathbf{C} \boldsymbol{\xi}| > \varepsilon\} \right| \longrightarrow 0 \quad \textit{in probability}.$$

Proof. Let $\mathbf{G}_\ell \in \mathbb{R}^{\ell \times r}$ have independent $N(0,1)$ entries, with rows $\mathbf{g}_i^\top$. The Gaussian representation of the Haar Stiefel matrix gives

$$\mathbf{U}_\ell \overset{d}{=} \mathbf{G}_\ell (\mathbf{G}_\ell^\top \mathbf{G}_\ell)^{-1/2}.$$

Hence

$$\mathbf{z}_i \overset{d}{=} \left(\frac{\mathbf{G}_\ell^\top \mathbf{G}_\ell}{\ell} \right)^{-1/2} \mathbf{g}_i.$$

Since r is fixed, $\ell^{-1}\mathbf{G}_\ell^\top \mathbf{G}_\ell \to \mathbf{I}_r$ almost surely. The empirical law of the independent vectors $\mathbf{g}_i$ converges weakly to $N(0, \mathbf{I}_r)$ almost surely, and the common multiplier above converges to the identity. More explicitly, if $\mathbf{R}_\ell = (\ell^{-1}\mathbf{G}_\ell^\top \mathbf{G}_\ell)^{-1/2}$, then $\|\mathbf{R}_\ell - \mathbf{I}_r\|_{\mathrm{op}} \to 0$ almost surely and

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \varphi(\mathbf{R}_\ell \mathbf{g}_i) - \frac{1}{\ell}\sum_{i=1}^{\ell} \varphi(\mathbf{g}_i) \longrightarrow 0$$

for every bounded Lipschitz φ, because the Gaussian vectors have bounded empirical first moment with probability tending to one. This proves weak convergence of μ_ℓ. We record the uniform-integrability step needed for polynomial growth. If $q > p$, then eventually almost surely $\|\mathbf{R}_\ell\|_{\mathrm{op}} \le 2$ and

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \|\mathbf{R}_\ell \mathbf{g}_i\|_2^p \mathbf{1}_{\{\|\mathbf{R}_\ell \mathbf{g}_i\|_2 > A\}} \le \frac{2^q}{A^{q-p}} \frac{1}{\ell}\sum_{i=1}^{\ell} \|\mathbf{g}_i\|_2^q.$$

The right-hand side is bounded almost surely for large ℓ and vanishes as $A \to \infty$. Thus truncation is uniform for every fixed polynomial-growth envelope. After this truncation, uniform continuity controls the common multiplier $\mathbf{R}_\ell$, and a finite net in $\mathcal{K}$ reduces the asserted uniform convergence to the bounded-Lipschitz case above.

For the indicator statement, first restrict to the compact annulus $\mathcal{K}_{\delta,R} := \{\mathbf{C} = \mathbf{C}^\top : \delta \le \|\mathbf{C}\|_{\mathrm{op}} \le R\}$. The Gaussian boundary mass is uniformly negligible there. Indeed, if this failed, there would be $\eta_n \downarrow 0$, $\mathbf{C}_n \in \mathcal{K}_{\delta,R}$, and a fixed $c > 0$ such that

$$\mathbb{P}\{\varepsilon - \eta_n \le |\boldsymbol{\xi}^\top \mathbf{C}_n \boldsymbol{\xi}| \le \varepsilon + \eta_n\} \ge c.$$

Passing to a convergent subsequence $\mathbf{C}_n \to \mathbf{C}_*$ preserves $\|\mathbf{C}_*\|_{\mathrm{op}} \ge \delta$. Since $\boldsymbol{\xi}^\top \mathbf{C}_n \boldsymbol{\xi} \to \boldsymbol{\xi}^\top \mathbf{C}_* \boldsymbol{\xi}$ almost surely and the latter has no atoms at $\pm\varepsilon$, the band probabilities tend to zero: for every

fixed $s > 0$, the perturbation bound $|\boldsymbol{\xi}^\top(\mathbf{C}_n - \mathbf{C}_*)\boldsymbol{\xi}| \le s$ fails with probability tending to zero, while the s-neighborhood of $\pm\varepsilon$ has probability tending to zero as $s \downarrow 0$. This is a contradiction. Choose continuous functions $\psi_\eta^- \le \mathbf{1}_{\{|t|>\varepsilon\}} \le \psi_\eta^+$ whose difference is supported on the band $\{\varepsilon - \eta \le |t| \le \varepsilon + \eta\}$. Applying the compact-uniform result to $f_\eta^\pm(\mathbf{C}, \mathbf{z}) = \psi_\eta^\pm(\mathbf{z}^\top \mathbf{C}\mathbf{z})$ and then letting $\eta \downarrow 0$ proves the claim uniformly on $\mathcal{K}_{\delta,R}$. On $\|\mathbf{C}\|_{\mathrm{op}} \le \delta$, both the empirical and Gaussian indicator averages are at most $\delta^2\varepsilon^{-2}$ times their corresponding fourth radial moments, since

$$\mathbf{1}_{\{|\mathbf{z}^\top \mathbf{C}\mathbf{z}|>\varepsilon\}} \le \delta^2 \|\mathbf{z}\|_2^4/\varepsilon^2.$$

Those moments are tight by the polynomial result, so letting $\delta \downarrow 0$ completes the proof. □

Lemma 7 (Gaussian quadratic-form bounds)**.** *Let $\mathbf{C} \in \mathbb{R}^{r\times r}$ be symmetric, $\boldsymbol{\xi} \sim N(0, \mathbf{I}_r)$, and $Y_{\mathbf{C}} := \boldsymbol{\xi}^\top \mathbf{C}\boldsymbol{\xi}$. Then*

$$\mathbb{E}Y_{\mathbf{C}}^2 = 2\,\mathrm{tr}(\mathbf{C}^2) + (\mathrm{tr}\,\mathbf{C})^2, \qquad \mathbb{E}|Y_{\mathbf{C}}| \ge \frac{\|Y_{\mathbf{C}}\|_2}{9} \ge \frac{\sqrt{2}}{9}\|\mathbf{C}\|_{\mathrm{op}}. \tag{27}$$

If $\mathbf{C} \neq 0$, the law of $Y_{\mathbf{C}}$ has no atoms.

Proof. Diagonalizing $\mathbf{C}$ and using the Gaussian second and fourth moments gives the identity for $\mathbb{E}Y_{\mathbf{C}}^2$, which is at least $2\|\mathbf{C}\|_{\mathrm{op}}^2$. Gaussian hypercontractivity and log-convexity of L^p norms give

$$\|Y_{\mathbf{C}}\|_4 \le 3\|Y_{\mathbf{C}}\|_2, \qquad \|Y_{\mathbf{C}}\|_2 \le \|Y_{\mathbf{C}}\|_1^{1/3}\|Y_{\mathbf{C}}\|_4^{2/3},$$

and hence $\mathbb{E}|Y_{\mathbf{C}}| \ge \|Y_{\mathbf{C}}\|_2/9$. Finally, when $\mathbf{C} \neq 0$, every level set of the nonconstant polynomial $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{C}\mathbf{x}$ has Lebesgue measure zero; the Gaussian law is absolutely continuous, so $Y_{\mathbf{C}}$ has no atoms. □

Lemma 8 (Haar-row extreme values)**.** *Let r be fixed and let $\mathbf{U}_\ell = [\mathbf{u}_1, \dots, \mathbf{u}_r] \sim \mathrm{Haar}(\mathrm{St}(\ell, r))$. Put*

$$\mathbf{z}_i := \sqrt{\ell}\,(u_1(i), \dots, u_r(i))^\top \in \mathbb{R}^r.$$

If $\mathbf{B}_\ell$ is a symmetric $r \times r$ random matrix, independent of $\mathbf{U}_\ell$, and $\rho_\ell := \|\mathbf{B}_\ell\|_{\mathrm{op}} > 0$ almost surely, then

$$\frac{1}{2\rho_\ell \log \ell} \max_{1\le i\le \ell} |\mathbf{z}_i^\top \mathbf{B}_\ell \mathbf{z}_i| \longrightarrow 1 \qquad \textit{in probability.} \tag{28}$$

More explicitly, for every $\eta > 0$,

$$\sup_{\substack{\mathbf{B}=\mathbf{B}^\top \\ \|\mathbf{B}\|_{\mathrm{op}}=1}} \mathbb{P}\left\{\left|\frac{\max_i |\mathbf{z}_i^\top \mathbf{B}\mathbf{z}_i|}{2\log \ell} - 1\right| > \eta\right\} \longrightarrow 0.$$

Thus the conclusion remains valid for coefficient matrices that vary with ℓ, provided they are independent of the token frame. In particular,

$$\max_{1\le i\le \ell} \|\mathbf{z}_i\|_2^2 = O_{\mathbb{P}}(\log \ell), \qquad \max_{1\le i\le \ell} \ell u_1(i)^2 = O_{\mathbb{P}}(\log \ell). \tag{29}$$

Proof. It is enough to prove the quadratic-form statement uniformly over deterministic symmetric matrices $\mathbf{B}$ with $\|\mathbf{B}\|_{\mathrm{op}} = 1$; conditioning on $\mathbf{B}_\ell/\rho_\ell$ then gives the random-matrix version. Use the Gaussian representation

$$\mathbf{U}_\ell \overset{d}{=} \mathbf{G}_\ell(\mathbf{G}_\ell^\top \mathbf{G}_\ell)^{-1/2}, \qquad \mathbf{z}_i \overset{d}{=} \mathbf{R}_\ell \mathbf{g}_i, \qquad \mathbf{R}_\ell = \left(\frac{\mathbf{G}_\ell^\top \mathbf{G}_\ell}{\ell}\right)^{-1/2},$$

where the rows $\mathbf{g}_i^\top$ correspond to independent $N(0, \mathbf{I}_r)$ column vectors $\mathbf{g}_i$. Since r is fixed, $\|\mathbf{R}_\ell - \mathbf{I}_r\|_{\mathrm{op}} \to 0$ in probability. Also

$$\max_i \|\mathbf{g}_i\|_2^2 = O_{\mathbb{P}}(\log \ell)$$

by a union bound and the chi-square tail. Hence, uniformly over $\|\mathbf{B}\|_{\mathrm{op}} = 1$,

$$\max_i \left|\mathbf{g}_i^\top \mathbf{R}_\ell \mathbf{B}\mathbf{R}_\ell \mathbf{g}_i - \mathbf{g}_i^\top \mathbf{B}\mathbf{g}_i\right| = o_{\mathbb{P}}(\log \ell). \tag{30}$$

We next prove the required estimate for the independent Gaussian quadratic forms $Q_i(\mathbf{B}) := \mathbf{g}_i^\top \mathbf{B}\mathbf{g}_i$, uniformly over $\|\mathbf{B}\|_{\mathrm{op}} = 1$:

$$\max_{1\le i\le \ell} |Q_i(\mathbf{B})| = (2 + o_{\mathbb{P}}(1)) \log \ell.$$

The upper bound follows from $|Q_i(\mathbf{B})| \le \|\mathbf{g}_i\|_2^2$. For the lower bound, diagonalize $\mathbf{B} = \mathbf{V}\,\mathrm{diag}(\lambda_1, \dots, \lambda_r)\mathbf{V}^\top$ and choose j_* with $|\lambda_{j_*}| = 1$. In the rotated coordinates $\widetilde{\mathbf{g}}_i = \mathbf{V}^\top \mathbf{g}_i$, fix $M > 0$ and define

$$E_i(a, M) := \left\{ \widetilde{g}_{i,j_*}^2 \ge a \log \ell + M, \quad \sum_{j \ne j_*} \widetilde{g}_{i,j}^2 \le M \right\}.$$

On $E_i(a, M)$, the orthogonal contribution has magnitude at most M, so $|Q_i(\mathbf{B})| \ge \widetilde{g}_{i,j_*}^2 - M \ge a \log \ell$. Rotational invariance makes the coordinates independent and gives, uniformly over $\mathbf{B}$,

$$\mathbb{P}(E_i(a, M)) = \mathbb{P}\{\chi_1^2 \ge a \log \ell + M\}\mathbb{P}\{\chi_{r-1}^2 \le M\} \ge c_{a,M,r} \frac{\ell^{-a/2}}{\sqrt{\log \ell}}$$

for every fixed $0 < a < 2$ and all large ℓ, with $\chi_0^2 := 0$ when $r = 1$. Therefore $\mathbb{P}\{\max_i |Q_i(\mathbf{B})| < a \log \ell\} \le \exp[-\ell \mathbb{P}(E_1(a, M))] \to 0$ uniformly in $\mathbf{B}$. Combining this lower bound for every $0 < a < 2$, the chi-square upper bound for every $a > 2$, and equation 30 proves equation 28.

Taking $\mathbf{B} = \mathbf{I}_r$ gives the first bound in equation 29, and taking $\mathbf{B} = \mathbf{e}_1\mathbf{e}_1^\top$ gives the second. □

C.2 Eigenvalues in the Low-Rank Model

The argument proceeds from an exact finite-dimensional quadratic-form representation. Coefficient convergence then yields the detailed limiting laws, while a final compactness argument gives the scale-uniform conclusions stated in the main theorem.

Roadmap Triangularity gives an exact diagonal quadratic form. Haar overlap convergence identifies finite-token limits, the empirical Haar-row law gives the limiting spectral measure and threshold counts, and Gaussian extreme values determine the stable-rank scale.

Exact diagonal formulas. For each token i, define the left singular loading vectors

$$\bar{\mathbf{u}}_{Q,i} := \mathbf{\Sigma}_Q \mathbf{U}_Q^\top \mathbf{e}_i \in \mathbb{R}^{r_Q}, \qquad \bar{\mathbf{u}}_{K,i} := \mathbf{\Sigma}_K \mathbf{U}_K^\top \mathbf{e}_i \in \mathbb{R}^{r_K}. \tag{31}$$

These vectors include both singular values and token-side leverage. With the overlap matrix $\mathbf{H}$ from equation 26, every entry of $\mathbf{Q}\mathbf{K}^\top$ has the form

$$(\mathbf{Q}\mathbf{K}^\top)_{ij} = \bar{\mathbf{u}}_{Q,i}^\top \mathbf{H}\, \bar{\mathbf{u}}_{K,j}. \tag{32}$$

Proposition 3 (Exact eigenvalues for lower-triangular masks)**.** *Assume* $\mathbf{L}$ *is lower triangular. Then the eigenvalues of* $\mathbf{M} = \mathbf{L} \odot \mathbf{Q}\mathbf{K}^\top$ *are exactly*

$$\lambda_i(\mathbf{M}) = L_{ii} \bar{\mathbf{u}}_{Q,i}^\top \mathbf{H}\, \bar{\mathbf{u}}_{K,i}, \qquad i = 1, \dots, \ell. \tag{33}$$

Equivalently,

$$\lambda_i(\mathbf{M}) = L_{ii} \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \sigma_a^Q \sigma_b^K\, u_a^Q(i) u_b^K(i)\, \langle \mathbf{v}_a^Q, \mathbf{v}_b^K \rangle. \tag{34}$$

If Assumption 1 also holds and r_Q, r_K *are fixed as* $d \to \infty$*, then with*

$$Z_{ab}^{(d)} := \sqrt{d}\, \langle \mathbf{v}_a^Q, \mathbf{v}_b^K \rangle, \qquad c_{ab}^{(d)}(i) := \frac{L_{ii} \sigma_a^Q \sigma_b^K u_a^Q(i) u_b^K(i)}{\sqrt{d}},$$

one has the exact representation

$$\lambda_i(\mathbf{M}) = \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} c_{ab}^{(d)}(i) Z_{ab}^{(d)}. \tag{35}$$

Moreover,

$$(Z_{ab}^{(d)})_{a,b} \Rightarrow (G_{ab})_{a,b},$$

where the entries G_{ab} *are independent* $N(0, 1)$.

Proof. Since $\mathbf{L}$ is lower triangular, $\mathbf{M} = \mathbf{L} \odot \mathbf{Q}\mathbf{K}^\top$ is lower triangular. The eigenvalues of a triangular matrix are its diagonal entries, which gives equation 33. Expanding equation 3 gives

$$\mathbf{Q}\mathbf{K}^\top = \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \sigma_a^Q \sigma_b^K \langle \mathbf{v}_a^Q, \mathbf{v}_b^K \rangle \mathbf{u}_a^Q (\mathbf{u}_b^K)^\top,$$

and taking the (i, i) entry gives equation 34. Also,

$$\bar{\mathbf{u}}_{Q,i}^\top \mathbf{H}\, \bar{\mathbf{u}}_{K,i} = \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \sigma_a^Q \sigma_b^K u_a^Q(i) u_b^K(i) \langle \mathbf{v}_a^Q, \mathbf{v}_b^K \rangle,$$

which identifies the compact loading-vector form with the expanded form. The exact representation equation 35 is a rewriting of equation 34, and the convergence of $(Z_{ab}^{(d)})_{a,b}$ follows from Lemma 4. □

Proposition 4 (Common-frame eigenvalue quadratic-form reduction)**.** *Assume that* $\mathbf{L}$ *is unit diagonal and lower triangular and that Assumptions 1 and 2 hold. For each pair* (ℓ, d)*, set*

$$\theta_{ab;\ell,d} := \frac{\sigma_{a;\ell,d}^Q \sigma_{b;\ell,d}^K}{\ell\sqrt{d}}, \qquad Z_{ab;\ell,d} := \sqrt{d}\, \langle \mathbf{v}_{a;\ell,d}^Q, \mathbf{v}_{b;\ell,d}^K \rangle.$$

Let $r := \max(r_Q, r_K)$ *and define the zero-padded matrices*

$$F_{ab;\ell,d} := \theta_{ab;\ell,d} Z_{ab;\ell,d}, \qquad 1 \le a \le r_Q, \quad 1 \le b \le r_K, \qquad \mathbf{B}_{\ell,d} := \frac{\mathbf{F}_{\ell,d} + \mathbf{F}_{\ell,d}^\top}{2}, \tag{36}$$

with all other entries of $\mathbf{F}_{\ell,d}$ *equal to zero. For* $\mathbf{z}_{i;\ell} := \sqrt{\ell}\,(u_1(i), \ldots, u_r(i))^\top$*, one has the exact representation*

$$\lambda_i(\mathbf{M}_{\ell,d}) = \mathbf{z}_{i;\ell}^\top \mathbf{B}_{\ell,d} \mathbf{z}_{i;\ell}, \qquad i = 1, \ldots, \ell. \tag{37}$$

Moreover, $\mathbf{B}_{\ell,d}$ *is independent of the token frame. If, as* $(\ell, d) \to (\infty, \infty)$*,* $\theta_{ab;\ell,d} \to \theta_{ab}$ *for every* a, b*, then*

$$\mathbf{B}_{\ell,d} \Rightarrow \mathbf{B}(\mathbf{G}) := \frac{\mathbf{F}(\mathbf{G}) + \mathbf{F}(\mathbf{G})^\top}{2}, \qquad F_{ab}(\mathbf{G}) = \theta_{ab} G_{ab}, \tag{38}$$

where $\mathbf{F}(\mathbf{G})$ *is zero outside its* $r_Q \times r_K$ *block and the entries of* $\mathbf{G}$ *are independent* $N(0, 1)$ *variables.*

Proof. Proposition 3, the unit diagonal, and the common token frame give

$$\lambda_i(\mathbf{M}_{\ell,d}) = \sum_{a,b} \theta_{ab;\ell,d} Z_{ab;\ell,d} \{\sqrt{\ell} u_a(i)\} \{\sqrt{\ell} u_b(i)\} = \mathbf{z}_{i;\ell}^\top \mathbf{F}_{\ell,d} \mathbf{z}_{i;\ell}.$$

A scalar quadratic form depends only on the symmetric part of its coefficient matrix, proving equation 37. Independence follows from the independence of the token frame and right singular spaces. Finally, Lemma 4 gives $(Z_{ab;\ell,d})_{a,b} \Rightarrow (G_{ab})_{a,b}$; coefficient convergence and the continuous mapping theorem yield equation 38. □

Theorem 9 (Eigenvalues of lower-triangular masked attention)**.** *Assume that* $\mathbf{L}$ *is lower triangular and unit diagonal and that Assumptions 1 and 2 hold. As* $(\ell, d) \to (\infty, \infty)$*, suppose*

$$\theta_{ab;\ell,d} := \frac{\sigma_{a;\ell,d}^Q \sigma_{b;\ell,d}^K}{\ell\sqrt{d}} \longrightarrow \theta_{ab}.$$

Then, for any fixed collection of distinct tokens $i_1, \ldots, i_m$*,*

$$(\lambda_{i_1}(\mathbf{M}_{\ell,d}), \ldots, \lambda_{i_m}(\mathbf{M}_{\ell,d})) \Rightarrow (\Lambda_{i_1}, \ldots, \Lambda_{i_m}), \qquad \Lambda_i := \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \theta_{ab} \xi_{i,a} \xi_{i,b} G_{ab}, \tag{39}$$

where the variables $\xi_{i,a}$ *are independent* $N(0, 1)$ *over token and singular-vector indices, and* (G_{ab}) *is an independent matrix of independent* $N(0, 1)$ *entries. The Gaussian matrix* (G_{ab}) *is shared across all tokens, whereas the token-side variables* $\xi_{i,a}$ *depend on* i*.*

Proof of Theorem 9. By Proposition 4,

$$\lambda_i(\mathbf{M}_{\ell,d}) = \mathbf{z}_{i;\ell}^\top \mathbf{B}_{\ell,d} \mathbf{z}_{i;\ell},$$

where $\mathbf{B}_{\ell,d}$ is independent of the token frame and converges to $\mathbf{B}(\mathbf{G})$. For any fixed distinct tokens $i_1, \dots, i_m$, Lemma 5 gives

$$(\mathbf{z}_{i_1;\ell}, \dots, \mathbf{z}_{i_m;\ell}) \Rightarrow (\boldsymbol{\xi}_{i_1}, \dots, \boldsymbol{\xi}_{i_m}),$$

where the limiting vectors are independent $N(0, \mathbf{I}_r)$. The token-frame variables are independent of $\mathbf{B}_{\ell,d}$, so this convergence is joint with equation 38, with independent limits. The continuous mapping theorem now gives

$$\mathbf{z}_{i_p;\ell}^\top \mathbf{B}_{\ell,d} \mathbf{z}_{i_p;\ell} \Rightarrow \boldsymbol{\xi}_{i_p}^\top \mathbf{B}(\mathbf{G}) \boldsymbol{\xi}_{i_p} = \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \theta_{ab} \xi_{i_p,a} \xi_{i_p,b} G_{ab}$$

simultaneously for $p = 1, \dots, m$. The same matrix $\mathbf{G}$ appears for every token because the right singular spaces are shared across all rows. □

Proposition 5 (Covariance of the limiting eigenvalues)**.** *In the setting of Theorem 9, define* $D_{\boldsymbol{\theta}} := \sum_{a=1}^{\min(r_Q,r_K)} \theta_{aa}^2, \qquad F_{\boldsymbol{\theta}} := \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \theta_{ab}^2$. *For any fixed collection of distinct tokens* $i_1, \dots, i_m$*, the limiting variables in equation 39 are independent conditionally on* $\mathbf{G}$*. Unconditionally, each has mean zero, and their covariance matrix satisfies*

$$\mathrm{Cov}\left(\Lambda_{i_p}, \Lambda_{i_q}\right) = \begin{cases} F_{\boldsymbol{\theta}} + 2D_{\boldsymbol{\theta}}, & p = q, \\ D_{\boldsymbol{\theta}}, & p \neq q. \end{cases} \tag{40}$$

Thus the limiting variables are generally dependent unconditionally because they share $\mathbf{G}$*. If* $D_{\boldsymbol{\theta}} = 0$*, distinct limiting variables are uncorrelated but may still be dependent through* $\mathbf{G}$*.*

Proof. Conditional on $\mathbf{G}$, each Λ_i is a function of the token-side vector $(\xi_{i,a})_a$, and these vectors are independent across distinct tokens. Hence the limiting variables are conditionally independent. Since $\mathbb{E}G_{ab} = 0$, one also has $\mathbb{E}\Lambda_i = 0$. For distinct tokens $i_p \neq i_q$, independence of the token-side vectors gives

$$\mathbb{E}\Lambda_{i_p}\Lambda_{i_q} = \sum_{a,b,c,d} \theta_{ab}\theta_{cd} \mathbb{E}[G_{ab}G_{cd}] \mathbb{E}[\xi_{i_p,a}\xi_{i_p,b}] \mathbb{E}[\xi_{i_q,c}\xi_{i_q,d}].$$

Using $\mathbb{E}[G_{ab}G_{cd}] = \mathbf{1}_{\{a=c,\ b=d\}}$ and $\mathbb{E}[\xi_{i,a}\xi_{i,b}] = \mathbf{1}_{\{a=b\}}$, only the diagonal terms $a = b$ survive. Therefore

$$\mathbb{E}\Lambda_{i_p}\Lambda_{i_q} = \sum_{a=1}^{\min(r_Q,r_K)} \theta_{aa}^2 = D_{\boldsymbol{\theta}}, \qquad p \neq q.$$

For a single token,

$$\mathbb{E}\Lambda_i^2 = \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \theta_{ab}^2 \mathbb{E}[\xi_{i,a}^2 \xi_{i,b}^2] = F_{\boldsymbol{\theta}} + 2D_{\boldsymbol{\theta}},$$

because $\mathbb{E}\xi^4 = 3$ on the diagonal $a = b$ and $\mathbb{E}[\xi_a^2\xi_b^2] = 1$ off the diagonal. This proves equation 40. □

For the coefficient matrices in Proposition 4, and for any symmetric matrix $\mathbf{C} \in \mathbb{R}^{r \times r}$, write

$$\nu_{\mathbf{C}} := \mathcal{L}_{\boldsymbol{\xi}}(\boldsymbol{\xi}^\top \mathbf{C} \boldsymbol{\xi}), \qquad \boldsymbol{\xi} \sim N(0, \mathbf{I}_r).$$

Proposition 6 (Non-collapse of the eigenvalue empirical law)**.** *Assume the setting of Theorem 9,* $\boldsymbol{\theta} \neq 0$*. Let* $\mu_{\mathrm{eig};\ell,d} := \ell^{-1} \sum_{i=1}^{\ell} \delta_{\lambda_i(\mathbf{M}_{\ell,d})}$*, and let* d_{BL} *denote the bounded-Lipschitz metric on probability measures. Then*

$$d_{\mathrm{BL}}\left(\mu_{\mathrm{eig};\ell,d}, \nu_{\mathbf{B}_{\ell,d}}\right) \longrightarrow 0 \quad \textit{in probability}, \qquad (\mathbf{B}_{\ell,d}, \mu_{\mathrm{eig};\ell,d}) \Rightarrow (\mathbf{B}(\mathbf{G}), \nu_{\mathbf{B}(\mathbf{G})}) \tag{41}$$

as a matrix–random-measure pair. Moreover $\nu_{\mathbf{B}(\mathbf{G})}$ *is nondegenerate and has unbounded support almost surely, so* $\mu_{\mathrm{eig};\ell,d}$ *does not converge weakly to* δ_0 *in probability.*

Proof. Let μ_ℓ be the empirical Haar-row measure in Lemma 6, and put $\gamma_r = N(0, \mathbf{I}_r)$. We first make the uniformity in the bounded-Lipschitz metric explicit. Let P_A be the Euclidean projection onto the ball of radius A. For $\|\mathbf{C}\|_{\mathrm{op}} \le R$, the map $\mathbf{x} \mapsto (P_A\mathbf{x})^\top \mathbf{C}(P_A\mathbf{x})$ is $2RA$-Lipschitz. Also, $\int \|\mathbf{x}\|_2^2 \, d\mu_\ell = r$ exactly, by orthonormality of the Haar columns. Thus truncating the quadratic forms by P_A gives

$$\sup_{\|\mathbf{C}\|_{\mathrm{op}} \le R} d_{\mathrm{BL}}\big((\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{C}\mathbf{x})_{\#}\mu_\ell, \nu_{\mathbf{C}}\big) \le \max\{1, 2RA\}\, d_{\mathrm{BL}}(\mu_\ell, \gamma_r) + \frac{4r}{A^2},$$

where $\#$ denotes pushforward and the supremum is over symmetric matrices. First let $\ell \to \infty$ and then $A \to \infty$. Lemma 6 and tightness of $\mathbf{B}_{\ell,d}$ prove the first assertion in equation 41. The map $\mathbf{C} \mapsto \nu_{\mathbf{C}}$ is continuous for the weak topology, so equation 38 and Slutsky's theorem give the joint convergence in equation 41. If $\boldsymbol{\theta} \neq 0$, the zero-padded symmetric Gaussian matrix $\mathbf{B}(\mathbf{G})$ is nonzero almost surely. Conditional on $\mathbf{G}$, its Gaussian quadratic form has variance $2\|\mathbf{B}(\mathbf{G})\|_F^2 > 0$, so the limiting law is nondegenerate. Along an eigenvector corresponding to a nonzero eigenvalue, the quadratic form is unbounded in magnitude. Since the Gaussian law has full support, $\nu_{\mathbf{B}(\mathbf{G})}(\{x : |x| > R\}) > 0$ for every finite $R > 0$, proving the unbounded-support assertion. □

Proposition 7 (Effective eigenvalue counts and tail probabilities)**.** *Assume the setting of Theorem 9, $\boldsymbol{\theta} \neq 0$. Define*

$$\overline{\Phi}(t) := \mathbb{P}\{Z > t\}, \qquad Z \sim N(0,1),$$

and

$$N_\varepsilon^{|\cdot|} := \#\{i : |\lambda_i(\mathbf{M}_{\ell,d})| > \varepsilon\}.$$

For $\boldsymbol{\xi} \sim N(0, \mathbf{I}_r)$, set

$$V_{\boldsymbol{\theta}}(\boldsymbol{\xi}) := \sum_{a=1}^{r_Q}\sum_{b=1}^{r_K} \theta_{ab}^2 \xi_a^2 \xi_b^2, \qquad s_{\boldsymbol{\theta}}^2 := \mathbb{E}V_{\boldsymbol{\theta}}(\boldsymbol{\xi}) = \sum_{a,b}(1 + 2\mathbf{1}_{\{a=b\}})\theta_{ab}^2.$$

For $\varepsilon > 0$, define

$$p_\varepsilon^{+}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi}}\overline{\Phi}\left(\frac{\varepsilon}{\sqrt{V_{\boldsymbol{\theta}}(\boldsymbol{\xi})}}\right), \qquad p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}) = 2p_\varepsilon^{+}(\boldsymbol{\theta}), \tag{42}$$

where the integrand is set equal to zero when $V_{\boldsymbol{\theta}}(\boldsymbol{\xi}) = 0$. Then, for every $\varepsilon > 0$,

$$\frac{\mathbb{E}N_\varepsilon^{|\cdot|}}{\ell} \longrightarrow p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}). \tag{43}$$

Moreover, for every $u > 0$,

$$\frac{3}{70}\overline{\Phi}\left(\frac{\sqrt{2}\,\varepsilon}{s_{\boldsymbol{\theta}}}\right) \le p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}) \le 2\overline{\Phi}(\varepsilon/u) + 2s_{\boldsymbol{\theta}}^2/u^2. \tag{44}$$

Also,

$$p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}) \le C\exp\left\{-c\left(\frac{\varepsilon}{s_{\boldsymbol{\theta}}}\right)^{2/3}\right\}, \tag{45}$$

for universal constants $c, C > 0$. For the finite coefficient matrix, set

$$q_{\varepsilon;\ell,d} := \mathbb{P}_{\boldsymbol{\xi}}\left\{\left|\boldsymbol{\xi}^\top \mathbf{B}_{\ell,d}\boldsymbol{\xi}\right| > \varepsilon\right\}.$$

Then

$$\frac{N_\varepsilon^{|\cdot|}}{\ell} - q_{\varepsilon;\ell,d} \longrightarrow 0 \quad \textit{in probability}, \qquad \left(\mathbf{B}_{\ell,d}, \frac{N_\varepsilon^{|\cdot|}}{\ell}\right) \Rightarrow \left(\mathbf{B}(\mathbf{G}), p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}; \mathbf{G})\right), \tag{46}$$

where

$$p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}; \mathbf{G}) := \mathbb{P}_{\boldsymbol{\xi}}\left\{\left|\sum_{a,b}\theta_{ab}\xi_a\xi_b G_{ab}\right| > \varepsilon\right\}.$$

In particular, $N_\varepsilon^{|\cdot|}/\ell \Rightarrow p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}; \mathbf{G})$. In the rank-one case,

$$p_\varepsilon^{+}(\theta) = \mathbb{E}_\xi\overline{\Phi}\left(\frac{\varepsilon}{|\theta|\xi^2}\right) = \int_0^\infty \overline{\Phi}\left(\frac{\varepsilon}{|\theta|x}\right)\frac{e^{-x/2}}{\sqrt{2\pi x}}\,dx. \tag{47}$$

Proof. Theorem 9 gives, for a fixed token,

$$\lambda_i(\mathbf{M}_{\ell,d}) \Rightarrow \sum_{a=1}^{r_Q}\sum_{b=1}^{r_K}\theta_{ab}\xi_a\xi_b G_{ab},$$

where (G_{ab}) is independent of $\boldsymbol{\xi}$. Conditional on $\boldsymbol{\xi}$, the right-hand side is a centered Gaussian random variable with variance $V_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. Hence

$$\mathbb{P}_{\mathbf{G}}\left\{\sum_{a,b}\theta_{ab}\xi_a\xi_b G_{ab} > \varepsilon \,\middle|\, \boldsymbol{\xi}\right\} = \overline{\Phi}\left(\frac{\varepsilon}{\sqrt{V_{\boldsymbol{\theta}}(\boldsymbol{\xi})}}\right),$$

with value zero when the variance is zero. Averaging over $\boldsymbol{\xi}$ gives the one-sided formula defining $p_\varepsilon^+(\boldsymbol{\theta})$. The limiting variable is symmetric about zero after averaging over $\mathbf{G}$, so the two-sided formula is twice the one-sided formula.

Linearity of expectation and exchangeability of the Haar token rows give

$$\frac{\mathbb{E}N_\varepsilon^{|\cdot|}}{\ell} = \mathbb{P}\{|\lambda_1(\mathbf{M}_{\ell,d})| > \varepsilon\} \longrightarrow p_\varepsilon^{|\cdot|}(\boldsymbol{\theta}),$$

which proves equation 43. The indicator discontinuity is harmless: a nonzero entry of $\boldsymbol{\theta}$ ensures $V_{\boldsymbol{\theta}}(\boldsymbol{\xi}) > 0$ almost surely, so the conditional Gaussian law, and hence its mixture, has no atoms.

For the upper estimate in equation 44, split on $V_{\boldsymbol{\theta}}(\boldsymbol{\xi}) \le u^2$. On this event the integrand is at most $\overline{\Phi}(\varepsilon/u)$, while the complement has probability at most $s_{\boldsymbol{\theta}}^2/u^2$ by Markov's inequality.

For the lower estimate, Paley–Zygmund applied to $V_{\boldsymbol{\theta}}(\boldsymbol{\xi})$ gives, with $K_{\boldsymbol{\theta}} := \mathbb{E}\{V_{\boldsymbol{\theta}}(\boldsymbol{\xi})^2\}/s_{\boldsymbol{\theta}}^4$,

$$\mathbb{P}\left\{V_{\boldsymbol{\theta}}(\boldsymbol{\xi}) \ge \frac{s_{\boldsymbol{\theta}}^2}{2}\right\} \ge \frac{1}{4K_{\boldsymbol{\theta}}}.$$

On this event, $\overline{\Phi}\{\varepsilon/V_{\boldsymbol{\theta}}(\boldsymbol{\xi})^{1/2}\} \ge \overline{\Phi}(\sqrt{2}\,\varepsilon/s_{\boldsymbol{\theta}})$. Multiplying by the outer factor two in $p_\varepsilon^{|\cdot|} = 2p_\varepsilon^+$ gives the lower bound in equation 44. Finally, $K_{\boldsymbol{\theta}} \le 35/3$ because

$$\frac{\mathbb{E}\{\xi_a^2\xi_b^2\xi_{a'}^2\xi_{b'}^2\}}{(1+2\mathbf{1}_{\{a=b\}})(1+2\mathbf{1}_{\{a'=b'\}})} \le \frac{35}{3}$$

for all indices; the maximum occurs when $a=b=a'=b'$.

The Gaussian chaos bound equation 45 is the standard hypercontractive tail estimate for a Gaussian polynomial of degree at most three, applied to $\sum_{a,b}\theta_{ab}\xi_a\xi_b G_{ab}$. For the empirical count, apply the compact-uniform indicator statement of Lemma 6 and use tightness of $\mathbf{B}_{\ell,d}$. This gives $N_\varepsilon^{|\cdot|}/\ell - q_{\varepsilon;\ell,d} \to 0$ in probability. Finally, the map $\mathbf{C} \mapsto \mathbb{P}_{\boldsymbol{\xi}}\{|\boldsymbol{\xi}^\top\mathbf{C}\boldsymbol{\xi}| > \varepsilon\}$ is continuous at every nonzero $\mathbf{C}$. Combining this fact with equation 38 proves the joint convergence in equation 46. In the rank-one case, write $x = \xi^2$ and use the $\chi^2(1)$ density to obtain equation 47. □

Proposition 8 (Eigenvalue rank asymptotics under coefficient convergence)**.** *Assume the setting of Theorem 9 with $\boldsymbol{\theta} \neq 0$. For any nonzero symmetric matrix $\mathbf{C} \in \mathbb{R}^{r\times r}$, define*

$$\rho(\mathbf{C}) := \|\mathbf{C}\|_{\mathrm{op}}, \qquad Y_{\mathbf{C}} := \boldsymbol{\xi}^\top\mathbf{C}\boldsymbol{\xi}, \qquad \mu_1(\mathbf{C}) := \mathbb{E}_{\boldsymbol{\xi}}|Y_{\mathbf{C}}|,$$
$$\eta_1(\mathbf{C}) := \mathbb{E}_{\boldsymbol{\xi}}\{|Y_{\mathbf{C}}|\log|Y_{\mathbf{C}}|\}, \qquad 0\log 0 := 0.$$

Write

$$\rho_*(\mathbf{G}) := \rho\{\mathbf{B}(\mathbf{G})\}, \qquad \mu_1(\mathbf{G}) := \mu_1\{\mathbf{B}(\mathbf{G})\}, \qquad \eta_1(\mathbf{G}) := \eta_1\{\mathbf{B}(\mathbf{G})\}.$$

Then

$$\frac{\log\ell}{\ell}\,\mathrm{srank}_\lambda(\mathbf{M}_{\ell,d}) \Rightarrow \frac{\mu_1(\mathbf{G})}{2\rho_*(\mathbf{G})}, \qquad \frac{\sqrt{2}}{18} \le \frac{\mu_1(\mathbf{G})}{2\rho_*(\mathbf{G})} \le \frac{r}{2} \quad \textit{almost surely}. \tag{48}$$

Moreover,

$$\log \mathrm{erank}_\lambda(\mathbf{M}_{\ell,d}) - \log\ell \Rightarrow \log\mu_1(\mathbf{G}) - \frac{\eta_1(\mathbf{G})}{\mu_1(\mathbf{G})}, \tag{49}$$

and the exponential of the random variable on the right belongs to $[1/81, 1]$ almost surely.

Proof. Let $\mathbf{B}_{\ell,d}$ be the finite coefficient matrix from Proposition 4 and put $\rho_{\ell,d} := \|\mathbf{B}_{\ell,d}\|_{\mathrm{op}}$. Since $\boldsymbol{\theta} \neq 0$, $\mathbf{B}_{\ell,d} \neq 0$ almost surely for all sufficiently large $\min\{\ell, d\}$, and $\mathbf{B}(\mathbf{G}) \neq 0$ almost surely. Indeed, the finite matrix has $B_{11;\ell,d} = \theta_{11;\ell,d} Z_{11;\ell,d} \neq 0$ almost surely by Lemma 4 and positivity of the nonzero singular values. Lemma 8 gives

$$\frac{\max_i |\lambda_i(\mathbf{M}_{\ell,d})|}{2\rho_{\ell,d} \log \ell} \longrightarrow 1 \qquad \text{in probability.}$$

Conditioning on $\mathbf{B}_{\ell,d}$ and applying Lemma 6 to $|\mathbf{z}^\top \mathbf{B}_{\ell,d} \mathbf{z}|$ and $|\mathbf{z}^\top \mathbf{B}_{\ell,d} \mathbf{z}| \log |\mathbf{z}^\top \mathbf{B}_{\ell,d} \mathbf{z}|$ gives

$$\frac{1}{\ell} \sum_{i=1}^{\ell} |\lambda_i(\mathbf{M}_{\ell,d})| - \mu_1(\mathbf{B}_{\ell,d}) \longrightarrow 0, \tag{50}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} |\lambda_i(\mathbf{M}_{\ell,d})| \log |\lambda_i(\mathbf{M}_{\ell,d})| - \eta_1(\mathbf{B}_{\ell,d}) \longrightarrow 0 \tag{51}$$

in probability. Here $t \mapsto |t| \log |t|$, extended by 0 at $t = 0$, is continuous and is bounded by a polynomial envelope after substituting the quadratic form (for example, by $C(1 + \|\mathbf{z}\|_2^4)$ on bounded coefficient sets), so the uniform-integrability clause of Lemma 6 applies. The convergence is uniform on compact coefficient sets, and $\mathbf{B}_{\ell,d}$ is tight. Thus these estimates hold jointly with $\mathbf{B}_{\ell,d} \Rightarrow \mathbf{B}(\mathbf{G})$. Both μ_1 and η_1 are continuous by the same polynomial envelope, and $\mu_1(\mathbf{B}(\mathbf{G})), \rho_*(\mathbf{G}) > 0$ almost surely. Consequently, $\mu_1(\mathbf{B}_{\ell,d})^{-1}$ and $\rho_{\ell,d}^{-1}$ are tight, which justifies the divisions below. The stable-rank identity gives the convergence in equation 48.

It remains to bound its random prefactor. Diagonalizing $\mathbf{B}(\mathbf{G})$ gives

$$Y_{\mathbf{B}(\mathbf{G})} \overset{d}{=} \sum_{k=1}^{r} \rho_k(\mathbf{G}) \chi_k^2,$$

where $\rho_k(\mathbf{G})$ are the eigenvalues of $\mathbf{B}(\mathbf{G})$. Therefore

$$\mu_1(\mathbf{G}) \leq \sum_{k=1}^{r} |\rho_k(\mathbf{G})| \leq r \rho_*(\mathbf{G}).$$

For the lower bound, apply Lemma 7 with $\mathbf{C} = \mathbf{B}(\mathbf{G})$ to obtain $\mu_1(\mathbf{G}) \geq \sqrt{2}\, \rho_*(\mathbf{G})/9$, proving both prefactor bounds.

For the entropy proxy, set $w_i := |\lambda_i(\mathbf{M}_{\ell,d})|$ and $p_i := w_i / \sum_j w_j$. Then

$$-\sum_{i=1}^{\ell} p_i \log p_i = \log \left(\sum_{i=1}^{\ell} w_i \right) - \frac{\sum_{i=1}^{\ell} w_i \log w_i}{\sum_{i=1}^{\ell} w_i}.$$

Substituting equation 50 and equation 51 first gives

$$\log \operatorname{erank}_\lambda(\mathbf{M}_{\ell,d}) - \log \ell = \log \mu_1(\mathbf{B}_{\ell,d}) - \frac{\eta_1(\mathbf{B}_{\ell,d})}{\mu_1(\mathbf{B}_{\ell,d})} + o_{\mathbb{P}}(1).$$

Using equation 38 then proves equation 49 by the continuous mapping theorem.

Finally, put $W = |Y_{\mathbf{B}(\mathbf{G})}|$ and $X = W/\mu_1(\mathbf{G})$. Jensen's inequality under the probability measure with density X gives

$$\frac{\eta_1(\mathbf{G})}{\mu_1(\mathbf{G})} - \log \mu_1(\mathbf{G}) = \mathbb{E}(X \log X) \leq \log \mathbb{E} X^2 \leq \log 81,$$

where the last inequality follows from $\mu_1(\mathbf{G}) \geq \|Y_{\mathbf{B}(\mathbf{G})}\|_2/9$. Hence the limiting entropy prefactor is at least $1/81$. It is at most one because the entropy of a probability vector on ℓ points is at most $\log \ell$. □

Proof of Theorem 1. For the proof only, along any cofinal sequence $(\ell_k, d_k) \to (\infty, \infty)$ write

$$\boldsymbol{\theta}^{(k)} := \boldsymbol{\theta}_{\ell_k, d_k}, \qquad s_k := s_{\ell_k, d_k}, \qquad \mathbf{M}_k := \mathbf{M}_{\ell_k, d_k}.$$

The scale s_k need not be bounded above. Put $\widetilde{\boldsymbol{\theta}}^{(k)} := \boldsymbol{\theta}^{(k)}/s_k$. These normalized arrays lie in the compact ellipsoid

$$\sum_{a,b}(1 + 2\mathbf{1}_{\{a=b\}})(\widetilde{\theta}^{(k)}_{ab})^2 = 1.$$

At the finite-dimensional quadratic-form level, multiplying every coefficient by s_k multiplies all eigenvalues by s_k, and therefore leaves both eigenvalue rank proxies unchanged.

For part (i), choose the cofinal sequence so that $\mathbb{E}\#\{i : |\lambda_i(\mathbf{M}_k)| > \varepsilon\}/\ell_k$ converges to the joint liminf. The normalization below is used only to express the fixed threshold relative to the common eigenvalue scale. Passing again to a subsequence, suppose $\widetilde{\boldsymbol{\theta}}^{(k)} \to \widetilde{\boldsymbol{\theta}}$ and $\varepsilon/s_k \to t$. If $s_- < \infty$, then $s_k \geq s_- - o(1)$ and $0 \leq t \leq \varepsilon/s_-$. If $s_- = \infty$, then $s_k \to \infty$ and $t = 0$; interpret $\varepsilon/s_- = 0$ in this case. Apply the normalized quadratic-form representation at threshold ε/s_k. The compactness and continuity argument in Proposition 7 extends its count limit to this convergent threshold sequence. Its limiting scale is one, so equation 44 gives

$$\frac{\mathbb{E}\#\{i : |\lambda_i(\mathbf{M}_k)| > \varepsilon\}}{\ell_k} \longrightarrow p_t^{|\cdot|}(\widetilde{\boldsymbol{\theta}}) \geq \frac{3}{70}\overline{\Phi}\left(\sqrt{2}\,t\right) \geq \frac{3}{70}\overline{\Phi}\left(\frac{\sqrt{2}\,\varepsilon}{s_-}\right)$$

along this further subsequence, with $p_0^{|\cdot|}(\widetilde{\boldsymbol{\theta}}) := 1$. When $t = 0$, use the atomlessness of the nonzero limiting quadratic form to obtain the same count convergence by approximation from positive thresholds. The positive expected tail bound rules out convergence of the empirical law to δ_0 in probability, since the tail fractions are bounded by one. Under coefficient convergence, Proposition 6 gives the asserted random empirical-law limit and its unbounded support. This proves part (i).

For part (ii), start from an arbitrary cofinal sequence and pass to a further subsequence on which $\widetilde{\boldsymbol{\theta}}^{(k)} \to \widetilde{\boldsymbol{\theta}} \neq 0$. The normalized coefficient matrix is tight and converges to the nonzero Gaussian coefficient matrix associated with $\widetilde{\boldsymbol{\theta}}$. The proof of Proposition 8, which uses only scale-invariant ratios, therefore applies verbatim to this normalized array. It gives the bounds in equation 48 and equation 49, with deterministic lower and upper bounds independent of the chosen coefficient limit. Rescaling back by s_k does not change either rank proxy. Hence

$$\mathrm{srank}_\lambda(\mathbf{M}_k) = \Theta_{\mathbb{P}}\left(\frac{\ell_k}{\log \ell_k}\right), \qquad \mathrm{erank}_\lambda(\mathbf{M}_k) = \Theta_{\mathbb{P}}(\ell_k)$$

along every such further subsequence. The subsequence criterion proves the joint-limit conclusions in equation 5. □

Remark 5 (The fixed-token eigenvalue scale). For the unit-diagonal causal masks above, a singular-value pair contributes at order one to a fixed-token eigenvalue when $\sigma_a^Q \sigma_b^K \asymp \ell\sqrt{d}$. More generally, the natural fixed-token scale is $s_{\ell,d}$: all fixed-token eigenvalues converge to zero in probability if every product is $o(\ell\sqrt{d})$, whereas $s_{\ell,d} \to \infty$ gives diverging fixed-token magnitudes in probability. The latter follows by normalizing the coefficient array and applying the same compactness argument as above; every subsequential normalized limit has no atom at zero. This statement concerns fixed tokens; the maximum eigenvalue magnitude already grows on the $\log \ell$ scale when the coefficients converge to a nonzero array.

Remark 6 (The random spectral-radius prefactor). The quantity $\rho_*(\mathbf{G})$ above is a random scale in this fixed-rank setting. It is the largest absolute eigenvalue of the finite symmetric variance-profile matrix $\mathbf{B}(\mathbf{G}) = \{\mathbf{F}(\mathbf{G}) + \mathbf{F}(\mathbf{G})^\top\}/2$, where $F_{ab}(\mathbf{G}) = \theta_{ab}G_{ab}$. The circular law concerns the eigenvalues of large non-Hermitian Ginibre matrices after a $1/\sqrt{r}$ normalization. It does not apply here.

C.3 Singular Values in the Low-Rank Model

We first isolate the algebra common to every fixed rank. We then prove the mask facts used by the sharper rank-one analysis, establish the rank-one empirical-law and constant-level refinements, and finally extend the rank bounds and principal-inverse empirical nondegeneracy to arbitrary fixed r_Q, r_K.

Roadmap The first proposition factors the low-rank score matrix into a finite coefficient matrix and diagonal token loadings. The next subsections establish the mask norm and inverse estimates, use them to prove the rank-one statements, and then replace the scalar coefficient by a fixed-dimensional symmetric coefficient part for the arbitrary-rank extension.

Proposition 9 (Low-rank singular-value reduction)**.** *Under the decomposition equation 3,*

$$\mathbf{M} = \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \sigma_a^Q \sigma_b^K \langle \mathbf{v}_a^Q, \mathbf{v}_b^K \rangle \operatorname{diag}(\mathbf{u}_a^Q)\, \mathbf{L}\, \operatorname{diag}(\mathbf{u}_b^K). \tag{52}$$

In the rank-one case,

$$s_j(\mathbf{M}) = |\sigma^Q \sigma^K \langle \mathbf{v}^Q, \mathbf{v}^K \rangle|\, s_j\{\operatorname{diag}(\mathbf{u}^Q)\mathbf{L} \operatorname{diag}(\mathbf{u}^K)\}.$$

Under Assumptions 1 and 2, write

$$\mathbf{u} := \mathbf{u}^Q = \mathbf{u}^K, \qquad w_i := \sqrt{\ell}\,|u_i|, \qquad \mathbf{T}_\ell := \operatorname{diag}(\mathbf{w})\mathbf{L} \operatorname{diag}(\mathbf{w}),$$

and define

$$\alpha_{\ell,d} := \frac{\sigma_{\ell,d}^Q \sigma_{\ell,d}^K}{\ell\sqrt{d}}, \qquad Z_{\ell,d} := \sqrt{d}\,\langle \mathbf{v}_{\ell,d}^Q, \mathbf{v}_{\ell,d}^K \rangle.$$

Then

$$s_j(\mathbf{M}) = |\alpha_{\ell,d} Z_{\ell,d}| s_j(\mathbf{T}_\ell), \qquad j = 1, \ldots, \ell. \tag{53}$$

Consequently, almost surely, all three scale-invariant singular-value rank proxies of $\mathbf{M}$ *equal those of* $\mathbf{T}_\ell$*.*

Proof. From the singular value decompositions,

$$\mathbf{Q}\mathbf{K}^\top = \sum_{a=1}^{r_Q} \sum_{b=1}^{r_K} \sigma_a^Q \sigma_b^K \langle \mathbf{v}_a^Q, \mathbf{v}_b^K \rangle \mathbf{u}_a^Q (\mathbf{u}_b^K)^\top$$

because $(\mathbf{v}_a^Q)^\top \mathbf{v}_b^K = \langle \mathbf{v}_a^Q, \mathbf{v}_b^K \rangle$. For any pair of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^\ell$, the Hadamard product with a rank-one outer product can be written entrywise as

$$\{\mathbf{L} \odot (\mathbf{x}\mathbf{y}^\top)\}_{ij} = L_{ij} x_i y_j = \{\operatorname{diag}(\mathbf{x})\mathbf{L} \operatorname{diag}(\mathbf{y})\}_{ij}.$$

Applying this identity with $\mathbf{x} = \mathbf{u}_a^Q$ and $\mathbf{y} = \mathbf{u}_b^K$ gives

$$\mathbf{L} \odot (\mathbf{u}_a^Q (\mathbf{u}_b^K)^\top) = \operatorname{diag}(\mathbf{u}_a^Q)\mathbf{L} \operatorname{diag}(\mathbf{u}_b^K)$$

term by term, proving equation 52. In the rank-one case the sum has a single term, so $\mathbf{M}$ is a scalar multiple of $\operatorname{diag}(\mathbf{u}^Q)\mathbf{L} \operatorname{diag}(\mathbf{u}^K)$. Singular values are homogeneous under scalar multiplication, which gives the first rank-one formula. Under the common-frame assumption, left and right multiplication by the diagonal sign matrix of $\mathbf{u}$ does not change singular values, and

$$\operatorname{diag}(\mathbf{u})\mathbf{L} \operatorname{diag}(\mathbf{u}) \quad \text{has the same singular values as} \quad \ell^{-1}\mathbf{T}_\ell.$$

Since $\sigma_{\ell,d}^Q \sigma_{\ell,d}^K \langle \mathbf{v}_{\ell,d}^Q, \mathbf{v}_{\ell,d}^K \rangle / \ell = \alpha_{\ell,d} Z_{\ell,d}$, this proves equation 53. Finally, $Z_{\ell,d} \neq 0$ almost surely, so the common scalar is nonzero almost surely and cancels from every scale-invariant rank proxy. □

C.3.1 Mask Properties

The rank-one refinements and fixed-rank extension use three elementary properties of causal masks. Proposition 10 identifies exactly when a nonexpansive semiseparable mask satisfies Assumption 3, Proposition 11 controls the principal inverses needed for empirical-law nondegeneracy, and Lemma 9 supplies the causal Volterra growth used below.

Roadmap We first record the exact singular values of the causal summation matrix. We then prove the semiseparable gain criterion and the principal-inverse bound; these two estimates are the only mask-specific inputs needed in the rank-one and fixed-rank arguments that follow.

Lemma 9 (Causal summation-matrix norms)**.** *Let $\mathbf{J}_m \in \mathbb{R}^{m\times m}$ be defined by $(\mathbf{J}_m)_{ij} = \mathbf{1}_{\{i\ge j\}}$. Its singular values are*

$$s_k(\mathbf{J}_m) = \left\{ 2\sin\left(\frac{(2k-1)\pi}{4m+2}\right)\right\}^{-1}, \qquad k = 1,\dots,m. \tag{54}$$

In particular, for a universal constant $c_J > 0$,

$$\|\mathbf{J}_m\|_{\mathrm{op}} \ge \frac{m}{2}, \qquad \|\mathbf{J}_m\|_* \ge c_J m\log(m+1). \tag{55}$$

Proof. The inverse of $\mathbf{J}_m$ is the first-difference matrix. Diagonalizing $(\mathbf{J}_m^{-1})^\top \mathbf{J}_m^{-1}$ gives equation 54. The operator-norm bound follows from the $k = 1$ singular value and $\sin x \le x$. The same inequality and $\sum_{k=1}^m (2k-1)^{-1} \ge \frac{1}{2}\log(m+1)$ give

$$\|\mathbf{J}_m\|_* \ge \frac{2m+1}{\pi}\sum_{k=1}^m \frac{1}{2k-1} \ge c_J m\log(m+1).$$

□

Proposition 10 (Equivalent uniform-forgetting conditions)**.** *For each ℓ, let $\mathbf{L}^{(\ell)}$ be a semiseparable mask of the form equation 2, with coefficients $a_t^{(\ell)} \in [0,1]$, and write*

$$P_{j,i}^{(\ell)} := \prod_{t=j+1}^{i} a_t^{(\ell)}, \qquad R_\ell := \max_{1\le i\le \ell}\sum_{j=1}^{i} P_{j,i}^{(\ell)}, \qquad C_\ell := \max_{1\le j\le \ell}\sum_{i=j}^{\ell} P_{j,i}^{(\ell)}.$$

The following conditions are equivalent:

(a) There are an integer $m \ge 1$ and a constant $\rho < 1$ such that

$$\sup_{\ell\ge m+1}\max_{1\le j\le \ell-m} P_{j,j+m}^{(\ell)} \le \rho. \tag{56}$$

(b) $\sup_\ell R_\ell < \infty$.

(c) $\sup_\ell C_\ell < \infty$.

(d) $\sup_\ell \|\mathbf{L}^{(\ell)}\|_{\mathrm{op}} < \infty$.

Under condition (a)*, the entries obey the exponential envelope*

$$P_{j,i}^{(\ell)} \le \rho^{\lfloor (i-j)/m\rfloor}, \qquad 1 \le j \le i \le \ell, \tag{57}$$

where $\rho^0 := 1$ also when $\rho = 0$, and

$$R_\ell \le \frac{m}{1-\rho}, \qquad C_\ell \le \frac{m}{1-\rho}, \qquad \|\mathbf{L}^{(\ell)}\|_{\mathrm{op}} \le \frac{m}{1-\rho}. \tag{58}$$

In particular, the pointwise condition $a_t^{(\ell)} \le a < 1$ is the special case $m = 1$ and $\rho = a$; it is sufficient but not necessary for uniform forgetting.

Proof. Suppose first that equation 56 holds. Partition the $i - j$ transition factors into $\lfloor (i-j)/m\rfloor$ consecutive full blocks of length m and one remainder block. Every full block contributes at most ρ, while the remainder contributes at most one. This proves equation 57. Summing the envelope over the lag $i - j$ gives

$$\sum_{k=0}^{\infty} \rho^{\lfloor k/m\rfloor} = \frac{m}{1-\rho},$$

which proves the row- and column-memory bounds in equation 58. Schur's test then proves its operator-norm bound. Thus (a) implies (b)–(d).

Conversely, suppose $\sup_\ell R_\ell \le K_R$. For any block $j, \dots, j+m$, the row ending at $j+m$ contains the $m+1$ suffix products of that block. Since all transition coefficients lie in $[0,1]$, each suffix product is at least the full product, and hence

$$(m+1)P^{(\ell)}_{j,j+m} \le R_\ell \le K_R.$$

Choosing m so that $m+1 > K_R$ proves (a) with $\rho = K_R/(m+1) < 1$. The same argument with the $m+1$ prefix products proves (c) implies (a).

It remains to show that (d) implies (a). Suppose $\sup_\ell \|\mathbf{L}^{(\ell)}\|_{\mathrm{op}} \le K$, set $I := \{j, \dots, j+m\}$, and let $\mathbf{L}_I$ be the corresponding principal $(m+1)\times(m+1)$ block. Put $p = P^{(\ell)}_{j,j+m}$. Every lower-triangular entry of $\mathbf{L}_I$ is a subproduct of the full block product and is therefore at least p. For $\mathbf{x} = (m+1)^{-1/2}\mathbf{1}$,

$$K \ge \|\mathbf{L}_I\|_{\mathrm{op}} \ge \mathbf{x}^\top \mathbf{L}_I \mathbf{x} \ge \frac{m+2}{2}\, p.$$

Choose m so that $m+2 > 2K$. Then (a) holds with $\rho = 2K/(m+2) < 1$, completing the equivalence. □

Proposition 11 (Principal inverse bounds for causal masks)**.** *Let $\mathbf{L}_S$ denote the principal submatrix of $\mathbf{L}$ indexed by $\varnothing \ne S \subset \{1, \dots, \ell\}$, with the inherited order. For every semiseparable mask equation 2 with $0 \le a_t \le 1$,*

$$\|\mathbf{L}_S^{-1}\|_{\mathrm{op}} \le 2. \tag{59}$$

This includes the causal linear mask by taking $a_t = 1$. More sharply, the decay mask with $0 \le \gamma < 1$ satisfies $\|\mathbf{L}_S^{-1}\|_{\mathrm{op}} \le 1+\gamma$, and the pointwise-stable semiseparable mask with $a_t \le a < 1$ satisfies $\|\mathbf{L}_S^{-1}\|_{\mathrm{op}} \le 1+a$.

Proof. Write $S = \{i_1 < \cdots < i_m\}$. For a general semiseparable mask, $\mathbf{L}_S^{-1}$ is lower bidiagonal with diagonal entries one and subdiagonal entries

$$-b_p, \qquad b_p := \prod_{t=i_{p-1}+1}^{i_p} a_t, \qquad p = 2, \dots, m.$$

The strictly subdiagonal matrix with weights b_p is a weighted shift whose operator norm is at most one (and is zero when $m = 1$). The triangle inequality proves equation 59. For the decay mask, $b_p = \gamma^{i_p - i_{p-1}} \le \gamma$; under pointwise stability, $b_p \le a$. These give the two sharper bounds. □

Remark 7 (Decay, bounded memory, and nondegeneracy). Proposition 10 is an equivalence for bounded mask gain. The uniform one-step condition $a_t^{(\ell)} \le a < 1$ is only a convenient sufficient condition: pointwise strict inequalities are not enough. For example, if $a_t = (t-1)/t$, then $P_{j,i} = j/i \to 0$ for each fixed j, but the row-memory sums grow linearly. Conversely, $a_{2k} = 1/2$ and $a_{2k-1} = 1$ satisfy uniform two-step contraction even though $\sup_t a_t = 1$.

Bounded mask gain controls the nearly linear rank rates through part (ii) of Theorem 2. Fixed-rank empirical nondegeneracy instead follows from the uniform principal-inverse condition alone by Theorem 3. Proposition 11 continues to hold when $a_t \equiv 1$, even though the corresponding mask gain grows linearly, so the higher-rank nondegeneracy theorem also covers the causal linear mask. The eigenvalue results likewise depend only on the unit diagonal, while the i.i.d. Gaussian results in Section 2.2 have near-linear rank growth even for the undiscounted causal mask. Thus decay controls the strong low-rank rates, but is not required for absolute spectral nondegeneracy.

C.3.2 Rank-One Refinements

The rank-one model permits two refinements beyond the higher-rank results: a sharper conditional singular-value count with an overlap-dependent threshold, and explicit constants in the general mask-gain bounds.

Roadmap The inverse bound converts a positive fraction of Haar coordinates into a uniform singular-value lower bound. A principal-submatrix argument then gives the empirical-law statement, while the causal summation block supplies the operator, nuclear, and stable-rank estimates for the sharper mask-gain bounds.

Theorem 10 (Rank-one empirical-law and mask-gain rank bounds)**.** *Assume that $r_Q = r_K = 1$ and that Assumptions 1 and 2 hold, and let $\mathbf{L} = \mathbf{L}_\ell$ be a unit-diagonal causal mask.*

(i) Empirical-law nondegeneracy. As $(\ell, d) \to (\infty, \infty)$, define

$$\mu_{\mathrm{sing};\ell,d} := \frac{1}{\ell}\sum_{j=1}^{\ell} \delta_{s_j(\mathbf{M})},$$
$$\alpha_{\ell,d} := \frac{\sigma^Q_{\ell,d}\sigma^K_{\ell,d}}{\ell\sqrt{d}},$$
$$\alpha_- := \liminf_{(\ell,d)\to(\infty,\infty)} \alpha_{\ell,d} > 0.$$

Suppose that there is a constant C_{inv} such that every nonempty principal submatrix $\mathbf{L}_S$ satisfies $\|\mathbf{L}_S^{-1}\|_{\mathrm{op}} \le C_{\mathrm{inv}}$. One may take $C_{\mathrm{inv}} = 2$ for the causal linear and nonexpansive semiseparable masks; see Proposition 11. Let

$$Z_{\ell,d} := \sqrt{d}\,\langle \mathbf{v}^Q_{\ell,d}, \mathbf{v}^K_{\ell,d}\rangle,$$

so that $Z_{\ell,d} \Rightarrow G \sim N(0,1)$. For every $t > 0$, write $p_t := \mathbb{P}\{|N(0,1)| > t\} = 2\overline{\Phi}(t)$. Then, for every $0 < \eta < p_t$,

$$\mathbb{P}\left\{\mu_{\mathrm{sing};\ell,d}\left(\left(\frac{|\alpha_{\ell,d}Z_{\ell,d}|t^2}{2C_{\mathrm{inv}}}, \infty\right)\right) \ge p_t - \eta\right\} \longrightarrow 1. \tag{60}$$

Moreover, define $\underline{\alpha} := \min\{1, \alpha_-/2\} > 0$, with the usual extended-real convention if $\alpha_- = \infty$, and, for every $g_0 > 0$, set $\varepsilon_{t,g_0} := \underline{\alpha} g_0 t^2/(2C_{\mathrm{inv}})$. Conditional on $|Z_{\ell,d}| \ge g_0$, $\mu_{\mathrm{sing};\ell,d}((\varepsilon_{t,g_0}, \infty)) \ge p_t - \eta$ with probability tending to one. Unconditionally,

$$\liminf_{(\ell,d)\to(\infty,\infty)} \mathbb{P}\left\{\mu_{\mathrm{sing};\ell,d}((\varepsilon_{t,g_0}, \infty)) \ge p_t - \eta\right\} \ge \mathbb{P}\{|G| \ge g_0\} > 0. \tag{61}$$

In particular, $\mu_{\mathrm{sing};\ell,d}$ does not converge weakly to δ_0 in probability.

(ii) Mask-gain rank bounds at fixed model width. Fix any $d \ge 1$, let $\ell \to \infty$, and set $\kappa_\ell := \|\mathbf{L}\|_{\mathrm{op}}$. Then

$$\mathrm{srank}(\mathbf{M}),\ \mathrm{erank}_2(\mathbf{M}) \ge \left(\frac{3}{4} + o_{\mathbb{P}}(1)\right)\frac{\ell}{\kappa_\ell^2(\log \ell)^2}, \tag{62}$$

$$\mathrm{erank}_1(\mathbf{M}) \ge \left(\frac{1}{2} + o_{\mathbb{P}}(1)\right)\frac{\ell}{\kappa_\ell \log \ell}. \tag{63}$$

Consequently, $\mathrm{srank}(\mathbf{M})$ and $\mathrm{erank}_2(\mathbf{M})$ diverge whenever $\kappa_\ell = o(\sqrt{\ell}/\log \ell)$, while $\mathrm{erank}_1(\mathbf{M})$ diverges whenever $\kappa_\ell = o(\ell/\log \ell)$.

Proof of Theorem 10. Use the normalized rank-one reduction equation 53, with

$$w_i = \sqrt{\ell}\,|u_i|, \qquad \mathbf{T}_\ell = \mathrm{diag}(\mathbf{w})\mathbf{L}\,\mathrm{diag}(\mathbf{w}).$$

The rank proxies of $\mathbf{M}$ and $\mathbf{T}_\ell$ agree almost surely.

Part (i)*: empirical-law nondegeneracy.* Fix $t > 0$ and set $S_t = \{i : w_i > t\}$. Lemma 6, including its threshold-indicator conclusion, gives $|S_t|/\ell \to p_t$ because the Gaussian boundary has zero mass. With probability tending to one S_t is nonempty, and its principal submatrix satisfies

$$s_{\min}\{(\mathbf{T}_\ell)_{S_t}\} \ge \left(\min_{i\in S_t} w_i^2\right) s_{\min}(\mathbf{L}_{S_t}) > \frac{t^2}{C_{\mathrm{inv}}}.$$

By singular-value interlacing for submatrices, $\mathbf{T}_\ell$ has at least $|S_t|$ singular values larger than t^2/C_{inv}. Equation equation 53 therefore implies equation 60.

For the deterministic-threshold consequence, fix $g_0 > 0$. For all sufficiently large $\min\{\ell, d\}$, $\alpha_{\ell,d} \geq \underline{\alpha}$. Hence, on $\{|Z_{\ell,d}| \geq g_0\}$, each of the $|S_t|$ singular values just found is larger than

$$\frac{\underline{\alpha} g_0 t^2}{C_{\mathrm{inv}}} = 2\varepsilon_{t,g_0} > \varepsilon_{t,g_0}.$$

The token-side vector $\mathbf{u}$ is independent of $Z_{\ell,d}$, so conditioning on $\{|Z_{\ell,d}| \geq g_0\}$ does not change the convergence $|S_t|/\ell \to p_t$ in probability. This proves the conditional assertion. Finally, $\mathbb{P}\{|Z_{\ell,d}| \geq g_0\} \to \mathbb{P}\{|G| \geq g_0\}$, and independence gives equation 61.

Part (ii)*: mask-gain bounds.* First,

$$\|\mathbf{T}_\ell\|_{\mathrm{op}} \leq \kappa_\ell \max_i w_i^2. \tag{64}$$

The one-dimensional case of Lemma 8 gives

$$\frac{\max_i w_i^2}{2\log \ell} \longrightarrow 1$$

in probability. Since $L_{ii} = 1$,

$$\|\mathbf{T}_\ell\|_F^2 \geq \sum_{i=1}^{\ell} w_i^4, \qquad \|\mathbf{T}_\ell\|_* \geq |\operatorname{Tr} \mathbf{T}_\ell| = \sum_{i=1}^{\ell} w_i^2.$$

Lemma 6 gives $\ell^{-1} \sum_i w_i^4 \to \mathbb{E}N(0,1)^4 = 3$, while $\sum_i w_i^2 = \ell$ exactly. Therefore

$$\operatorname{srank}(\mathbf{T}_\ell) \geq \frac{\sum_i w_i^4}{\kappa_\ell^2 (\max_i w_i^2)^2}, \tag{65}$$

$$\operatorname{erank}_1(\mathbf{T}_\ell) \geq \frac{\|\mathbf{T}_\ell\|_*}{\|\mathbf{T}_\ell\|_{\mathrm{op}}} \geq \frac{\sum_i w_i^2}{\kappa_\ell \max_i w_i^2}, \tag{66}$$

$$\operatorname{erank}_2(\mathbf{T}_\ell) \geq \frac{\|\mathbf{T}_\ell\|_F^2}{\|\mathbf{T}_\ell\|_{\mathrm{op}}^2} = \operatorname{srank}(\mathbf{T}_\ell). \tag{67}$$

Here both entropy inequalities follow because Shannon entropy dominates min-entropy, first for the singular-value weights and then for their squares. Substituting the Haar asymptotics proves equation 62 and equation 63, including the stated divergence criteria. □

Corollary 1 (Semiseparable memory-dependent rank bounds)**.** *Assume $r_Q = r_K = 1$ and Assumptions 1 and 2 hold. Let $d \geq 1$ be fixed and let $\ell \to \infty$. Let $\mathbf{L}^{(\ell)}$ be a nonexpansive semiseparable mask, and let R_ℓ, C_ℓ be its row- and column-memory sums from Proposition 10. Then*

$$\kappa_\ell := \|\mathbf{L}^{(\ell)}\|_{\mathrm{op}} \leq \sqrt{R_\ell C_\ell}, \tag{68}$$

and

$$\operatorname{srank}(\mathbf{M}),\ \operatorname{erank}_2(\mathbf{M}) \geq \left(\frac{3}{4} + o_{\mathbb{P}}(1)\right) \frac{\ell}{R_\ell C_\ell (\log \ell)^2}, \tag{69}$$

$$\operatorname{erank}_1(\mathbf{M}) \geq \left(\frac{1}{2} + o_{\mathbb{P}}(1)\right) \frac{\ell}{\sqrt{R_\ell C_\ell}\log \ell}. \tag{70}$$

In particular, under Assumption 3, equivalently the uniform block-contraction condition equation 56,

$$\operatorname{srank}(\mathbf{M}),\ \operatorname{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{(\log \ell)^2}\right), \qquad \operatorname{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\log \ell}\right). \tag{71}$$

The Toeplitz-decay mask is the special case $m = 1$ and $\rho = \gamma$.

Proof. Equation equation 68 is Schur's test. Substitution into the mask-gain bounds in part (ii) of Theorem 10 gives the first two estimates. Under uniform block contraction, Proposition 10 gives $R_\ell \vee C_\ell \leq m/(1-\rho)$ uniformly, which yields the stated rates. □

Proposition 12 (Rank-one causal-Volterra spread)**.** *For the causal linear mask $\mathbf{L}_{ij} = \mathbf{1}_{\{i \geq j\}}$, write $\mathbf{w} = \sqrt{\ell}(|u(1)|, \ldots, |u(\ell)|)$, where $\mathbf{u} = (u(1), \ldots, u(\ell))$ is uniformly distributed on the unit sphere in $\mathbb{R}^\ell$, and put $\mathbf{T}_\ell = \operatorname{diag}(\mathbf{w})\mathbf{L}\operatorname{diag}(\mathbf{w})$. There exist deterministic constants $c_V, C_V > 0$, independent of ℓ, such that, with probability tending to one,*

$$c_V \ell \leq \|\mathbf{T}_\ell\|_{\mathrm{op}} \leq C_V \ell, \qquad \|\mathbf{T}_\ell\|_* \geq c_V \ell \log \ell. \tag{72}$$

In addition,

$$\frac{\|\mathbf{T}_\ell\|_F^2}{\ell^2} \xrightarrow{\mathbb{P}} \frac{1}{2}.$$

Consequently, under Assumptions 1 and 2, the corresponding rank-one matrix $\mathbf{M}$ in Proposition 9 satisfies

$$\operatorname{srank}(\mathbf{M}) = \Theta_{\mathbb{P}}(1), \qquad \operatorname{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}(\log \ell), \qquad \operatorname{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}(1). \tag{73}$$

Proof. Since $\sum_i w_i^2 = \ell$ exactly and Lemma 6 gives $\ell^{-1} \sum_i w_i^4 \to 3$ in probability,

$$\|\mathbf{T}_\ell\|_F^2 = \sum_{i \geq j} w_i^2 w_j^2 = \frac{1}{2}\ell^2 + \frac{1}{2}\sum_i w_i^4 = \left(\frac{1}{2} + o_{\mathbb{P}}(1)\right)\ell^2.$$

Consequently, $\|\mathbf{T}_\ell\|_{\mathrm{op}} \leq \|\mathbf{T}_\ell\|_F \leq C_V \ell$ with probability tending to one. For the matching lower bound, test $\mathbf{T}_\ell$ on the normalized all-one vector. The empirical Haar-row law gives

$$\sum_{i \geq j} w_i w_j = \frac{1}{2}\left\{\left(\sum_i w_i\right)^2 + \sum_i w_i^2\right\},$$

and hence

$$\frac{1}{\ell^2}\sum_{i \geq j} w_i w_j = \frac{1}{2}\left\{\left(\frac{1}{\ell}\sum_i w_i\right)^2 + \frac{1}{\ell^2}\sum_i w_i^2\right\} \longrightarrow \frac{1}{2}\{\mathbb{E}|N(0,1)|\}^2 = \frac{1}{\pi}$$

in probability, and hence $\|\mathbf{T}_\ell\|_{\mathrm{op}} \geq c_V \ell$ with probability tending to one.

It remains to prove the nuclear-norm lower bound. Fix $t = 1$, let $S_t = \{i : w_i > t\}$, and put $m = |S_t|$. Then

$$\frac{m}{\ell} \longrightarrow p_t := \mathbb{P}\{|N(0,1)| > t\} > 0$$

in probability. With the indices in inherited order,

$$(\mathbf{T}_\ell)_{S_t} = \mathbf{D}_{S_t}\mathbf{J}_m\mathbf{D}_{S_t}, \qquad \mathbf{D}_{S_t} := \operatorname{diag}((w_i)_{i \in S_t}),$$

on the event $m \geq 1$, whose probability tends to one. Every diagonal entry of $\mathbf{D}_{S_t}$ exceeds t. Compression is contractive for the nuclear norm. Since $\mathbf{J}_m = \mathbf{D}_{S_t}^{-1}(\mathbf{D}_{S_t}\mathbf{J}_m\mathbf{D}_{S_t})\mathbf{D}_{S_t}^{-1}$ and $\|\mathbf{D}_{S_t}^{-1}\|_{\mathrm{op}} \leq t^{-1}$, the ideal property gives $\|\mathbf{J}_m\|_* \leq t^{-2}\|(\mathbf{T}_\ell)_{S_t}\|_*$; equivalently, in the lower-bound direction,

$$\|\mathbf{T}_\ell\|_* \geq \|(\mathbf{T}_\ell)_{S_t}\|_* \geq t^2\|\mathbf{J}_m\|_*.$$

Lemma 9 and $m/\ell \to p_t$ now give $\|\mathbf{T}_\ell\|_* \geq c_V \ell \log \ell$ with probability tending to one.

The Frobenius and operator estimates imply $\operatorname{srank}(\mathbf{T}_\ell) = \Theta_{\mathbb{P}}(1)$. By Proposition 13, $\operatorname{erank}_2(\mathbf{T}_\ell) \geq \operatorname{srank}(\mathbf{T}_\ell)$. Shannon entropy also gives the following nuclear-norm bound:

$$\operatorname{erank}_1(\mathbf{T}_\ell) \geq \frac{\|\mathbf{T}_\ell\|_*}{\|\mathbf{T}_\ell\|_{\mathrm{op}}} = \Omega_{\mathbb{P}}(\log \ell).$$

Thus

$$\operatorname{erank}_2(\mathbf{T}_\ell) = \Omega_{\mathbb{P}}(1).$$

Equation equation 53 transfers these scale-invariant conclusions to $\mathbf{M}$. □

C.3.3 Extension to Arbitrary Fixed Rank

The rank-one argument isolates the mask geometry. At arbitrary fixed rank, the scalar coefficient is replaced by a finite random coefficient matrix. The principal new issue is to show that its symmetric part retains a nonvanishing fraction of the coefficient mass; the independent Haar right spaces provide exactly this non-cancellation.

Roadmap Normalize away the common singular-value scale, express diagonal entries as fixed-dimensional Gaussian quadratic forms, and prove non-cancellation of the symmetric coefficient part from Haar right-space randomness. The resulting quadratic-form law gives empirical nondegeneracy, while the mask gain and principal-inverse estimates transfer the rank bounds to arbitrary fixed rank.

Proof of Theorem 2. Normalized coefficient form. Set

$$s_{Q,\ell} := \left(\sum_{a=1}^{r_Q} (\sigma_{a,\ell}^Q)^2\right)^{1/2}, \qquad s_{K,\ell} := \left(\sum_{b=1}^{r_K} (\sigma_{b,\ell}^K)^2\right)^{1/2},$$

$$p_{a,\ell} := \frac{\sigma_{a,\ell}^Q}{s_{Q,\ell}}, \qquad q_{b,\ell} := \frac{\sigma_{b,\ell}^K}{s_{K,\ell}}, \qquad \gamma_\ell := \frac{s_{Q,\ell} s_{K,\ell}}{\ell\sqrt{d}}.$$

The scalar γ_ℓ may vary arbitrarily: by Proposition 16, it multiplies every term below and cancels from all three rank proxies. Let $r := \max(r_Q, r_K)$ and let $\mathbf{z}_i := \sqrt{\ell}\,(u_1(i), \dots, u_r(i)) \in \mathbb{R}^r$ be the scaled token rows. Define the zero-padded coefficient matrix $\mathbf{F}_\ell \in \mathbb{R}^{r\times r}$ by

$$F_{ab}^{(\ell)} := \gamma_\ell p_{a,\ell} q_{b,\ell} Z_{ab;\ell,d}, \qquad Z_{ab;\ell,d} := \sqrt{d}\,\langle \mathbf{v}_{a,\ell}^Q, \mathbf{v}_{b,\ell}^K \rangle,$$

on the $r_Q \times r_K$ block and by zero elsewhere, and put

$$\mathbf{B}_\ell := \frac{\mathbf{F}_\ell + \mathbf{F}_\ell^\top}{2}.$$

The common-frame representation gives the exact formula

$$M_{ij} = L_{ij}\, \mathbf{z}_i^\top \mathbf{F}_\ell \mathbf{z}_j, \qquad M_{ii} = \mathbf{z}_i^\top \mathbf{B}_\ell \mathbf{z}_i. \tag{74}$$

Moreover, $\mathbf{F}_\ell$ and $\mathbf{B}_\ell$ are independent of the token frame.

Right-space non-cancellation. We first record the non-cancellation supplied by the random right spaces. Set

$$\rho_\ell := \|\mathbf{B}_\ell\|_{\mathrm{op}}, \qquad \chi_\ell := \frac{\rho_\ell}{\|\mathbf{F}_\ell\|_*},$$

with $\chi_\ell = 0$ when $\mathbf{F}_\ell = 0$. Then

$$\chi_\ell = \Omega_{\mathbb{P}}(1). \tag{75}$$

To prove this, remove the common scalar and define

$$\mathbf{C}_\ell := \gamma_\ell^{-1} \mathbf{F}_\ell, \qquad \mathbf{D}_\ell := \frac{\mathbf{C}_\ell + \mathbf{C}_\ell^\top}{2}.$$

Here $\gamma_\ell > 0$ because $\mathbf{Q}, \mathbf{K} \neq 0$; individual singular values may vanish. View $\mathbf{p}_\ell = (p_{a,\ell})_a$ and $\mathbf{q}_\ell = (q_{b,\ell})_b$ as zero-padded vectors in $\mathbb{R}^r$. They lie on fixed-dimensional unit spheres, so from every subsequence one may extract a further subsequence on which $\mathbf{p}_\ell \to \mathbf{p}$ and $\mathbf{q}_\ell \to \mathbf{q}$. Along that subsequence, the fixed-d Haar law gives

$$(\mathbf{C}_\ell, \mathbf{D}_\ell) \Rightarrow (\mathbf{C}_d(\mathbf{H}), \mathbf{D}_d(\mathbf{H})),$$

$$C_{d,ab}(\mathbf{H}) = p_a q_b \sqrt{d}\, H_{ab},$$

$$\mathbf{D}_d(\mathbf{H}) = \frac{\mathbf{C}_d(\mathbf{H}) + \mathbf{C}_d(\mathbf{H})^\top}{2},$$

with zero padding outside the $r_Q \times r_K$ block and $\mathbf{H} = \mathbf{V}_Q^\top \mathbf{V}_K$ for independent Haar Stiefel frames in $\mathbb{R}^d$. Lemma 4 gives $\mathbb{P}\{H_{ab} = 0\} = 0$ for every entry. If the supports of $\mathbf{p}$ and $\mathbf{q}$ intersect, choose

a with $p_a q_a > 0$; then $D_{d,aa} = p_a q_a \sqrt{d}\, H_{aa} \neq 0$ almost surely. If the supports are disjoint, choose a, b with $p_a q_b > 0$; then $a \neq b$, $p_b q_a = 0$, and $D_{d,ab} = p_a q_b \sqrt{d}\, H_{ab}/2 \neq 0$ almost surely. Thus both $\mathbf{C}_d(\mathbf{H})$ and $\mathbf{D}_d(\mathbf{H})$ are nonzero almost surely, and hence

$$\frac{\|\mathbf{D}_\ell\|_{\mathrm{op}}}{\|\mathbf{C}_\ell\|_*} \Rightarrow \frac{\|\mathbf{D}_d(\mathbf{H})\|_{\mathrm{op}}}{\|\mathbf{C}_d(\mathbf{H})\|_*} > 0 \quad \text{almost surely}$$

along every such coefficient-convergent subsequence. Compactness makes this lower bound uniform over the normalized singular-value profiles. Indeed, if uniformity failed, there would be a fixed $\eta_0 > 0$, profiles $(\mathbf{p}_n, \mathbf{q}_n)$, and $t_n \downarrow 0$ for which the corresponding ratios χ_n satisfy $\mathbb{P}\{\chi_n \leq t_n\} \geq \eta_0$. Extract a coefficient-convergent subsequence. The preceding argument gives $\chi_n \Rightarrow \chi_*$ with $\mathbb{P}\{\chi_* > 0\} = 1$. For each $a > 0$, Portmanteau's inequality then yields

$$\eta_0 \leq \limsup_n \mathbb{P}\{\chi_n \leq t_n\} \leq \limsup_n \mathbb{P}\{\chi_n \leq a\} \leq \mathbb{P}\{\chi_* \leq a\}.$$

Letting $a \downarrow 0$ gives a contradiction. Hence, for every $\eta > 0$, there is a $c_\eta > 0$, depending only on η, r_Q, r_K, d, such that $\mathbb{P}\{\chi_\ell \geq c_\eta\} \geq 1 - \eta$ uniformly in the profiles. This proves equation 75.

General mask-gain bounds. Let $\mathbf{X}_\ell := \sqrt{\ell}\,\mathbf{U}$ so that its rows are $\mathbf{z}_i^\top$. If $\mathbf{F}_\ell = \sum_k s_k \mathbf{a}_k \mathbf{b}_k^\top$ is a singular-value decomposition, then equation 74 gives

$$\mathbf{M} = \sum_k s_k \,\mathrm{diag}(\mathbf{X}_\ell \mathbf{a}_k)\, \mathbf{L}\, \mathrm{diag}(\mathbf{X}_\ell \mathbf{b}_k).$$

Therefore

$$\frac{\|\mathbf{M}\|_{\mathrm{op}}}{\kappa_\ell \|\mathbf{F}_\ell\|_* \log \ell} = O_{\mathbb{P}}(1), \tag{76}$$

because $\|\mathbf{M}\|_{\mathrm{op}} \leq \kappa_\ell \|\mathbf{F}_\ell\|_* \max_i \|\mathbf{z}_i\|_2^2$ and Lemma 8 gives $\max_i \|\mathbf{z}_i\|_2^2 = O_{\mathbb{P}}(\log \ell)$.

Because $\mathbf{L}$ is unit diagonal, the second formula in equation 74, nuclear-norm duality, and the uniform form of Lemma 6 give

$$\frac{\|\mathbf{M}\|_F^2}{\ell \rho_\ell^2} \geq \frac{1}{\ell \rho_\ell^2} \sum_{i=1}^{\ell} (\mathbf{z}_i^\top \mathbf{B}_\ell \mathbf{z}_i)^2 = \Omega_{\mathbb{P}}(1), \tag{77}$$

$$\frac{\|\mathbf{M}\|_*}{\ell \rho_\ell} \geq \frac{1}{\ell \rho_\ell} \sum_{i=1}^{\ell} |\mathbf{z}_i^\top \mathbf{B}_\ell \mathbf{z}_i| = \Omega_{\mathbb{P}}(1). \tag{78}$$

Lemma 7 supplies the two uniform Gaussian quadratic-form lower bounds used here. The required uniform Haar-row law follows from compactness of the symmetric operator-norm unit sphere. For the squared quadratic forms it follows by a finite expansion into degree-four coordinate monomials and Lemma 6; for the absolute quadratic forms, use

$$|\mathbf{z}^\top \mathbf{C} \mathbf{z} - \mathbf{z}^\top \mathbf{D} \mathbf{z}| \leq \|\mathbf{C} - \mathbf{D}\|_{\mathrm{op}} \|\mathbf{z}\|_2^2$$

together with a finite net and $\ell^{-1} \sum_i \|\mathbf{z}_i\|_2^2 = r$. Also, $\rho_\ell > 0$ almost surely: $p_{1,\ell} q_{1,\ell} > 0$ and $(B_\ell)_{11} = \gamma_\ell p_{1,\ell} q_{1,\ell} Z_{11;\ell,d}$, where $\mathbb{P}\{Z_{11;\ell,d} = 0\} = 0$. Apply these inequalities conditionally to $\mathbf{B}_\ell / \rho_\ell$. The nuclear bound uses the diagonal sign matrix matching the signs of M_{ii} as a dual certificate.

Combining equation 75– equation 78, and using Shannon entropy's min-entropy lower bound, yields

$$\mathrm{srank}(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\kappa_\ell^2 (\log \ell)^2}\right), \qquad \mathrm{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{\ell}{\kappa_\ell \log \ell}\right),$$

and Proposition 13 gives $\mathrm{erank}_2(\mathbf{M}) \geq \mathrm{srank}(\mathbf{M})$. This proves part (i). Part (ii) follows immediately from Assumption 3; Proposition 10 supplies the semiseparable characterization of this assumption.

The causal linear mask. It remains to treat the causal linear mask. Fix $\eta > 0$. By equation 75, there is a deterministic $c_0 \in (0, 1]$ such that

$$\mathbb{P}\{\rho_\ell \geq c_0 \|\mathbf{F}_\ell\|_*\} \geq 1 - \eta \tag{79}$$

for all sufficiently large ℓ. Let $\mathcal{N} \subset S^{r-1}$ be a fixed finite $1/4$-net. For every symmetric matrix $\mathbf{C}$, some $\mathbf{v} \in \mathcal{N}$ satisfies

$$|\mathbf{v}^\top \mathbf{C}\mathbf{v}| \geq \frac{1}{2}\|\mathbf{C}\|_{\mathrm{op}}.$$

Set $\delta := c_0/32$. For each $\mathbf{v} \in \mathcal{N}$, let

$$S_{\mathbf{v}} := \{i : \|\mathbf{z}_i - \mathbf{v}\|_2 \leq \delta\}.$$

Since $\mathcal{N}$ is finite, Lemma 6 gives, simultaneously over $\mathbf{v} \in \mathcal{N}$,

$$\frac{|S_{\mathbf{v}}|}{\ell} \longrightarrow q_\delta := \mathbb{P}\{\|\boldsymbol{\xi} - \mathbf{e}_1\|_2 \leq \delta\} > 0$$

in probability; rotational invariance makes q_δ independent of the choice of the unit vector $\mathbf{v}$.

On the event in equation 79, choose $\mathbf{v} \in \mathcal{N}$ such that, with $c := \mathbf{v}^\top \mathbf{F}_\ell \mathbf{v} = \mathbf{v}^\top \mathbf{B}_\ell \mathbf{v}$,

$$|c| \geq \frac{1}{2}\rho_\ell \geq \frac{c_0}{2}\|\mathbf{F}_\ell\|_*.$$

Write $S_{\mathbf{v}} = \{i_1 < \cdots < i_m\}$ and $\mathbf{z}_{i_k} = \mathbf{v} + \mathbf{e}_k$, so $\|\mathbf{e}_k\|_2 \leq \delta$. For the causal linear mask, the corresponding principal block satisfies

$$\mathbf{M}_{S_{\mathbf{v}}} = c\mathbf{J}_m + \mathbf{E}_m, \qquad (\mathbf{J}_m)_{kl} = \mathbf{1}_{\{k \geq l\}}.$$

Expanding $(\mathbf{v} + \mathbf{e}_k)^\top \mathbf{F}_\ell (\mathbf{v} + \mathbf{e}_l)$ gives an explicit form for the perturbation. Let $\mathbf{R} \in \mathbb{R}^{m \times r}$ have kth row $\mathbf{e}_k^\top$, and write $\mathbf{F}_\ell = \sum_t s_t \mathbf{a}_t \mathbf{b}_t^\top$. Then

$$\mathbf{E}_m = \operatorname{diag}(\mathbf{R}\mathbf{F}_\ell \mathbf{v})\mathbf{J}_m + \mathbf{J}_m \operatorname{diag}(\mathbf{R}\mathbf{F}_\ell^\top \mathbf{v}) + \sum_t s_t \operatorname{diag}(\mathbf{R}\mathbf{a}_t)\mathbf{J}_m \operatorname{diag}(\mathbf{R}\mathbf{b}_t).$$

The ideal property of the operator and nuclear norms therefore gives, for $\nu \in \{\mathrm{op}, *\}$,

$$\|\mathbf{E}_m\|_\nu \leq \left(2\delta\|\mathbf{F}_\ell\|_{\mathrm{op}} + \delta^2\|\mathbf{F}_\ell\|_*\right)\|\mathbf{J}_m\|_\nu \leq \frac{|c|}{2}\|\mathbf{J}_m\|_\nu. \tag{80}$$

Compression is contractive for both norms. Also Lemma 9 gives the required operator- and nuclear-norm lower bounds for $\mathbf{J}_m$. The reverse triangle inequality and equation 80 first give

$$\|\mathbf{M}\|_\nu \geq \|\mathbf{M}_{S_{\mathbf{v}}}\|_\nu \geq \frac{|c|}{2}\|\mathbf{J}_m\|_\nu, \qquad \nu \in \{\mathrm{op}, *\}.$$

Consequently, for every $\varepsilon > 0$ there is a constant $c_\varepsilon > 0$ such that, for all sufficiently large ℓ, both normalized norms below exceed c_ε with probability at least $1 - \varepsilon$. Equivalently,

$$\frac{\|\mathbf{M}\|_{\mathrm{op}}}{\ell\|\mathbf{F}_\ell\|_*} = \Omega_{\mathbb{P}}(1), \qquad \frac{\|\mathbf{M}\|_*}{\ell \log \ell\,\|\mathbf{F}_\ell\|_*} = \Omega_{\mathbb{P}}(1). \tag{81}$$

On the other hand, $\mathbf{M}$ is the lower-triangular truncation of $\mathbf{X}_\ell \mathbf{F}_\ell \mathbf{X}_\ell^\top$ for the causal all-one mask, so truncation can only decrease the Frobenius norm. Since $\mathbf{X}_\ell^\top \mathbf{X}_\ell = \ell \mathbf{I}_r$,

$$\|\mathbf{M}\|_F \leq \|\mathbf{X}_\ell \mathbf{F}_\ell \mathbf{X}_\ell^\top\|_F = \ell\|\mathbf{F}_\ell\|_F \leq \ell\|\mathbf{F}_\ell\|_*,$$

and hence $\|\mathbf{M}\|_{\mathrm{op}} \leq \ell\|\mathbf{F}_\ell\|_*$. The operator lower bound and the Frobenius upper bound give $\operatorname{srank}(\mathbf{M}) = O_{\mathbb{P}}(1)$, while $\operatorname{srank}(\mathbf{M}) \geq 1$ whenever $\mathbf{M} \neq 0$; the operator lower bound shows that this event has probability tending to one. Thus $\operatorname{srank}(\mathbf{M}) = \Theta_{\mathbb{P}}(1)$. The nuclear lower bound and operator upper bound give $\operatorname{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}(\log \ell)$. By Proposition 13, $\operatorname{erank}_2(\mathbf{M}) \geq \operatorname{srank}(\mathbf{M}) = \Omega_{\mathbb{P}}(1)$. This proves part (iii). □

Proof of Theorem 3. Use the normalized coefficient form from the preceding proof and set

$$\mathbf{C}_\ell := \gamma_\ell^{-1}\mathbf{F}_\ell, \qquad X_\ell := \|\mathbf{C}_\ell\|_F^2.$$

Thus $(C_\ell)_{ab} = p_{a,\ell} q_{b,\ell} Z_{ab;\ell,d}$ on the $r_Q \times r_K$ block. Since the normalized singular-value vectors have unit Euclidean norm and $\mathbb{E}Z_{ab;\ell,d}^2 = 1$,

$$\mathbb{E}X_\ell = \sum_{a,b} p_{a,\ell}^2 q_{b,\ell}^2 = 1.$$

Also $|Z_{ab;\ell,d}| \le \sqrt{d}$, so $0 \le X_\ell \le d$ and $\mathbb{E}X_\ell^2 \le d$. Paley–Zygmund therefore gives

$$\mathbb{P}\left\{\|\mathbf{C}_\ell\|_* \ge \frac{1}{\sqrt{2}}\right\} \ge \mathbb{P}\left\{X_\ell \ge \frac{1}{2}\right\} \ge \frac{1}{4d} =: p_C. \tag{82}$$

Recall $\rho_\ell = \|\mathbf{B}_\ell\|_{\mathrm{op}}$ and $\chi_\ell = \rho_\ell / \|\mathbf{F}_\ell\|_*$. By equation 75, there is a constant $a_H \in (0,1]$ such that, for all sufficiently large ℓ,

$$\mathbb{P}\{\chi_\ell \ge a_H\} \ge 1 - \frac{p_C}{2}.$$

Consequently, the coefficient event

$$\mathcal{E}_\ell := \left\{\chi_\ell \ge a_H, \quad \|\mathbf{C}_\ell\|_* \ge \frac{1}{\sqrt{2}}\right\}$$

satisfies

$$\liminf_{\ell\to\infty} \mathbb{P}(\mathcal{E}_\ell) \ge \frac{p_C}{2}. \tag{83}$$

Enlarging the inverse-bound constant if necessary, take $C_{\mathrm{inv}} \ge 1$. Let $\mathcal{N} \subset S^{r-1}$ be a fixed finite $1/4$-net and set

$$\beta := \frac{a_H}{2}, \qquad T := \max\{2, C_{\mathrm{inv}}^{4r-1}\}.$$

Choose a deterministic $\delta > 0$ small enough that

$$\delta \le \frac{\beta}{2}, \qquad \delta^2 T(1+\beta^{-1}) \le \frac{\beta}{8C_{\mathrm{inv}}}, \tag{84}$$

and define

$$q_\delta := \mathbb{P}\{\|\boldsymbol{\xi} - \mathbf{e}_1\|_2 \le \delta\} > 0, \qquad \boldsymbol{\xi} \sim N(0, \mathbf{I}_r).$$

For $\mathbf{v} \in \mathcal{N}$, define $S_{\mathbf{v}} := \{i : \|\mathbf{z}_i - \mathbf{v}\|_2 \le \delta\}$. The empirical Haar-row law and rotational invariance give

$$\mathcal{A}_\ell := \left\{\min_{\mathbf{v}\in\mathcal{N}} |S_{\mathbf{v}}| \ge \frac{q_\delta}{2}\ell\right\}, \qquad \mathbb{P}(\mathcal{A}_\ell) \longrightarrow 1. \tag{85}$$

On $\mathcal{E}_\ell \cap \mathcal{A}_\ell$, choose $\mathbf{v} \in \mathcal{N}$ such that

$$|c| := |\mathbf{v}^\top \mathbf{F}_\ell \mathbf{v}| = |\mathbf{v}^\top \mathbf{B}_\ell \mathbf{v}| \ge \frac{1}{2}\rho_\ell \ge \beta \|\mathbf{F}_\ell\|_*.$$

Write $S_{\mathbf{v}} = \{i_1 < \cdots < i_m\}$ and $\mathbf{z}_{i_k} = \mathbf{v} + \mathbf{e}_k$, where $\|\mathbf{e}_k\|_2 \le \delta$. Put $\mathbf{L}_S := (\mathbf{L}_\ell)_{S_{\mathbf{v}}}$, let $\mathbf{E} \in \mathbb{R}^{m\times r}$ have rows $\mathbf{e}_k^\top$, and define

$$g_k := 1 + \frac{\mathbf{e}_k^\top \mathbf{F}_\ell \mathbf{v}}{c}, \qquad h_k := 1 + \frac{\mathbf{v}^\top \mathbf{F}_\ell \mathbf{e}_k}{c}, \qquad \mathbf{R}_\ell := \mathbf{F}_\ell - \frac{\mathbf{F}_\ell \mathbf{v}\mathbf{v}^\top \mathbf{F}_\ell}{c}.$$

Writing $\mathbf{D}_g := \mathrm{diag}(g_1, \ldots, g_m)$ and $\mathbf{D}_h := \mathrm{diag}(h_1, \ldots, h_m)$, direct expansion gives the exact factorization

$$\mathbf{M}_{S_{\mathbf{v}}} = c\mathbf{D}_g \mathbf{L}_S \mathbf{D}_h + \mathbf{L}_S \odot (\mathbf{E}\mathbf{R}_\ell \mathbf{E}^\top).$$

Since $|c| \ge \beta\|\mathbf{F}_\ell\|_*$ and $\|\mathbf{F}_\ell\|_{\mathrm{op}} \le \|\mathbf{F}_\ell\|_*$,

$$\min_k\{|g_k|, |h_k|\} \ge 1 - \frac{\delta}{\beta} \ge \frac{1}{2}.$$

The principal-inverse assumption therefore yields

$$s_{\min}(c\mathbf{D}_g \mathbf{L}_S \mathbf{D}_h) \ge \frac{|c|}{4C_{\mathrm{inv}}} \ge \frac{\beta\|\mathbf{F}_\ell\|_*}{4C_{\mathrm{inv}}}. \tag{86}$$

Moreover,

$$\|\mathbf{R}_\ell\|_* \le \|\mathbf{F}_\ell\|_* + \frac{\|\mathbf{F}_\ell \mathbf{v}\|_2 \|\mathbf{F}_\ell^\top \mathbf{v}\|_2}{|c|} \le (1+\beta^{-1})\|\mathbf{F}_\ell\|_*. \tag{87}$$

Let $s_1(\mathbf{L}_S) \geq \cdots \geq s_m(\mathbf{L}_S)$ and let $k_T := \#\{j : s_j(\mathbf{L}_S) > T\}$. Since $\mathbf{L}_S$ is unit lower triangular, $|\det \mathbf{L}_S| = 1$, while the inverse bound gives $s_j(\mathbf{L}_S) \geq C_{\mathrm{inv}}^{-1}$ for every j. If $C_{\mathrm{inv}} = 1$, these facts force all the singular values to equal one, so $k_T = 0$. If $C_{\mathrm{inv}} > 1$, then

$$1 = \prod_{j=1}^{m} s_j(\mathbf{L}_S) \geq T^{k_T} C_{\mathrm{inv}}^{-(m-k_T)},$$

and hence

$$k_T \leq \frac{m \log C_{\mathrm{inv}}}{\log(C_{\mathrm{inv}} T)} \leq \frac{m}{4r}. \tag{88}$$

The final bound also holds when $C_{\mathrm{inv}} = 1$.

Split the singular-value decomposition of the mask at T as $\mathbf{L}_S = \mathbf{L}_{>T} + \mathbf{L}_{\leq T}$, and write the residual term as

$$\mathbf{H}_m := \mathbf{L}_{>T} \odot (\mathbf{E}\mathbf{R}_\ell \mathbf{E}^\top),$$
$$\mathbf{G}_m := \mathbf{L}_{\leq T} \odot (\mathbf{E}\mathbf{R}_\ell \mathbf{E}^\top).$$

The Hadamard-product rank inequality and $\mathrm{rank}(\mathbf{E}\mathbf{R}_\ell \mathbf{E}^\top) \leq r$ give

$$\mathrm{rank}(\mathbf{H}_m) \leq r k_T \leq \frac{m}{4}. \tag{89}$$

For the low-mask part, take a singular-value decomposition $\mathbf{R}_\ell = \sum_t \tau_t \mathbf{a}_t \mathbf{b}_t^\top$. Since $\mathbf{L} \odot (\mathbf{x}\mathbf{y}^\top) = \mathrm{diag}(\mathbf{x})\mathbf{L}\,\mathrm{diag}(\mathbf{y})$ and every row of $\mathbf{E}$ has norm at most δ,

$$\|\mathbf{G}_m\|_{\mathrm{op}} \leq \delta^2 \|\mathbf{L}_{\leq T}\|_{\mathrm{op}} \sum_t \tau_t \leq \delta^2 T \|\mathbf{R}_\ell\|_*$$
$$\leq \frac{\beta \|\mathbf{F}_\ell\|_*}{8 C_{\mathrm{inv}}}, \tag{90}$$

where the last step uses equation 84 and equation 87. Combining equation 86 and equation 90 shows that

$$s_{\min}(c\mathbf{D}_g \mathbf{L}_S \mathbf{D}_h + \mathbf{G}_m) \geq \frac{\beta \|\mathbf{F}_\ell\|_*}{8 C_{\mathrm{inv}}}.$$

Adding the rank-at-most-$m/4$ matrix $\mathbf{H}_m$ and using the min–max principle leaves at least $3m/4$ singular values of $\mathbf{M}_{S_{\mathrm{v}}}$ no smaller than the same threshold.

Set

$$\underline{\gamma} := \min\left\{1, \frac{\gamma_-}{2}\right\} > 0,$$

with the usual extended-real convention if $\gamma_- = \infty$. For all sufficiently large ℓ, $\gamma_\ell \geq \underline{\gamma}$. Thus, on $\mathcal{E}_\ell \cap \mathcal{A}_\ell$, these singular values are larger than

$$\varepsilon_0 := \frac{\beta \underline{\gamma}}{16\sqrt{2}\, C_{\mathrm{inv}}}.$$

Singular-value interlacing for submatrices and equation 85 then show that at least $3q_\delta \ell/8$ singular values of $\mathbf{M}_\ell$ exceed ε_0. Combining this conclusion with equation 83 and equation 85 proves equation 9 with $c_{\mathrm{mass}} = 3q_\delta/8$ and, for example, $p_{\mathrm{nd}} = p_C/4$. The empirical-law conclusion follows immediately. □

Remark 8 (Why the causal-linear rank rates need the Volterra argument). Part (iii) is not obtained by simply substituting the causal linear mask gain into part (i). Indeed, for $\mathbf{L} = \mathbf{J}_\ell$, Lemma 9 gives

$$\kappa_\ell = \|\mathbf{J}_\ell\|_{\mathrm{op}} = \left\{2 \sin\left(\frac{\pi}{4\ell + 2}\right)\right\}^{-1} \sim \frac{2}{\pi}\ell.$$

Consequently, the general bounds in part (i) specialize only to

$$\mathrm{srank}(\mathbf{M}),\ \mathrm{erank}_2(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{1}{\ell(\log \ell)^2}\right), \qquad \mathrm{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}\left(\frac{1}{\log \ell}\right),$$

which are asymptotically uninformative since every nonzero matrix has these rank proxies at least one. In contrast, part (ii) follows directly from part (i) because its mask families have uniformly bounded κ_ℓ.

The causal rank-rate proof uses finer information than the largest mask singular value. A positive-density principal block is a perturbation of $c\mathbf{J}_m$ with $m \asymp \ell$, and the Volterra profile satisfies

$$\|\mathbf{J}_m\|_{\mathrm{op}} \asymp m, \qquad \|\mathbf{J}_m\|_* \asymp m \log m.$$

The first estimate supplies the operator-norm lower bound needed for stable rank control, while the second supplies the logarithmic nuclear-norm gain that leads to $\mathrm{erank}_1(\mathbf{M}) = \Omega_{\mathbb{P}}(\log \ell)$. Thus the full causal singular-value profile, rather than κ_ℓ alone, is essential for the rates in part (iii).

Remark 9 (Why random right-space non-cancellation is needed). The higher-rank bounds use Assumption 1 through equation 75. Without such a non-cancellation condition the conclusion is false for arbitrary deterministic coefficient matrices. For example, take $r = 2$, let the coefficient matrix be nonzero and skew-symmetric, and take the identity mask, which is the semiseparable mask with all transition coefficients equal to zero. Then $M_{ii} = \mathbf{z}_i^\top \mathbf{F} \mathbf{z}_i = 0$ and every off-diagonal entry is masked out, so $\mathbf{M} = 0$. Independent Haar right spaces rule out this exact skew cancellation almost surely and yield the quantitative bound equation 75.

D Causal Softmax Attention Comparisons

This appendix records the corresponding causal row-softmax comparisons. These operators are not instances of the entrywise masked model equation 1: the softmax denominator couples all accessible entries in a row. Given scores

$$\mathbf{S} = \frac{\mathbf{Q}\mathbf{K}^\top}{\tau_\ell}, \qquad \tau_\ell > 0,$$

the causal self-attention matrix is

$$\mathbf{A} = \mathrm{softmax}_{\mathrm{row}}^{\mathrm{causal}}(\mathbf{S}), \qquad A_{ij} = \frac{\mathbf{1}_{\{j \le i\}} e^{S_{ij}}}{\sum_{m=1}^{i} e^{S_{im}}}. \tag{91}$$

The causal all-one mask selects the support of equation 91 by excluding future tokens from the denominator, but the resulting matrix is a normalized stochastic operator rather than $\mathbf{L} \odot \mathbf{S}$.

D.1 Bounded-Score Non-Collapse and Sink Collapse

The bounded-oscillation model and the trained GPT-2 behavior are consistent with different asymptotic regimes. The next result makes the obstruction precise: bounded within-row logit variation forces logarithmically diverging scale-invariant rank proxies, even when the score matrix has fixed rank. We then give a complementary attention-sink condition under which the energy-sensitive ranks do collapse.

Roadmap Bounded row oscillation makes every accessible softmax weight comparable to the reciprocal row length. Summing the squared weights yields logarithmic Frobenius-energy growth, and a Hardy-type estimate controls the operator norm. For sink collapse, a single column asymptotically dominates the Frobenius energy; a slightly stronger decay rate for the residual also forces erank_2 to converge to one. Collapse of erank_1 requires separate control of the nuclear-norm tail.

Theorem 11 (Bounded causal scores do not exhibit relative rank collapse)**.** *Let*

$$\mathbf{A}_\ell = \mathrm{softmax}_{\mathrm{row}}^{\mathrm{causal}}(\mathbf{S}_\ell), \qquad A_{ij} = \frac{\mathbf{1}_{\{j \le i\}} e^{S_{ij}}}{\sum_{m=1}^{i} e^{S_{im}}}.$$

Assume that the accessible score oscillation is uniformly bounded: for a constant $B < \infty$ independent of ℓ,

$$\max_{1 \le i \le \ell} \left\{ \max_{1 \le j \le i} S_{ij} - \min_{1 \le j \le i} S_{ij} \right\} \le B. \tag{92}$$

In particular, this holds with $B = 2R^2$ under the fixed-rank bounded feature model

$$S_{ij} = \mathbf{x}_i^\top \mathbf{y}_j, \qquad \mathbf{x}_i, \mathbf{y}_j \in \mathbb{R}^r, \qquad \max_i \|\mathbf{x}_i\|_2 \vee \max_j \|\mathbf{y}_j\|_2 \le R,$$

where r and R do not depend on ℓ. Let

$$\mu^{\mathrm{csm}}_{\mathrm{eig},\ell} := \frac{1}{\ell}\sum_{j=1}^{\ell}\delta_{\lambda_j(\mathbf{A}_\ell)}, \qquad \mu^{\mathrm{csm}}_{\mathrm{sing},\ell} := \frac{1}{\ell}\sum_{j=1}^{\ell}\delta_{s_j(\mathbf{A}_\ell)}.$$

Then $\mathrm{rank}(\mathbf{A}_\ell) = \ell$ *for every* ℓ*, and*

$$\mu^{\mathrm{csm}}_{\mathrm{eig},\ell} \Rightarrow \delta_0, \qquad \mu^{\mathrm{csm}}_{\mathrm{sing},\ell} \Rightarrow \delta_0.$$

More quantitatively, for every $\varepsilon > 0$*,*

$$\frac{1}{\ell}\#\{j : |\lambda_j(\mathbf{A}_\ell)| > \varepsilon\} \le \frac{e^{2B}(1+\log\ell)}{\varepsilon^2\ell}, \qquad \frac{1}{\ell}\#\{j : s_j(\mathbf{A}_\ell) > \varepsilon\} \le \frac{e^{2B}(1+\log\ell)}{\varepsilon^2\ell}. \tag{93}$$

Furthermore,

$$c_B\log\ell \le \mathrm{srank}(\mathbf{A}_\ell) \le C_B(1+\log\ell), \qquad c_B\log\ell \le \mathrm{erank}_1(\mathbf{A}_\ell), \qquad c_B\log\ell \le \mathrm{erank}_2(\mathbf{A}_\ell), \tag{94}$$

for constants $0 < c_B < C_B < \infty$*. The diagonal eigenvalues satisfy*

$$\frac{e^{-B}}{i} \le \lambda_i(\mathbf{A}_\ell) = A_{ii} \le \frac{e^B}{i}, \qquad i = 1,\dots,\ell,$$

and hence the eigenvalue proxies obey

$$\mathrm{srank}_\lambda(\mathbf{A}_\ell) = \Theta_B(\log\ell), \qquad \mathrm{erank}_\lambda(\mathbf{A}_\ell) \ge \ell^{c_B} \tag{95}$$

for all sufficiently large ℓ*. Thus the two empirical laws degenerate at their unrescaled absolute scale, but every scale-invariant rank proxy displayed above diverges. In the terminology introduced above, this is not relative rank collapse. By Proposition 15, the optimal relative Frobenius error of a rank-one approximation obeys*

$$\min_{\mathrm{rank}(\mathbf{R})\le 1}\frac{\|\mathbf{A}_\ell - \mathbf{R}\|_F^2}{\|\mathbf{A}_\ell\|_F^2} = 1 - \frac{1}{\mathrm{srank}(\mathbf{A}_\ell)} \longrightarrow 1. \tag{96}$$

Proof. For each row, subtract its smallest accessible score; this does not change its softmax. Assumption equation 92 then gives

$$e^{-B}\frac{\mathbf{1}_{\{j\le i\}}}{i} \le A_{ij} \le e^B\frac{\mathbf{1}_{\{j\le i\}}}{i}. \tag{97}$$

Since $\mathbf{A}_\ell$ is lower triangular, its eigenvalues are its diagonal entries, and every diagonal entry is strictly positive. Hence $\mathrm{rank}(\mathbf{A}_\ell) = \ell$ and

$$e^{-B}/i \le \lambda_i(\mathbf{A}_\ell) = A_{ii} \le e^B/i. \tag{98}$$

Moreover

$$\|\mathbf{A}_\ell\|_F^2 = \sum_{i=1}^{\ell}\sum_{j=1}^{i}A_{ij}^2 \asymp_B \sum_{i=1}^{\ell}\frac{1}{i} = \Theta_B(\log\ell). \tag{99}$$

The lower bound $\|\mathbf{A}_\ell\|_{\mathrm{op}} \ge 1$ follows because $\mathbf{A}_\ell$ is row-stochastic and hence has the right eigenvector $\mathbf{1}$. For the upper bound, let $C_{ij} = \mathbf{1}_{\{j\le i\}}/i$. By equation 97, $0 \le \mathbf{A}_\ell \le e^B\mathbf{C}_\ell$ entrywise. Since both matrices have nonnegative entries, $|\mathbf{A}_\ell x| \le e^B\mathbf{C}_\ell|x|$ entrywise for every vector x, and therefore

$$\|\mathbf{A}_\ell\|_{\mathrm{op}} \le e^B\|\mathbf{C}_\ell\|_{\mathrm{op}}.$$

The Cesàro matrix $\mathbf{C}_\ell$ has operator norm at most two by Hardy's inequality,

$$\sum_{i=1}^{\ell}\left(\frac{1}{i}\sum_{j=1}^{i}x_j\right)^2 \le 4\sum_{j=1}^{\ell}x_j^2.$$

Therefore $\|\mathbf{A}_\ell\|_{\mathrm{op}} = O_B(1)$. Combining this with equation 99 gives the two-sided stable-rank bound in equation 94. Eckart–Young–Mirsky gives equation 96 because $s_1(\mathbf{A}_\ell) = \|\mathbf{A}_\ell\|_{\mathrm{op}}$, and the stable-rank lower bound sends the displayed error to one.

The two counting bounds in equation 93 follow from Markov's inequality applied to squared eigenvalue moduli and squared singular values, using $\sum_j |\lambda_j(\mathbf{A}_\ell)|^2 \le \|\mathbf{A}_\ell\|_F^2$ and equation 99. Since every eigenvalue and singular value is bounded by $\|\mathbf{A}_\ell\|_{\mathrm{op}} = O_B(1)$, the counting bounds imply weak convergence of both empirical measures to δ_0.

For erank_2, apply the entropy lower bound by min-entropy:

$$\mathrm{erank}_2(\mathbf{A}_\ell) \ge \frac{\|\mathbf{A}_\ell\|_F^2}{\|\mathbf{A}_\ell\|_{\mathrm{op}}^2} = \mathrm{srank}(\mathbf{A}_\ell) \ge c_B \log \ell.$$

The same min-entropy bound for singular-value weights gives

$$\mathrm{erank}_1(\mathbf{A}_\ell) \ge \frac{\|\mathbf{A}_\ell\|_*}{\|\mathbf{A}_\ell\|_{\mathrm{op}}} \ge \frac{\|\mathbf{A}_\ell\|_F^2}{\|\mathbf{A}_\ell\|_{\mathrm{op}}^2} = \mathrm{srank}(\mathbf{A}_\ell) \ge c_B \log \ell,$$

where $\|\mathbf{A}_\ell\|_* \ge \|\mathbf{A}_\ell\|_F^2 / \|\mathbf{A}_\ell\|_{\mathrm{op}}$.

Finally, equation 98 gives

$$\sum_i |\lambda_i(\mathbf{A}_\ell)| = \Theta_B(\log \ell), \qquad \max_i |\lambda_i(\mathbf{A}_\ell)| = \Theta_B(1),$$

so $\mathrm{srank}_\lambda(\mathbf{A}_\ell) = \Theta_B(\log \ell)$. For the eigenvalue entropy, the weights $p_i = |\lambda_i| / \sum_j |\lambda_j|$ are comparable, up to constants depending only on B, to

$$\frac{1/i}{H_\ell}, \qquad H_\ell := \sum_{i=1}^{\ell} \frac{1}{i}.$$

More precisely, there are constants $0 < c_B < C_B < \infty$ such that

$$\frac{c_B}{iH_\ell} \le p_i \le \frac{C_B}{iH_\ell}.$$

Consequently, $-\log p_i \ge \log i + \log H_\ell - \log C_B$. Choose a fixed $i_0 = i_0(B)$ so that this lower bound is nonnegative for all $i \ge i_0$ and all sufficiently large ℓ. On those indices we may use $p_i \ge c_B/(iH_\ell)$ without reversing the inequality, and obtain

$$\begin{aligned} -\sum_i p_i \log p_i &\ge \frac{c_B}{H_\ell} \sum_{i=i_0}^{\ell} \frac{\log i + \log H_\ell - \log C_B}{i} \\ &\ge \frac{c_B}{H_\ell} \sum_{i=i_0}^{\ell} \frac{\log i}{i} - O_B(1) \ge c_B' \log \ell \end{aligned}$$

for all sufficiently large ℓ, because $H_\ell \asymp \log \ell$ and $\sum_{i \le \ell} (\log i)/i = (1/2 + o(1))(\log \ell)^2$. Hence $\mathrm{erank}_\lambda(\mathbf{A}_\ell) \ge \ell^{c_B}$. This proves all claims. □

The fixed-B regime above, and hence the finite-R feature condition, is therefore incompatible with asymptotic energy-rank collapse. One sufficient way to leave that regime is a coherent score advantage for a shared early token. The following result separates the energy in this attention-sink column from the energy in the remaining, possibly broad, attention tail.

Theorem 12 (A sink-energy criterion for approximate rank-one collapse)**.** *Let $\ell \ge 2$, let all accessible scores be finite, and let $\mathbf{A}_\ell$ be their causal row-softmax. Write $a_i := A_{i1}$. For $i \ge 2$, define the conditional non-sink distribution and its collision probability by*

$$r_{ij} := \frac{A_{ij}}{1 - a_i}, \quad 2 \le j \le i, \qquad \chi_i := \sum_{j=2}^{i} r_{ij}^2,$$

and set $\chi_1 := 0$. Define

$$\mathcal{S}_\ell := \sum_{i=1}^{\ell} a_i^2, \qquad \mathcal{D}_\ell := \sum_{i=2}^{\ell} (1 - a_i)^2 \chi_i, \qquad \rho_{2,\ell} := \frac{\mathcal{D}_\ell}{\mathcal{S}_\ell + \mathcal{D}_\ell}. \tag{100}$$

Let $\mathbf{a}_\ell = (a_1, \dots, a_\ell)^\top$ *and*

$$\mathbf{R}_{\mathrm{sink},\ell} := \mathbf{a}_\ell \mathbf{e}_1^\top, \qquad \mathbf{E}_\ell := \mathbf{A}_\ell - \mathbf{R}_{\mathrm{sink},\ell}.$$

Then $\mathrm{rank}(\mathbf{A}_\ell) = \ell$*, but*

$$\|\mathbf{E}_\ell\|_F^2 = \mathcal{D}_\ell, \qquad \|\mathbf{A}_\ell\|_F^2 = \mathcal{S}_\ell + \mathcal{D}_\ell, \tag{101}$$

$$1 \le \mathrm{srank}(\mathbf{A}_\ell) \le 1 + \frac{\mathcal{D}_\ell}{\mathcal{S}_\ell}, \qquad \min_{\mathrm{rank}(\mathbf{R})\le 1} \frac{\|\mathbf{A}_\ell - \mathbf{R}\|_F^2}{\|\mathbf{A}_\ell\|_F^2} \le \rho_{2,\ell}. \tag{102}$$

Let

$$h_{\mathrm{bin}}(t) := -t \log t - (1-t)\log(1-t),$$

with the usual convention at zero and one. Whenever $\rho_{2,\ell} \le 1/2$*,*

$$\mathrm{erank}_2(\mathbf{A}_\ell) \le \exp\left\{h_{\mathrm{bin}}(\rho_{2,\ell}) + \rho_{2,\ell}\log(\ell - 1)\right\}. \tag{103}$$

Consequently,

$$\frac{\mathcal{D}_\ell}{\mathcal{S}_\ell} \longrightarrow 0 \quad \Longrightarrow \quad \mathrm{srank}(\mathbf{A}_\ell) \longrightarrow 1 \quad \textit{and} \quad \min_{\mathrm{rank}(\mathbf{R})\le 1} \frac{\|\mathbf{A}_\ell - \mathbf{R}\|_F}{\|\mathbf{A}_\ell\|_F} \longrightarrow 0,$$

while

$$\frac{\mathcal{D}_\ell}{\mathcal{S}_\ell}\log \ell \longrightarrow 0 \quad \Longrightarrow \quad \mathrm{erank}_2(\mathbf{A}_\ell) \longrightarrow 1. \tag{104}$$

The same conclusion for erank_1 *requires stronger control of the weak singular-value tail. In fact, set*

$$\rho_{1,\ell} := \frac{\|\mathbf{E}_\ell\|_*}{\sqrt{\mathcal{S}_\ell} + \|\mathbf{E}_\ell\|_*}.$$

Whenever $\rho_{1,\ell} \le 1/2$*,*

$$\mathrm{erank}_1(\mathbf{A}_\ell) \le \exp\left\{h_{\mathrm{bin}}(\rho_{1,\ell}) + \rho_{1,\ell}\log(\ell - 1)\right\}. \tag{105}$$

Thus $\|\mathbf{E}_\ell\|_* \log \ell = o(\sqrt{\mathcal{S}_\ell})$ *is sufficient for* $\mathrm{erank}_1(\mathbf{A}_\ell) \to 1$*; Frobenius-energy collapse alone does not make the bound in equation 105 vanish.*

Proof. The first column of $\mathbf{A}_\ell$ is $\mathbf{a}_\ell$, while $\mathbf{E}_\ell$ is supported on columns $2, \dots, \ell$. Thus $\mathbf{R}_{\mathrm{sink},\ell}$ and $\mathbf{E}_\ell$ are orthogonal in the Frobenius inner product, and row i of $\mathbf{E}_\ell$ has squared norm $(1-a_i)^2\chi_i$. This proves equation 101. Moreover,

$$\|\mathbf{A}_\ell\|_{\mathrm{op}} \ge \|\mathbf{A}_\ell \mathbf{e}_1\|_2 = \|\mathbf{a}_\ell\|_2 = \sqrt{\mathcal{S}_\ell}.$$

Combining this inequality with the Frobenius split proves the stable-rank bound. Taking $\mathbf{R} = \mathbf{R}_{\mathrm{sink},\ell}$ proves the relative rank-one approximation bound.

By the Eckart–Young–Mirsky theorem,

$$\sum_{k=2}^{\ell} s_k(\mathbf{A}_\ell)^2 \le \|\mathbf{A}_\ell - \mathbf{R}_{\mathrm{sink},\ell}\|_F^2 = \mathcal{D}_\ell.$$

Hence the total mass outside the largest normalized squared singular value is at most $\rho_{2,\ell}$. For a fixed tail mass ρ, entropy is maximized when that mass is uniform on the remaining $\ell - 1$ coordinates, giving $h_{\mathrm{bin}}(\rho) + \rho\log(\ell - 1)$. This function is increasing for $0 \le \rho \le 1/2$, which proves equation 103 and the two limiting claims.

The nuclear-norm form of Eckart–Young–Mirsky similarly gives

$$\sum_{k=2}^{\ell} s_k(\mathbf{A}_\ell) \le \|\mathbf{E}_\ell\|_*, \qquad s_1(\mathbf{A}_\ell) \ge \sqrt{\mathcal{S}_\ell}.$$

Thus the normalized singular-value tail mass is at most $\rho_{1,\ell}$, and the same entropy argument proves equation 105. Finally, positivity of every diagonal entry again gives $\mathrm{rank}(\mathbf{A}_\ell) = \ell$. □

The separate nuclear-tail requirement is consistent with the experiment: GPT-2 has nearly constant srank and erank_2 but a growing erank_1, which can be produced by many weak singular directions carrying little squared energy.

The quantities in equation 100 are directly measurable from an attention matrix. The following pre-softmax condition is sufficient and allows the sink probability a_i to vanish: coherent accumulation down one column can dominate Frobenius energy even when most probability remains in a diffuse background.

Corollary 2 (A logarithmic sink margin with a diffuse background)**.** *Let all accessible entries of* $\mathbf{S}_\ell$ *be finite, set*

$$\mathbf{A}_\ell := \mathrm{softmax}_{\mathrm{row}}^{\mathrm{causal}}(\mathbf{S}_\ell), \qquad a_i := A_{i1},$$

and, for $i \geq 2$*, let*

$$G_i := S_{i1} - \max_{2\leq j\leq i} S_{ij}, \qquad \chi_i := \frac{\sum_{j=2}^{i} e^{2S_{ij}}}{\left(\sum_{j=2}^{i} e^{S_{ij}}\right)^2}.$$

Define

$$\mathcal{S}_\ell := \sum_{i=1}^{\ell} a_i^2, \qquad \mathcal{D}_\ell := \sum_{i=2}^{\ell} (1-a_i)^2 \chi_i.$$

Suppose that for constants $c > 1/2$ *and* $1 \leq C < \infty$*, independent of* i *and* ℓ*,*

$$G_i \geq c \log i - C, \qquad \chi_i \leq \frac{C}{i-1}, \qquad i = 2, \ldots, \ell. \tag{106}$$

Set $\beta := \min\{2c-1, 1\} > 0$*. Then*

$$\mathcal{S}_\ell = \Omega_{c,C}(\ell^\beta), \qquad \mathcal{D}_\ell = O_C(1 + \log \ell),$$

and consequently

$$0 \leq \mathrm{srank}(\mathbf{A}_\ell) - 1 = O_{c,C}\left(\frac{\log \ell}{\ell^\beta}\right),$$

$$\min_{\mathrm{rank}(\mathbf{R})\leq 1} \frac{\|\mathbf{A}_\ell - \mathbf{R}\|_F}{\|\mathbf{A}_\ell\|_F} = O_{c,C}\left(\sqrt{\frac{\log \ell}{\ell^\beta}}\right), \qquad \mathrm{erank}_2(\mathbf{A}_\ell) \longrightarrow 1.$$

Proof. Let $z_i := \sum_{j=2}^{i} e^{S_{ij} - S_{i1}}$. The first inequality in equation 106 gives

$$z_i \leq e^C i^{1-c}, \qquad a_i = \frac{1}{1+z_i} \geq \frac{1}{1 + e^C i^{1-c}}.$$

It follows by summation that $\mathcal{S}_\ell = \Omega_{c,C}(\ell^{2c-1})$ when $1/2 < c < 1$, and $\mathcal{S}_\ell = \Omega_C(\ell)$ when $c \geq 1$.

The displayed pre-softmax formula for χ_i is exactly the collision probability in Theorem 12. Therefore $\mathcal{D}_\ell \leq C H_{\ell-1} = O_C(1 + \log \ell)$. The conclusions now follow from Theorem 12. □

The collision condition is a Rényi-2 diffuseness requirement, not a demand that every non-sink score be nearly equal. One transparent but stronger score-only certificate is

$$\max_{2\leq j\leq i} S_{ij} - \min_{2\leq j\leq i} S_{ij} \leq C_0,$$

which gives $\chi_i \leq e^{C_0}/(i-1)$. The threshold $c > 1/2$ is weaker than the full $\log i$ advantage that is sufficient in the worst case to keep the sink probability itself bounded away from zero; the exact $\mathcal{S}_\ell$–$\mathcal{D}_\ell$ theorem is the more flexible condition to test empirically.

Condition equation 106 has a natural SVD interpretation, but it requires singular-vector orientation as well as singular values. Row-center the scores with

$$\mathbf{\Pi}_\ell := \mathbf{I}_\ell - \ell^{-1}\mathbf{1}\mathbf{1}^\top, \qquad \widetilde{\mathbf{S}}_\ell := \mathbf{S}_\ell \mathbf{\Pi}_\ell.$$

This operation subtracts one constant from each row and hence leaves the causal softmax unchanged. It is an offline analytical diagnostic and may use future score entries; it is not part of the causal computation. If $\mathbf{S}_\ell = \mathbf{Q}_\ell \mathbf{K}_\ell^\top / \tau_\ell$, then

$$\widetilde{\mathbf{S}}_\ell = \frac{\mathbf{Q}_\ell (\mathbf{\Pi}_\ell \mathbf{K}_\ell)^\top}{\tau_\ell},$$

so this is precisely the score matrix formed with token-centered keys.

Corollary 3 (An aligned score singular mode is sufficient)**.** *Write a leading singular-mode decomposition*

$$\widetilde{\mathbf{S}}_\ell = \sigma_1 \mathbf{u}\mathbf{v}^\top + \mathbf{E}, \qquad \|\mathbf{E}\|_{\mathrm{op}} = \sigma_2(\widetilde{\mathbf{S}}_\ell),$$

and assume that, after a simultaneous sign choice of $\mathbf{u}, \mathbf{v}$*, one has* $u_i \geq 0$ *for every* i*. Suppose that for constants* $c > 1/2$ *and* $1 \leq C < \infty$*, every* $i \geq 2$ *satisfies*

$$\sigma_1 u_i \left(v_1 - \max_{2 \leq j \leq i} v_j \right) - 2\|\mathbf{E}_{i,:}\|_\infty \geq c \log i - C, \tag{107}$$

$$\sigma_1 u_i \left(\max_{2 \leq j \leq i} v_j - \min_{2 \leq j \leq i} v_j \right) + 2\|\mathbf{E}_{i,:}\|_\infty \leq C. \tag{108}$$

Then the conclusions of Corollary 2 hold. Since $\|\mathbf{E}_{i,:}\|_\infty \leq \|\mathbf{E}\|_{\mathrm{op}} = \sigma_2$*, a stronger but simpler singular-value residual certificate is obtained by replacing each row residual norm above by* $\sigma_2(\widetilde{\mathbf{S}}_\ell)$*.*

Proof. Score differences are unchanged by row-centering. For $2 \leq j \leq i$,

$$S_{i1} - S_{ij} = \sigma_1 u_i (v_1 - v_j) + E_{i1} - E_{ij}.$$

The first aligned-SVD inequality therefore implies the lower bound on G_i in equation 106. The second bounds the non-sink score range by C, and hence gives $\chi_i \leq e^C/(i-1)$. Thus Corollary 2 applies after enlarging its constant to $C' := e^C$. □

For example, if $u_i \gtrsim \ell^{-1/2}$ across the query rows, $\mathbf{v}$ has a first-token-localized spike (e.g., at BOS) separated from a nearly flat non-sink background, and the residual is rowwise bounded, then equation 107 asks for a centered singular spike of at least order $\sqrt{\ell} \log \ell$. More precisely, when $\sigma_1 u_i \asymp \log \ell$, the non-sink spread in $\mathbf{v}$ must be $O(1/\log \ell)$ in equation 108. The operator-norm substitution by σ_2 is convenient but conservative; the entrywise row residual in equation 107 is the milder condition to test on a trained score matrix. The logarithmic score advantage necessarily violates the fixed-B hypothesis of Theorem 11. The more general $\mathcal{S}_\ell$–$\mathcal{D}_\ell$ criterion in Theorem 12 also permits exceptional rows, provided their total residual energy remains negligible relative to $\mathcal{S}_\ell / \log \ell$.

Remark 10 (Why score singular values alone cannot imply collapse). Even the singular values of the row-centered scores do not retain the causal location of a dominant key. Let $m_\ell := 2 \log \ell$ and, for $a \in \{1, \ell\}$, set

$$\mathbf{S}_\ell^{(a)} := m_\ell \mathbf{1} \left(\mathbf{e}_a - \ell^{-1} \mathbf{1} \right)^\top.$$

Let

$$\mathbf{A}_\ell^{(a)} := \mathrm{softmax}_{\mathrm{row}}^{\mathrm{causal}} \left(\mathbf{S}_\ell^{(a)} \right).$$

Both matrices are row-centered and rank one, and both have the same sole nonzero singular value $m_\ell \sqrt{\ell - 1}$. This also defeats conditions on the singular values of $\mathbf{Q}$ and $\mathbf{K}$ separately: with head dimension one, take $\mathbf{Q} = m_\ell \mathbf{1}$ and $\mathbf{K} = \mathbf{e}_a - \ell^{-1}\mathbf{1}$; $\mathbf{Q}$ is identical in the two examples and the two key vectors have the same norm. For $a = 1$, every row gives token one a score advantage m_ℓ, so

$$1 - A_{i1}^{(1)} \leq \ell e^{-m_\ell} = \ell^{-1}.$$

Theorem 12 then gives $\mathrm{srank}(\mathbf{A}_\ell^{(1)}) \to 1$ and $\mathrm{erank}_2(\mathbf{A}_\ell^{(1)}) \to 1$.

For $a = \ell$, by contrast, the first $\ell - 1$ causal rows see equal accessible scores. Their principal block is the Cesàro matrix $\mathbf{C}_{\ell-1}$, so

$$\|\mathbf{A}_\ell^{(\ell)}\|_F^2 = H_{\ell-1} + O(1), \qquad 1 \leq \|\mathbf{A}_\ell^{(\ell)}\|_{\mathrm{op}} \leq 3,$$

and hence $\mathrm{srank}(\mathbf{A}_\ell^{(\ell)}) = \Theta(\log \ell)$. Thus identical score singular values can yield opposite effective-rank behavior. Thus singular values alone cannot determine collapse. The sufficient aligned-mode condition above additionally specifies a shared early causal token and same-sign loading across query rows. Raw score singular values are even less informative because softmax is unchanged by arbitrary row shifts $\mathbf{S} \mapsto \mathbf{S} + \mathbf{z}\mathbf{1}^\top$.

Some Properties of Spectral Ranks

For completeness, we record a few deterministic relations among the spectral rank proxies. Let $X \neq 0$ have positive singular values $s_1 \geq \cdots \geq s_m > 0$, where $m = \mathrm{rank}(X)$, and define

$$q_i := \frac{s_i}{\sum_j s_j}, \qquad p_i := \frac{s_i^2}{\sum_j s_j^2}.$$

Then

Proposition 13 (Ordering of spectral rank proxies)**.**

$$1 \leq \mathrm{srank}(X) \leq \mathrm{erank}_2(X) \leq \mathrm{erank}_1(X) \leq \mathrm{rank}(X). \tag{109}$$

Proof. Indeed, the squared-weight distribution p majorizes the singular-value distribution q. To see this, assume both are sorted decreasingly and let $A_k = \sum_{i \leq k} q_i$. For $1 \leq k < m$, sortedness gives

$$(1 - A_k) \sum_{i \leq k} q_i^2 - A_k \sum_{j > k} q_j^2 = \sum_{i \leq k < j} q_i q_j (q_i - q_j) \geq 0.$$

Rearranging, and including the equality case $k = m$, yields

$$\frac{\sum_{i \leq k} q_i^2}{\sum_i q_i^2} \geq \sum_{i \leq k} q_i = A_k.$$

Thus $p \succ q$. Since Shannon entropy is Schur-concave, $H(p) \leq H(q)$, which gives $\mathrm{erank}_2(X) \leq \mathrm{erank}_1(X)$. Also, Shannon entropy dominates min-entropy:

$$H(p) \geq -\log \max_i p_i, \qquad \mathrm{erank}_2(X) = e^{H(p)} \geq \frac{1}{\max_i p_i} = \mathrm{srank}(X).$$

Finally, $H(p), H(q) \leq \log m$, while $\|X\|_F^2 \leq m \|X\|_{\mathrm{op}}^2$. These give the remaining bounds in equation 109. A useful norm form of the same argument is

$$\mathrm{erank}_1(X) \geq \frac{\|X\|_*}{\|X\|_{\mathrm{op}}} \geq \frac{\|X\|_F^2}{\|X\|_{\mathrm{op}}^2} = \mathrm{srank}(X).$$

□

Proposition 14 (Entropy and Schatten-norm lower bounds)**.** *For the same X,*

$$\mathrm{erank}_1(X) \geq \frac{\|X\|_*^2}{\|X\|_F^2}, \qquad \mathrm{erank}_2(X) \geq \frac{\|X\|_F^4}{\sum_i s_i(X)^4} \geq \mathrm{srank}(X). \tag{110}$$

Proof. The first inequality is the Shannon-versus-Rényi-2 entropy bound applied to q. Applying the same bound to p gives the first inequality for erank_2. Finally, $\sum_i s_i^4 \leq s_1^2 \sum_i s_i^2$, which gives the last inequality. □

Proposition 15 (Stable rank and best rank-one approximation)**.** *The optimal relative Frobenius error of a rank-one approximation is*

$$\min_{\mathrm{rank}(R) \leq 1} \frac{\|X - R\|_F^2}{\|X\|_F^2} = 1 - \frac{1}{\mathrm{srank}(X)}. \tag{111}$$

Proof. The Eckart–Young–Mirsky theorem gives the minimum squared error as $\|X\|_F^2 - s_1(X)^2$. Divide by $\|X\|_F^2$. □

Proposition 16 (Scale invariance)**.** *For every nonzero scalar c,*

$$\mathrm{srank}(cX) = \mathrm{srank}(X), \qquad \mathrm{erank}_1(cX) = \mathrm{erank}_1(X), \qquad \mathrm{erank}_2(cX) = \mathrm{erank}_2(X).$$

Proof. All singular values are multiplied by $|c|$, so the normalized distributions q and p and the stable-rank ratio are unchanged. □