Supplementary Material
The Residual Stream's Effective Depth

Barak Gahtan BARAKGAHTAN@CS.TECHNION.AC.IL
Technion Israel Institute of Technology

Ido Galil IGALIL@NVIDIA.COM
Nvidia

Alex M. Bronstein BRON@CS.TECHNION.AC.IL
Technion Israel Institute of Technology & ISTA Institute of Science and Technology Austria

Editors: Andy Song, Bo Han and Sarah Erfani

Appendix Roadmap

The appendix is organised as an evidence map for the main-text claims. The full per-model numeric table underlying the main regime figure is kept in the main paper (Appendix A); this supplement collects the remaining material. Appendices S1–S2 contain the mathematical and synthetic foundations of D_{eff}: basic properties, the orthogonal-update reference F_L, and controlled constructions that separate accumulated states from layer updates. Appendix S3 gives the matched-reference calibration that separates the universal F_L gap from update-size, h_0, and measured update-correlation effects. Appendix S4 expands the position-0, token-normalisation, and top-PC controls from the main representative panel to all sixteen models, adds the Gemma-3 size extension, and reports Gemma-style architectural probes at initialisation and under a short pretraining budget. Appendix S5 gives the CKA convergence proof. Appendices S6–S9 then collect robustness checks for the aggregation itself, including lag truncation, alternative aggregations, the norm-process mixing analogy, and alternative tapers. Appendix S10 documents extraction and CKA details and the trained-vs-random null baseline figure, while Appendices S11–S14 report auxiliary diagnostics and representation-similarity robustness checks. Appendices S15–S17 support the interpretation of D_{eff} as a global residual-stream diagnostic: they compare against ShortGPT-style shallow behaviour, report the per-update diagnostic, and give the controlled residual-carry intervention. The final appendices provide statistical stability and boundary checks: bootstrap confidence intervals (Appendix S18), pruning boundary checks (Appendix S19), and the external capability-scaling check in Appendix S20.

Appendix S1. Properties of D_{eff}

We state and prove the basic properties of the effective depth diagnostic. Throughout, $\hat{\varrho}(k)$ denotes the CKA-based layer similarity autocorrelation (Definition 1 of the main paper), which satisfies $\hat{\varrho}(k) \in [0, 1]$ since CKA $\in [0, 1]$.

Theorem 1 (Properties of D_{eff}) *Let $\hat{\varrho}(k) \in [0, 1]$ for all $k = 1, \ldots, L-1$. Then:*

© 2026 B. Gahtan, I. Galil & A.M. Bronstein.

(i) ***Bounds:*** $1 \le D_{\text{eff}} \le L$.

(ii) ***Independence limit:*** *If* $\hat{\varrho}(k) = 0$ *for all* $k \ge 1$, *then* $D_{\text{eff}} = L$.

(iii) ***Maximum redundancy:*** *If* $\hat{\varrho}(k) = 1$ *for all* k, *then* $D_{\text{eff}} = 1$ *for every* $L \ge 1$.

(iv) ***Algebraic structure:*** $D_{\text{eff}} = L/\hat{S}$ *where* $\hat{S} = 1 + 2\sum_{k=1}^{L-1}(1 - k/L)\hat{\varrho}(k)$. *This has the same algebraic form as the Bartlett spectral density estimator at frequency zero, but we do not claim that* $\hat{\varrho}$ *has the Toeplitz positive-semidefinite structure of a true autocorrelation sequence.*

Proof **(i)** The denominator is $\hat{S} = 1+2\sum_{k=1}^{L-1}(1-k/L)\hat{\varrho}(k)$. Since $\hat{\varrho}(k) \ge 0$ and $(1-k/L) > 0$ for $k < L$, every term in the sum is non-negative, so $\hat{S} \ge 1$ and therefore $D_{\text{eff}} \le L$. For the upper bound on $\hat{S}$: since $\hat{\varrho}(k) \le 1$,

$$\hat{S} \;\le\; 1 + 2\sum_{k=1}^{L-1}\left(1 - \frac{k}{L}\right) \;=\; 1 + 2\left((L-1) - \frac{(L-1)L/2}{L}\right) \;=\; L.$$

Hence $D_{\text{eff}} = L/\hat{S} \ge 1$.

(ii) If $\hat{\varrho}(k) = 0$ for $k \ge 1$, then $\hat{S} = 1$ and $D_{\text{eff}} = L$.

(iii) If $\hat{\varrho}(k) = 1$ for all k, then

$$\hat{S} = 1 + 2\sum_{k=1}^{L-1}\left(1 - \frac{k}{L}\right) = 1 + 2(L-1) - 2 \cdot \frac{(L-1)L/2}{L} = 1 + 2(L-1) - (L-1) = L.$$

So $D_{\text{eff}} = L/L = 1$ exactly.

(iv) This is immediate from the definition. ∎

Appendix S2. Controlled Synthetic Constructions and Empirical Validation

To clarify what D_{eff} does and does not measure, we evaluate controlled synthetic constructions using the same linear-CKA and Bartlett aggregation pipeline as the main experiments. Unless otherwise noted, these controls use $N = 10{,}000$, $L = 32$, and $d = 1024$. The goal is qualitative validation of the diagnostic: can it rise on genuinely decorrelated states, and can accumulated states remain highly redundant even when the underlying updates are diverse? Table S1 summarizes the three constructions, and Figure S1 contrasts them with real models; the per-construction analyses below expand on each.

Independent-state baseline. For i.i.d. Gaussian states $H_1, \ldots, H_L$, the empirical diagnostic gives $D_{\text{eff}}/L = 0.258$ and $\hat{\varrho}(1) = 0.093$. This is well above the real-model range, showing that the estimator does not collapse on unrelated states. It is below the idealized independence limit of 1 because finite-sample linear CKA remains positive when d/N is non-negligible, even for unrelated random matrices.

Orthogonal-state construction. When the states themselves are explicitly orthogonal across layers, the metric returns $D_{\text{eff}}/L = 0.996$ with $\hat{\varrho}(1) = 0.0001$. This rules out the failure mode in which the diagnostic is intrinsically unable to approach full depth.

Table S1: Controlled synthetic constructions. Independent and orthogonal states show that the metric rises when states are genuinely decorrelated; orthogonal-update sweeps show that $D_{\text{eff}}(h)$ can remain low even when the updates themselves are diverse.

Construction	$D_{\text{eff}}(h)/L$	$D_{\text{eff}}(f)/L$	$\hat{\varrho}_h(1)$
Independent states	0.258	–	0.093
Orthogonal states	0.996	–	0.0001
Orthogonal updates (scale sweep)	0.031–0.056	0.258	0.911–0.9999

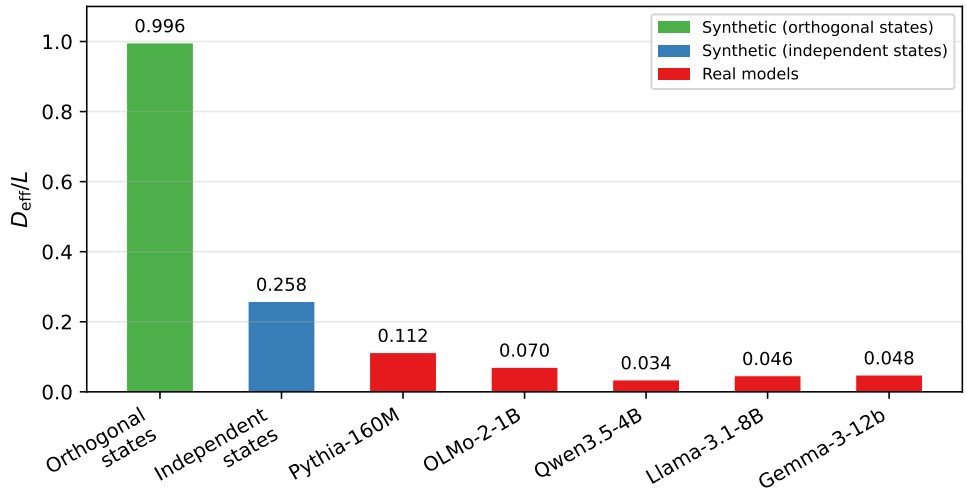


Figure S1: D_{eff}/L for synthetic constructions (green, blue) and real models (red). Orthogonal states recover $D_{\text{eff}}/L = 0.996$; independent states give 0.258; real models fall well below both, confirming that the low values reflect genuine redundancy in the accumulated residual stream.

Orthogonal-update construction. The key control separates state redundancy from update diversity. We construct mutually orthogonal per-layer updates f_ℓ and define accumulated states by $h_\ell = \sum_{j \le \ell} f_j$. In this construction the updates remain diverse by design, yet the accumulated states can still be highly similar. Table S2 shows the resulting sweep over update scale.

The population version of this construction has a simple closed form. If the updates have common covariance and zero cross-covariance across layers, then the update CKA is zero for distinct layers and, for $i < j$,

$$\mathrm{CKA}(h_i, h_j) = \frac{\| \sum_{r \le i} \Sigma_{f_r f_r} \|_F^2}{\| \sum_{r \le i} \Sigma_{f_r f_r} \|_F \| \sum_{r \le j} \Sigma_{f_r f_r} \|_F} = \frac{i}{j}.$$

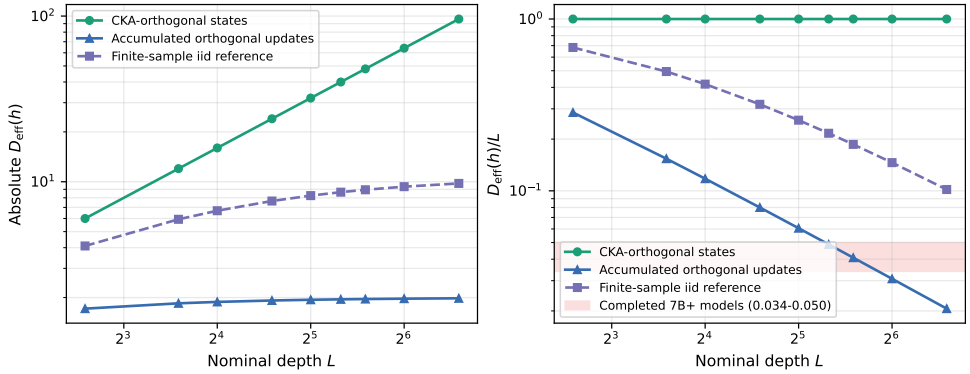


Figure S2: Depth sweep for synthetic residual constructions. Orthogonal states recover $D_{\text{eff}} = L$, whereas accumulated orthogonal updates obey $D_{\text{eff}} = 2L/(L+1) < 2$ and therefore have $D_{\text{eff}}/L = O(1/L)$. The finite-sample iid reference matches the positive CKA bias observed in the sampled iid control; it is shown only as an estimator calibration.

Table S2: Orthogonal-update construction (N=10,000, L=32, d=1024). Diverse updates can coexist with very low accumulated-state depth.

$\|f\|/\|h\|$ scale	$D_{\text{eff}}(h)/L$	$D_{\text{eff}}(f)/L$	$\hat{\varrho}_h(1)$
0.01	0.031	0.258	0.9999
0.05	0.032	0.258	0.998
0.10	0.034	0.258	0.992
0.30	0.044	0.258	0.964
0.50	0.049	0.258	0.945
1.00	0.054	0.258	0.924
2.00	0.055	0.258	0.915
5.00	0.056	0.258	0.911

Thus the lag-k profile is $\hat{\varrho}(k) = \frac{1}{L-k}\sum_{i=1}^{L-k} i/(i+k)$, and direct substitution into Eq. (5) of the main paper gives $D_{\text{eff}} = 2L/(L+1) < 2$. This is why the normalized ratio D_{eff}/L falls with nominal depth even under maximally diverse updates (Figure S2).

Implications for interpreting real models. These constructions establish three points. First, D_{eff} is not broken: it rises substantially on independent states and approaches L on orthogonal states (Figure S1). Second, low $D_{\text{eff}}(h)$ does not imply that layers compute identical updates: even perfectly orthogonal updates can yield very low accumulated-state depth (Figure S3). Third, the real-model results combine both phenomena. Residual accumulation keeps the sequence of states seen by downstream layers highly similar, and the real-model updates are also correlated, as shown by $D_{\text{eff}}^{\text{update}}$ in Appendix S16.

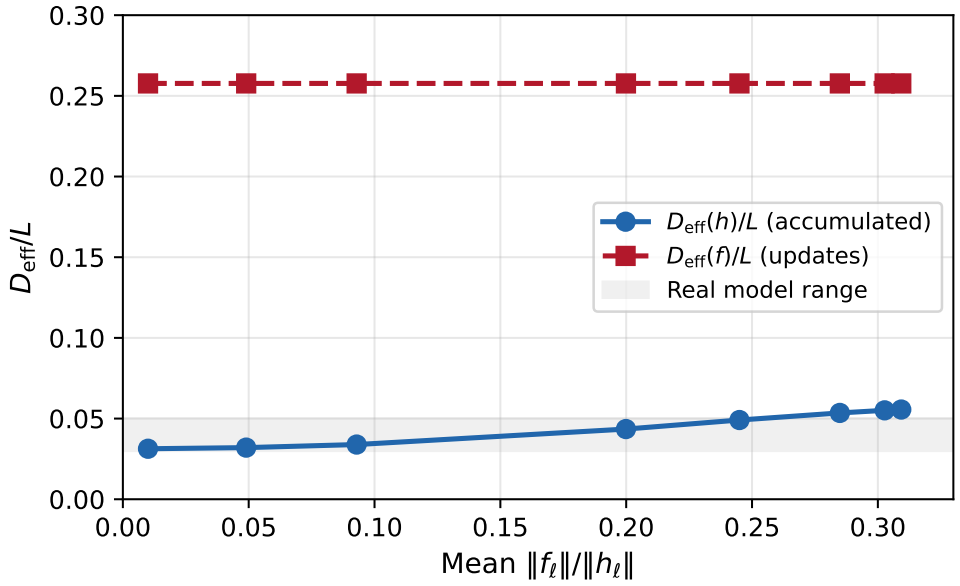


Figure S3: Orthogonal-update sweep (N=10,000, L=32, d=1024). $D_{\text{eff}}(f)/L$ (updates, red) remains constant at 0.258 regardless of scale, while $D_{\text{eff}}(h)/L$ (accumulated, blue) stays in the real-model range (gray band). Diverse updates do not imply diverse accumulated states.

Appendix S3. Matched Reference Baselines

The universal $F_L = 2L/(L+1)$ reference is intentionally simple: it assumes equal-norm, mutually orthogonal updates and no persistent initial component. Table S3 asks which model-specific quantities change that reference. F_{norm} retains the measured update covariance-norm profile while setting update cross-covariances to zero. $F_{h_0+\text{norm}}$ additionally retains the persistent initial state. $F_{h_0+\text{corr}}$ keeps the measured update-size profile, h_0, and the measured update similarity lag profile in a scalar accumulated-state reference.

Two points matter for the interpretation. First, matching update sizes and h_0 does not remove the sub-F_L finding: all sixteen models remain below F_{norm} and $F_{h_0+\text{norm}}$. Second, once the measured update-similarity profile is retained, the sign flips for every model: $F_{h_0+\text{corr}}$ predicts lower D_{eff} than observed. Thus the gap to the universal F_L reference is largely explained by correlated updates, while the observed accumulated states retain more diversity than a scalar reference that preserves only update sizes and average update similarities. This is why we treat F_L as a first structural yardstick, not as a definitive null.

Appendix S4. Full Position-0 and Norm Controls

Table S4 expands the main-text position-0 and norm-control analysis to all sixteen models. The controls are symmetric: they are applied to default outliers and non-outliers alike. The strongest invariant result is that all sixteen models are below the orthogonal-update

Table S3: Matched reference baselines for D_{eff}. F_L is the universal equal-update reference; F_{norm} matches measured update covariance norms; $F_{h_0+\text{norm}}$ additionally retains the persistent initial state; $F_{h_0+\text{corr}}$ also retains the measured update-similarity lag profile. Gaps are relative to each reference, in percent.

Model	L	D_{eff}	gap to F_L	gap to F_{norm}	gap to $F_{h_0+\text{norm}}$	gap to $F_{h_0+\text{corr}}$
Pythia-70M	6	1.23	+28.3	+59.0	+59.0	-14.7
Pythia-160M	12	1.34	+27.4	+54.5	+54.5	-24.6
Pythia-410M	24	1.46	+23.7	+46.6	+46.6	-39.0
Pythia-1B	16	1.45	+23.0	+34.9	+34.9	-38.6
Pythia-1.4B	24	1.42	+26.0	+42.1	+42.1	-36.9
OLMo-2-1B	16	1.12	+40.4	+64.1	+61.7	-5.8
OLMo-2-7B	32	1.14	+41.2	+69.4	+68.1	-7.1
OLMo-2-13B	40	1.18	+39.7	+63.4	+62.3	-9.0
Llama-3.1-8B	32	1.48	+23.9	+40.5	+40.5	-40.1
Mistral-7B	32	1.53	+21.1	+16.7	+16.7	-43.8
Gemma-3-12b	48	2.29	-17.1	+48.6	+48.6	-89.1
Qwen3.5-0.8B	24	1.07	+44.1	+64.4	+63.4	-3.0
Qwen3.5-2B	24	1.08	+43.6	+71.6	+71.3	-4.2
Qwen3.5-4B	32	1.08	+44.4	+73.6	+73.4	-2.9
Qwen3.5-9B	32	1.15	+40.9	+76.8	+76.8	-9.0
Qwen3.5-27B	64	1.35	+31.4	+68.7	+68.6	-27.7

reference after token-normalisation and after top-1-PC removal. First-non-BOS pooling is also sub-reference for fifteen of sixteen models; the exception is Pythia-70M, a six-layer boundary case.

Table S5 extends the same controls to two additional Gemma-3 sizes. The default-pooling anomaly is not isolated to Gemma-3-12b: Gemma-3-27B is also above-reference by default, while Gemma-3-4B sits close to the reference. In both cases, first-non-BOS, token-normalised, and top-PC-removed variants are sub-reference.

To test whether two Gemma-style block features directly account for this behaviour, we also run Gemma-scale *forward-pass* interventions on the trained Gemma-3-4B, 12B, and 27B checkpoints. The intervention is not retraining: post-normalisation replaces the post-attention and post-MLP RMSNorms by identities, while no-QK replaces the query/key RMSNorms by identities. Table S6 shows that post-normalisation is not the source of the above-reference default measurement: removing it drives every Gemma size into a highly sub-reference regime, with gaps near +47–+49% across controls. QK normalisation instead modulates the position/default-pooling channel. Removing QK normalisation makes the default anomaly much larger at all three sizes, and for Gemma-3-27B it even makes the first-non-BOS and token-normalised variants above-reference. Removing both features is dominated by the post-normalisation ablation and again collapses the stream. Thus the Gemma behaviour is not explained by post-normalisation alone; QK normalisation appears to suppress, rather than cause, the position-0/default-pooling anomaly.

Table S4: Position-0 and norm control variants of D_{eff} for the full model set (N=10,000, FineWeb-Edu). "Default" columns reproduce the canonical pooling used in Table 2 of the main paper. "First non-BOS" replaces the pooled mean with position 1 only; "token-normed" divides each token's hidden state by its ℓ_2 norm before mean pooling; "top-1/3 PC" projects out dominant cross-sample directions from each layer's pooled matrix. The only above-reference model in the canonical setup, Gemma-3-12b, becomes strongly sub-reference under every control variant.

		Default		Control variants (gap to F_L, %)				
Model	L	D_{eff}/L	gap%	first non-BOS	token-normed	top-1 PC	top-3 PC	max pos-0 norm ratio
Pythia-70M	6	0.2047	+28.3	-6.5	+30.1	+20.0	+11.4	10.0× at L_3
Pythia-160M	12	0.1117	+27.4	+22.1	+33.7	+27.7	+23.3	18.4× at L_4
Pythia-410M	24	0.0610	+23.7	+8.7	+38.4	+34.3	+27.9	42.1× at L_9
Pythia-1B	16	0.0906	+23.0	+13.7	+40.0	+35.3	+29.7	35.3× at L_6
Pythia-1.4B	24	0.0592	+26.0	+14.1	+41.8	+37.2	+30.9	24.6× at L_5
OLMo-2-1B	16	0.0702	+40.4	+30.7	+40.3	+34.9	+30.6	2.7× at L_9
OLMo-2-7B	32	0.0357	+41.2	+34.5	+41.1	+36.4	+33.5	1.5× at L_1
OLMo-2-13B	40	0.0294	+39.7	+34.7	+39.8	+36.0	+29.5	1.5× at L_1
Llama-3.1-8B	32	0.0461	+23.9	+41.0	+41.3	+38.6	+35.1	309.7× at L_2
Mistral-7B	32	0.0478	+21.1	+33.4	+42.0	+34.8	+33.0	202.7× at L_2
Gemma-3-12b	48	0.0478	-17.0	+41.6	+34.4	+31.4	+26.6	11.0× at L_7
Qwen3.5-0.8B	24	0.0447	+44.1	+37.0	+43.9	+41.8	+34.1	5.7× at L_7
Qwen3.5-2B	24	0.0451	+43.6	+34.6	+43.2	+41.6	+37.3	4.5× at L_{15}
Qwen3.5-4B	32	0.0337	+44.4	+26.5	+44.6	+43.0	+39.3	3.9× at L_7
Qwen3.5-9B	32	0.0358	+40.9	+16.4	+44.3	+39.2	+35.5	3.4× at L_7
Qwen3.5-27B	64	0.0211	+31.4	+16.4	+44.3	+38.4	+35.2	3.5× at L_{22}

Table S5: Additional Gemma-3 control variants. Values are gaps to F_L in percent. The default-pooling anomaly persists across larger Gemma-3 sizes, but all controls move both extra models into the sub-reference regime.

Model	Default	first non-BOS	token-normed	top-1 PC	top-3 PC
Gemma-3-4B	+2.2	+35.0	+30.8	+23.4	+25.4
Gemma-3-27B	−8.3	+25.5	+7.4	+3.0	+28.0

The preceding interventions act on trained checkpoints. They therefore show that Gemma's post-branch and QK normalisers modulate the measured profile, but they do not by themselves show whether the anomaly is an architectural property present before training. We next run two deliberately limited probes. First, we instantiate Gemma-3-4B from its configuration, apply the same norm interventions, and measure D_{eff} at initialisation only. Second, we train Gemma-3-270M variants from scratch on a short FineWeb-Edu token stream. These runs are not meant to reproduce Gemma's full pretraining, instruction tuning, or post-training pipeline. They are controlled stress tests of whether the norm switches are sufficient to induce the anomaly under simplified conditions.

Table S6: Gemma-scale forward-pass interventions. Values are gaps to F_L in percent under the same N=10,000 FineWeb-Edu extraction as Table S5. "No post" replaces the post-attention and post-MLP RMSNorms by identities; "no QK" replaces the query/key RMSNorms by identities.

Model	Intervention	Default	first non-BOS	token-normed	top-1 PC	top-3 PC
Gemma-3-4B	none	+2.2	+35.0	+30.8	+23.4	+25.4
	no post	+48.1	+48.5	+48.3	+40.8	+47.5
	no QK	−60.2	+31.3	+6.3	−8.4	+8.6
	no post/no QK	+48.2	+48.5	+48.4	+47.7	+47.6
Gemma-3-12B	none	−17.0	+35.3	+34.4	+31.4	+26.6
	no post	+47.6	+48.9	+48.3	+47.2	+46.0
	no QK	−46.9	+36.3	+1.6	+5.2	+8.8
	no post/no QK	+46.1	+48.9	+46.6	+45.7	+40.3
Gemma-3-27B	none	−8.3	+25.5	+7.4	+3.0	+28.0
	no post	+47.5	+49.1	+47.1	+46.8	+42.9
	no QK	−65.7	−27.3	−74.7	−78.1	+8.8
	no post/no QK	+48.1	+49.1	+48.1	+47.3	+47.3

Table S7: Gemma-3-4B initialisation-only norm probes. Values are gaps to F_L in percent, averaged over three random seeds with standard deviations. All jobs use `max_steps=0`; no training is performed. The pretrained Gemma-style default anomaly is absent at random initialisation.

Variant	Default	first pos-1	token-normed	top-1 PC	top-3 PC	max pos-0 ratio
none	+43.7±0.2	+43.9±0.1	+43.7±0.2	+42.8±0.1	+42.2±0.1	1.00±0.00
no post	+47.1±0.1	+43.6±0.1	+47.1±0.1	+46.0±0.1	+44.7±0.2	1.20±0.00
no QK	+43.7±0.2	+43.9±0.0	+43.7±0.2	+42.8±0.1	+42.2±0.1	1.00±0.00
no post/no QK	+47.1±0.1	+43.6±0.1	+47.1±0.1	+46.0±0.1	+44.7±0.2	1.19±0.00

Table S7 shows that the large-checkpoint anomaly is not present simply because the Gemma-3-4B block contains post-branch normalisation and QK normalisation. All initialisation variants are strongly sub-reference. Removing post-normalisation raises the gap further, while removing QK normalisation has essentially no effect at initialisation. This contrasts with the trained-checkpoint intervention in Table S6, where removing QK normalisation amplifies the default-pooling anomaly.

The short-pretraining result should be read as a boundary, not as a negative proof about Gemma. A few hundred million FineWeb-Edu tokens at 270M scale are far from the full training and post-training history of the released Gemma checkpoints. Within this simplified setting, however, the same norm switches are not sufficient to produce the above-reference default behaviour: all variants remain near a +41% sub-reference gap. Post-normalisation removal does create a position-0 norm spike, but the default, token-normalised, and top-PC controlled gaps remain sub-reference. Together with Table S7, this

Table S8: Gemma-3-270M short-pretraining norm probes. Values are final checkpoint gaps to F_L in percent, averaged over three seeds with standard deviations. These runs use a short FineWeb-Edu pretraining budget and are not intended to match the full Gemma training pipeline. None of the variants reproduces the pretrained Gemma anomaly.

Variant	Default	first pos-1	token-normed	top-1 PC	top-3 PC	max pos-0 ratio	val loss
none	+41.3±0.6	+33.0±0.5	+41.0±0.6	+35.7±2.7	+32.6±2.1	1.48±0.04	5.06±0.01
no post	+40.7±0.3	+22.8±1.7	+40.6±0.4	+36.8±1.7	+33.9±3.0	5.50±1.17	5.08±0.01
no QK	+41.7±0.3	+31.3±2.3	+41.6±0.3	+36.2±2.9	+32.9±2.7	1.79±0.18	5.07±0.01
no post/no QK	+40.7±0.3	+26.7±0.9	+40.8±0.3	+36.6±2.0	+34.3±2.8	5.06±1.24	5.09±0.01

suggests that the pretrained Gemma anomaly is not a pure initialisation-level architectural identity. The safer interpretation is that Gemma's architecture modulates the profile of trained checkpoints, while the anomaly itself likely depends on scale, long training history, or later checkpoint-specific dynamics.

Appendix S5. Proof of CKA Convergence (Proposition 4 of the main paper)

Proof Fix layers $\ell, m \in \{1, \dots, L\}$ and let $x_1, \dots, x_n$ be i.i.d. draws from the data distribution. Denote the i-th row of the centered representation matrix at layer ℓ by $\tilde{h}_\ell^{(i)} = h_\ell^{(i)} - \bar{h}_\ell$, where $\bar{h}_\ell = n^{-1} \sum_{j=1}^n h_\ell^{(j)}$.

The empirical linear CKA is

$$\widehat{\mathrm{CKA}}_n(\ell, m) \;=\; \frac{\|\tilde{H}_m^\top \tilde{H}_\ell\|_F^2}{\|\tilde{H}_\ell^\top \tilde{H}_\ell\|_F \cdot \|\tilde{H}_m^\top \tilde{H}_m\|_F}.$$

Step 1: Convergence of Gram matrices. Each entry of $\tilde{H}_\ell^\top \tilde{H}_m / n$ is a sample covariance. By the strong law of large numbers (the assumption $\mathbb{E}[\|h_j\|^4] < \infty$ for all layers j ensures $\mathrm{Var}(h_{\ell,j} h_{m,k}) < \infty$),

$$\left[\frac{\tilde{H}_\ell^\top \tilde{H}_m}{n}\right]_{jk} \xrightarrow{a.s.} \mathrm{Cov}(h_{\ell,j},\, h_{m,k}) \;=:\; [\Sigma_{\ell m}]_{jk}.$$

The sample mean $\bar{h}_\ell \to \mu_\ell$ a.s., so sample centering and population centering agree asymptotically: $\tilde{H}_\ell^\top \tilde{H}_m / n \to \Sigma_{\ell m}$ entrywise.

Step 2: Continuous mapping. Define the population CKA functional:

$$\mathrm{CKA}^*(\ell, m) \;=\; \frac{\|\Sigma_{\ell m}\|_F^2}{\|\Sigma_{\ell\ell}\|_F \cdot \|\Sigma_{mm}\|_F}.$$

The map $(A, B, C) \mapsto \|A\|_F^2 / (\|B\|_F \cdot \|C\|_F)$ is continuous on the domain where $\|B\|_F, \|C\|_F > 0$. By the non-degeneracy assumption, $\|\Sigma_{\ell\ell}\|_F > 0$ and $\|\Sigma_{mm}\|_F > 0$. By the continuous mapping theorem, $\widehat{\mathrm{CKA}}_n(\ell, m) \xrightarrow{p} \mathrm{CKA}^*(\ell, m)$.

Table S9: Sensitivity of D_{eff}/L to maximum lag K. The paper uses full $K = L-1$ throughout. For most models values stabilize by $K = 25$, while for Qwen3.5-27B the gap between $K = 25$ and $K = L-1$ is 0.009.

Model	L	K=5	K=10	K=15	K=25	K=L−1
Pythia-70M	6	0.239	–	–	–	–
Pythia-160M	12	0.145	0.112	–	–	0.110
Pythia-1.4B	24	0.116	0.076	0.064	–	0.057
Qwen3.5-4B	32	0.102	0.060	0.045	0.035	0.034
OLMo-2-7B	32	0.103	0.061	0.047	0.037	0.035
Llama-3.1-8B	32	0.106	0.064	0.052	0.046	0.045
OLMo-2-13B	40	0.100	0.058	0.043	0.032	0.029
Qwen3.5-27B	64	0.103	0.059	0.043	0.030	0.021

Step 3: Lag-averaged consistency. For fixed L and lag k, the estimator $\hat{\varrho}_{\text{CKA}}(k)$ in Eq. (4) of the main paper is a finite average of $L-k$ convergent terms: $\hat{\varrho}_{\text{CKA}}(k) \xrightarrow{p} \varrho^*_{\text{CKA}}(k)$, where

$$\varrho^*_{\text{CKA}}(k) = \frac{1}{L-k} \sum_{\ell=1}^{L-k} \text{CKA}^*(\ell, \ell+k).$$

■

Appendix S6. Maximum Lag Sensitivity

Although the paper uses the full autocorrelation ($K = L-1$) throughout, it is still useful to examine how D_{eff}/L changes as a function of the maximum included lag. Table S9 reports this sensitivity for representative models.

At $K = 5$ all models show $D_{\text{eff}}/L \approx 0.10$, reflecting only short-range similarity. As K increases, D_{eff}/L decreases monotonically because the Bartlett taper includes progressively more high-lag terms (which remain close to 1 for deep models). The values stabilize by $K = 25$ for models with $L \leq 40$; for Qwen3.5-27B ($L = 64$), the difference between $K = 25$ and $K = 63$ is 0.009, reflecting the additional high-lag information available in very deep models. These results clarify why earlier truncated calculations often appeared numerically close for shallower models, while the full-K definition remains the correct paper setting.

Appendix S7. Alternative Aggregation Functions

The weighted full-lag formula (Eq. (5) of the main paper) is one way to aggregate the layer similarity structure into a scalar. To verify that the conclusions do not depend on this

specific choice, we apply three additional aggregations to the full $L \times L$ CKA similarity matrix $C_{ij} = \mathrm{CKA}(H_i, H_j)$.

Participation ratio (Roy and Vetterli, 2020). Given the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_L$ of C:

$$\mathrm{PR} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}.$$

PR equals L when all eigenvalues are equal (all layers independent) and equals 1 when a single eigenvalue dominates (all layers identical). It requires no lag structure and operates on the full matrix spectrum.

Effective rank (Roy and Vetterli, 2007).

$$\mathrm{ER} = \exp\left(-\sum_i p_i \log p_i\right), \quad p_i = \frac{\lambda_i}{\sum_j \lambda_j}.$$

ER is the exponential of the spectral entropy. Like PR, it ranges from 1 to L and captures the effective dimensionality of the similarity matrix.

Threshold counting.

$$\mathrm{TC}_\tau = 1 + \sum_{\ell=1}^{L-1} \mathbf{1}[\mathrm{CKA}(H_\ell, H_{\ell+1}) < \tau].$$

TC counts the number of adjacent-layer transitions where similarity drops below a threshold τ (we use $\tau = 0.95$). This is the simplest possible aggregation and requires no eigendecomposition.

Results. Figure S4 compares all four aggregation methods (Bartlett, PR, ER, $\mathrm{TC}_{0.95}$) across thirteen models with $L \geq 16$. Despite fundamentally different mathematical formulations, all four methods agree on the model ordering: deeper models consistently have lower effective depth under every aggregation. Bartlett, PR, and ER track closely (Spearman $\varrho > 0.98$ pairwise); $\mathrm{TC}_{0.95}$ is noisier but preserves the same qualitative pattern.

Appendix S8. Covariance Decay under Beta-Mixing (Norm Process)

The paper's main diagnostic D_{eff} uses CKA-based layer similarity (Definition 1 of the main paper). For the *norm-based* estimator $\hat{\varrho}^{\mathrm{norm}}$ (Eq. S1), we can bound the autocovariance using the mixing coefficients of the underlying scalar process.

Proposition 2 (Covariance bound for the norm process) *Let $Z_\ell = \|h_\ell\|_2$ for $\ell = 1, \ldots, L$, and assume $\mathrm{Var}(Z_\ell) > 0$ and $\mathbb{E}[|Z_\ell|^{2+\delta}] < \infty$ for some $\delta > 0$. If the scalar process (Z_ℓ) is stationary and beta-mixing with coefficients $\beta(k)$, then for all $k \geq 1$:*

$$|\mathrm{Cov}(Z_1, Z_{1+k})| \;\leq\; C_\delta \; \beta(k)^{\delta/(2+\delta)} \; \left(\mathbb{E}[|Z_1 - \mathbb{E}Z_1|^{2+\delta}]\right)^{2/(2+\delta)},$$

where C_δ is a constant depending only on δ. Dividing by $\mathrm{Var}(Z)$ yields a bound on $|\varrho^{\mathrm{norm}}(k)|$.

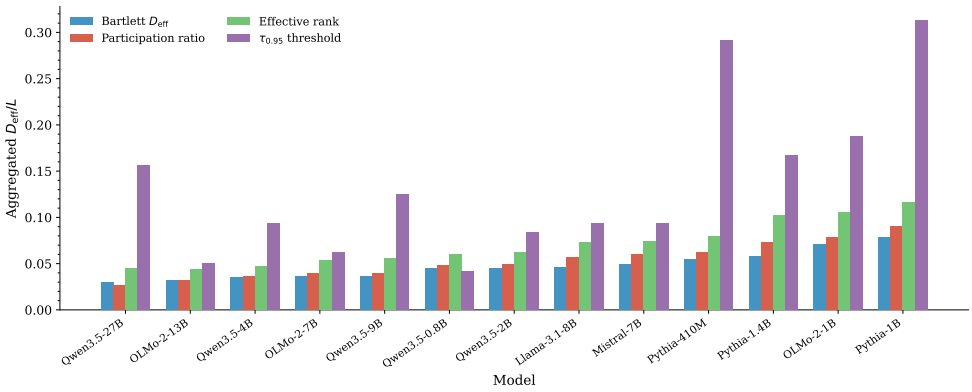


Figure S4: Four independent aggregation methods applied to the same CKA similarity matrices. Despite different mathematical formulations, all methods agree on the model ordering and confirm that deep models ($L \geq 32$) have effective depth below 10% of L.

Proof This is a standard covariance inequality for beta-mixing processes. We apply the result of Bradley (2005, Theorem 3.3): for real-valued random variables $U \in \mathcal{F}_0^t$ and $V \in \mathcal{F}_{t+k}^\infty$ with $\mathbb{E}[|U|^p] < \infty$ and $\mathbb{E}[|V|^q] < \infty$ where $1/p + 1/q < 1$,

$$|\mathrm{Cov}(U, V)| \;\leq\; C_{p,q}\, \beta(k)^{1-1/p-1/q}\, \|U\|_p\, \|V\|_q.$$

Setting $p = q = 2 + \delta$, $U = Z_1 - \mathbb{E}Z_1$, $V = Z_{1+k} - \mathbb{E}Z_{1+k}$ gives $1/p + 1/q = 2/(2+\delta) < 1$ (for $\delta > 0$), hence the exponent on $\beta(k)$ is $1 - 2/(2+\delta) = \delta/(2+\delta)$. ■

Remark 3 *This bound applies to the scalar norm process (Z_ℓ) and the norm-based estimator $\hat{\varrho}^{\text{norm}}$. It does not directly bound the CKA-based $\hat{\varrho}(k)$ used in the paper's primary D_{eff} definition. Establishing an analogous bound for CKA would require a multivariate extension beyond the present scope.*

Appendix S9. Alternative Tapers for D_{eff}

The Bartlett taper $w(k) = 1 - k/L$ in Eq. (5) of the main paper is one of several standard kernel choices for spectral density estimation at zero frequency. Alternatives include:
Parzen taper.

$$w_{\text{Parzen}}(k) = \begin{cases} 1 - 6(k/M)^2 + 6(k/M)^3 & \text{if } 0 \leq k \leq M/2, \\ 2(1 - k/M)^3 & \text{if } M/2 < k \leq M, \\ 0 & \text{if } k > M, \end{cases}$$

where $M \leq L$ is a bandwidth parameter.
Tukey–Hanning taper.

$$w_{\text{TH}}(k) = \tfrac{1}{2}\big(1 + \cos(\pi k/M)\big) \quad \text{for } k \leq M.$$

Table S10: Five-seed random-weight stability for D_{eff}/L (N=10,000, FineWeb-Edu, K=L−1). The narrow random-weight bands show that the null baseline is nearly deterministic for a fixed architecture and initialisation scheme; trained values fall outside the random range in all seven cases.

Model	Trained	Random mean	Random range	mean $\pm 2\sigma$
OLMo-2-1B	0.0702	0.0741	[0.0739, 0.0743]	[0.0737, 0.0745]
OLMo-2-7B	0.0369	0.0361	[0.0360, 0.0362]	[0.0359, 0.0362]
OLMo-2-13B	0.0294	0.0286	[0.0286, 0.0287]	[0.0285, 0.0287]
Qwen3.5-4B	0.0350	0.0469	[0.0468, 0.0470]	[0.0468, 0.0471]
Qwen3.5-9B	0.0369	0.0442	[0.0441, 0.0442]	[0.0441, 0.0443]
Qwen3.5-27B	0.0211	0.0217	[0.0216, 0.0218]	[0.0216, 0.0218]
Llama-3.1-8B	0.0468	0.0347	[0.0345, 0.0348]	[0.0344, 0.0349]

All kernel choices yield D_{eff} values that are qualitatively similar. We use the Bartlett taper throughout because it is a simple full-lag weighting scheme and requires no additional bandwidth parameter.

Appendix S10. Additional Experimental Details

Hardware. All hidden-state extractions run on 8×H100 80GB GPUs (university HPC cluster). Analysis (D_{eff} computation, CKA matrices, figures) runs on CPU.

Hidden-state extraction. Models are loaded in bfloat16. For each input passage of up to 512 tokens, we run a single forward pass with `output_hidden_states=True` and mean-pool each layer's output over valid (non-padding) token positions. The result is a tensor of shape $(n, L+1, d)$ per model-dataset pair, saved as `hidden_states.pt`. We discard layer 0 (embedding) before computing D_{eff}, using layers $1, \ldots, L$ as specified in Definition 1 of the main paper.

CKA computation. Column-centering is applied before computing linear CKA. For numerical stability, we divide the cross-covariance $\tilde{H}_m^\top \tilde{H}_\ell$ by n before computing the Frobenius norm. We use $N = 10{,}000$ samples so that $N > d$ for every evaluated model; the synthetic and unbiased-CKA checks above quantify the remaining finite-sample behaviour.

Norm-based and PCA estimators. The norm-based estimator uses the Frobenius norm $\|H_\ell\|_F/\sqrt{nd}$ at each layer. The PCA estimator projects to the top $\min(64, d, n)$ principal components computed from the concatenation of all layers.

Null baseline: seed stability and figure. For seven architectures we evaluate random-weight D_{eff}/L under five independent default-Hugging-Face initialisations each (N=10,000, K=L−1). Table S10 shows that random-weight D_{eff}/L is seed-stable, with standard deviation at most 2×10^{-4}. Trained D_{eff}/L lies outside the random-seed range and outside the mean $\pm 2\sigma$ band in every case, but the direction depends on architecture: Qwen3.5 and OLMo-2-1B move below their random-weight references, while larger OLMo-2 and Llama-3.1-8B move above them. Figure S5 visualises Table 1 of the main paper.

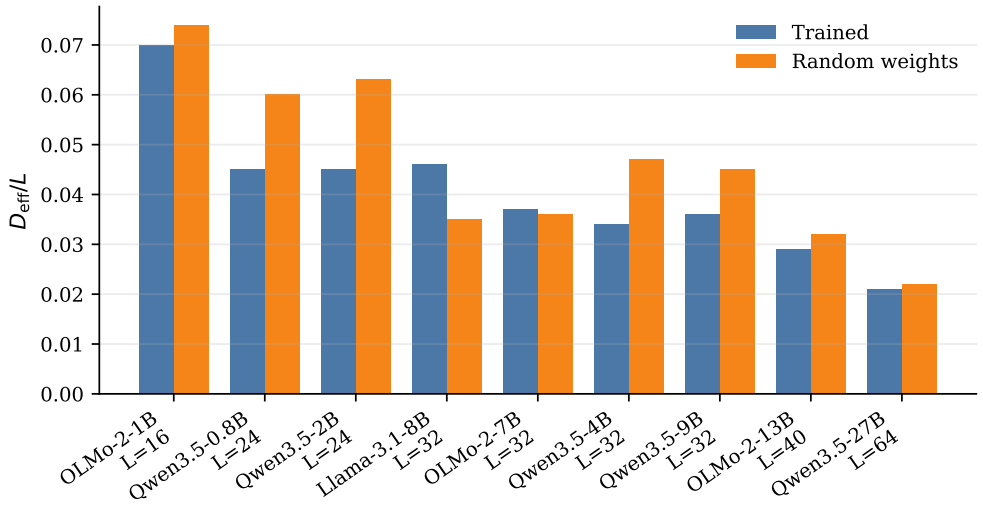


Figure S5: Trained vs. random-weight D_{eff}/L across model families for completed pairs (sorted by depth L). Bar heights remain in the same low range for trained and random, indicating a strong residual-geometric component.

Appendix S11. Geodesic Efficiency vs. D_{eff}/L

We compute the geodesic efficiency $\eta_{\text{geo}} = \|h_L - h_1\|_2 / \sum_{\ell=2}^{L} \|h_\ell - h_{\ell-1}\|_2$ for the same 1,000 input passages and plot it against D_{eff}/L in fig. S6. The scatter plot reveals a positive but moderate correlation, indicating that the two measures capture related but distinct aspects of layer utilisation.

The correlation ($r = 0.65$) is positive but moderate. Geodesic efficiency measures the *directness* of the residual-stream path, while D_{eff} measures the *statistical independence* of adjacent layers. A model can take a straight path (high η_{geo}) through representation space while still producing highly similar intermediate representations (low D_{eff}/L).

Appendix S12. Norm-Based and PCA Estimators

Estimator 1 (Norm-based, scalar). Let $s_\ell = \|H_\ell\|_F / \sqrt{nd}$ be the normalized Frobenius norm at layer ℓ. Define

$$\hat{\varrho}^{\text{norm}}(k) = \frac{\sum_{\ell=1}^{L-k} (s_\ell - \bar{s})(s_{\ell+k} - \bar{s})}{\sum_{\ell=1}^{L} (s_\ell - \bar{s})^2}, \quad \bar{s} = \frac{1}{L} \sum_{\ell=1}^{L} s_\ell. \tag{S1}$$

This is the cheapest estimator (one scalar per layer) but discards directional information.

Estimator 3 (PCA projection). Compute a shared PCA basis $V \in \mathbb{R}^{d \times r}$ from the concatenated hidden states $[H_1; \ldots; H_L]$. Project: $\tilde{H}_\ell = H_\ell V$. Then

$$\hat{\varrho}^{\text{PCA}}(k) = \frac{\text{tr}(C(k))}{\text{tr}(C(0))}, \tag{S2}$$

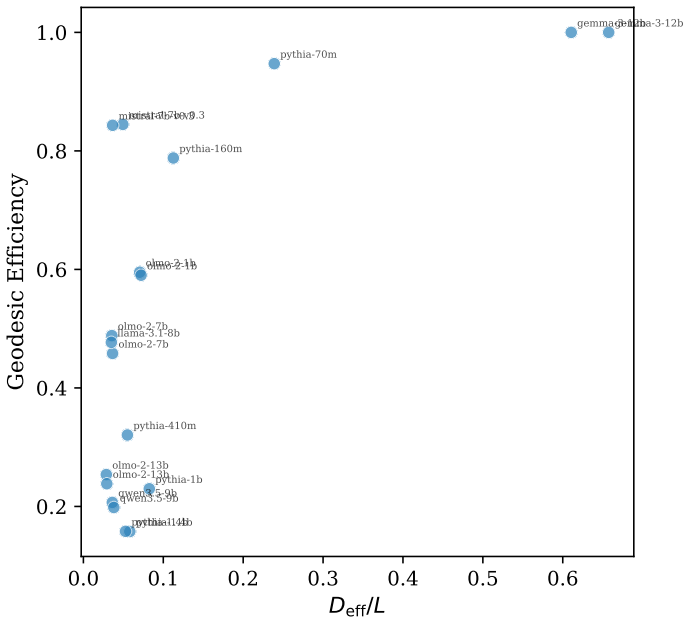


Figure S6: D_{eff}/L vs. geodesic efficiency across all model–dataset pairs. Pearson $r = 0.65$. The moderate correlation confirms that the CKA-based diagnostic and the geometric measure capture related but distinct aspects of layer utilisation.

where $C(k) = \frac{1}{L-k}\sum_{\ell=1}^{L-k} \tilde{H}_\ell^\top \tilde{H}_{\ell+k}$ is the lag-k cross-covariance.

The three estimators target related but distinct population quantities: $\hat{\varrho}^{\text{norm}}$ measures autocorrelation of a scalar summary, $\hat{\varrho}^{\text{CKA}}$ measures average representational similarity, and $\hat{\varrho}^{\text{PCA}}$ measures projected cross-covariance. They agree qualitatively in practice but are not interchangeable in general.

Appendix S13. Multi-Metric Robustness

This auxiliary check uses WikiText-103 rather than the FineWeb-Edu calibration set used in the main results, and is intended only to test whether the similarity metric changes the qualitative ordering. CKA and unbiased HSIC-CKA are nearly identical across all models (maximum difference 0.002), confirming that the choice between standard and unbiased CKA is negligible in the large-sample regime relevant to the paper. Cosine similarity produces slightly higher D_{eff}/L but exhibits the same ordering and the same qualitative conclusion: all 7B+ models have $D_{\text{eff}}/L < 0.055$.

Appendix S14. Instruction-Tuned vs. Base Models

This result is consistent with the null-baseline finding in the main text: if random weights and trained weights already yield similar D_{eff}, it is unsurprising that the relatively mild parameter changes introduced by instruction tuning also leave D_{eff} nearly unchanged.

Table S11: Auxiliary metric-robustness check: D_{eff}/L under three similarity metrics on WikiText-103. CKA and unbiased HSIC-CKA agree within 0.002; cosine similarity is slightly higher but follows the same pattern.

Model	CKA	Unbiased HSIC-CKA	Cosine
Pythia-70M	0.239	0.240	0.277
OLMo-2-1B	0.071	0.071	0.137
Llama-3.1-8B	0.035	0.035	0.051
Mistral-7B	0.036	0.036	0.047
OLMo-2-7B	0.037	0.038	0.054
OLMo-2-13B	0.033	0.033	0.046
Qwen3.5-9B	0.038	0.039	0.041

Table S12: Instruction-tuned vs. base model D_{eff}/L (N=10,000). Differences are negligible ($\leq$ 0.002), indicating that post-training alignment leaves the residual-stream geometry largely unchanged.

Model	L	Instruct	Base	$\lvert\Delta\rvert$
Qwen3.5-0.8B	24	0.045	0.045	<0.001
Qwen3.5-4B	32	0.035	0.035	<0.001
Llama-3.1-8B	32	0.047	0.047	0.002

Appendix S15. ShortGPT Analysis: Shallow Models

ShortGPT's BI score $\text{BI}(\ell) = 1 - \cos(h_\ell, h_{\ell+1})$ captures the lag-1 term of the CKA autocorrelation. Across the evaluated models, Pearson $r > 0.97$ between mean BI and $1 - \hat{\varrho}(1)$, confirming that BI and the lag-1 CKA view are closely related in deep, lag-1-dominated regimes. This is a limitation as well as a useful boundary check: D_{eff} is not meant to replace BI as a local layer score. Its added role is the calibrated model-level aggregation of the full lag profile, together with the F_L reference.

For Pythia-70M ($L = 6$), the Bartlett weights are nearly uniform across all 5 lags, and higher-lag terms contribute substantially to D_{eff}. A lag-1-only diagnostic would assign an incorrect effective depth to Pythia-70M because it ignores the rapid decorrelation at higher lags ($\hat{\varrho}(1) = 0.803$ but $\hat{\varrho}(5) < 0.4$, as the autocorrelation profile decays appreciably across the full 5-lag range). The Bartlett aggregation naturally interpolates between shallow and deep regimes: for large L the taper $(1-k/L)$ suppresses high-lag terms, recovering the lag-1-dominated regime; for small L the taper is nearly flat and all lags are weighted comparably. More broadly, D_{eff} correctly handles the boundary between shallow models (where most layers contribute independently) and deep models (where lag-1 dominates).

Table S13: Accumulated D_{eff}/L vs. update-based $D_{\text{eff}}^{\text{update}}/L$ (N=10,000, FineWeb-Edu). Synthetic orthogonal-update controls show that accumulation alone can keep $D_{\text{eff}}(h)$ low; this table asks whether real-model updates are also correlated. They are.

Model	L	D_{eff}/L	$D_{\text{eff}}^{\text{upd}}/L$	$\hat{\varrho}^{\text{upd}}(1)$	$\|f\|/\|h\|$
Pythia-160M	12	0.112	0.154	0.648	0.54
OLMo-2-1B	16	0.070	0.080	0.846	0.50
Qwen3.5-4B	32	0.034	0.045	0.787	0.31
Llama-3.1-8B	32	0.046	0.045	0.806	0.37
OLMo-2-13B	40	0.029	0.041	0.796	0.29
Qwen3.5-27B	64	0.021	0.033	0.601	0.18

Appendix S16. Per-Layer Update Diagnostic

The accumulated-state D_{eff} on h_ℓ is the primary diagnostic in this paper, but a complementary view computes the same Bartlett aggregation on the per-layer updates $f_\ell = h_\ell - h_{\ell-1}$. This update-level diagnostic, $D_{\text{eff}}^{\text{update}}$, asks whether the updates themselves are correlated rather than only their accumulated sums.

$D_{\text{eff}}^{\text{update}}/L$ is only modestly higher than the accumulated D_{eff}/L and remains below 0.16 for all completed models. Combined with the synthetic orthogonal-update construction (Appendix S2), this shows that both effects matter: residual accumulation compresses the sequence of states seen by downstream layers, and the updates themselves are also correlated. This is consistent with the "iterative inference" view of Jastrzebski et al. (2018) and the "curse of depth" identified by Sun et al. (2025), both of which show that deeper layers produce small, correlated updates. The mean update-to-accumulated norm ratio $\|f_\ell\|/\|h_\ell\|$ decreases in the deeper models, from 0.54 at L=12 to 0.29–0.37 for L=32–40 and 0.18 at L=64, reflecting the growing dominance of the accumulated residual stream.

Appendix S17. Controlled Residual-Carry Intervention

To probe the residual-carry mechanism causally, we train a 12-layer nanoGPT on the `shakespeare_char` corpus with the block equation

$$h_{\ell+1} \;=\; \gamma\, h_\ell + \text{attn}(\text{LN}(h_\ell)), \qquad h_{\ell+1} \;=\; \gamma\, h_{\ell+1} + \text{mlp}(\text{LN}(h_{\ell+1})), \tag{S3}$$

and sweep $\gamma \in \{1.0, 0.75, 0.5, 0.25\}$ over multiple random seeds (n_{layer}=12, n_{embd}=384, n_{head}=6, block_size=256, batch_size=64, max_iters=5000, lr=10^{-3} constant). We track $D_{\text{eff}}(h)$, D_{eff}(update), validation loss, and the update-to-state norm ratio at training checkpoints; full trajectories are in Figure S7.

Table S14: Residual-carry intervention on a 12-layer nanoGPT (`shakespeare_char`, iter 5000). Values are mean across seeds with [min, max] in brackets (n=3 seeds for $\gamma \in \{1.0, 0.75\}$, n=2 for $\gamma \in \{0.5, 0.25\}$). At γ=1.0, training has already begun to overfit by iter 5000 (best validation loss 1.56 at iter 1500).

γ	val. loss	$D_{\text{eff}}(h)/L$	$\hat{\varrho}_h(1)$	$\|f\|/\|h\|$
1.00	3.46 [3.30, 3.58]	0.094 [0.093, 0.095]	0.96	0.32
0.75	1.76 [1.71, 1.86]	0.092 [0.087, 0.098]	0.95	0.82
0.50	3.35 [3.35, 3.35]	0.124 [0.112, 0.137]	0.84	0.82
0.25	3.35 [3.35, 3.35]	0.268 [0.211, 0.325]	0.52	1.25

Result. Two regimes are visible. In the small-perturbation regime (γ=0.75), $D_{\text{eff}}(h)/L$ is statistically indistinguishable from γ=1.0 (mean 0.092 vs 0.094; per-seed ranges $[0.087, 0.098]$ vs $[0.093, 0.095]$ overlap fully) and the model trains slightly better at this iteration budget. In the large-perturbation regime ($\gamma \leq 0.5$), $D_{\text{eff}}(h)/L$ rises substantially (mean 0.124 at γ=0.5 and 0.268 at γ=0.25, vs 0.094 at γ=1.0) in every seed, while validation loss plateaus at the unigram baseline $\sim$ 3.35 in every seed. The same ordering between the trains-vs-fails-to-train regimes is visible from initialisation onward (Figure S7).

Interpretation. The intervention supports two compatible claims in this controlled setting: (i) residual carry is a causal contributor to accumulated-state effective depth, since holding everything else fixed, large reductions in γ shift $D_{\text{eff}}(h)/L$ upward in every seed; and (ii) escaping the low-$D_{\text{eff}}(h)$ regime can conflict with stable optimisation, since the only γ values that substantially raise $D_{\text{eff}}(h)/L$ also break optimisation. At standard γ=1.0, the model achieves a low $D_{\text{eff}}(h)/L$ qualitatively similar to the real-LLM values from the main experiments; mild weakening (γ=0.75) leaves $D_{\text{eff}}(h)/L$ unchanged, and stronger weakening diversifies accumulated states only at a steep optimisation cost. These claims are limited to this char-level nanoGPT setup; broader claims would require additional scales and tasks beyond this controlled experiment.

Appendix S18. Bootstrap Stability of D_{eff}/L

To verify that the D_{eff}/L values reported in Table 2 of the main paper are not artefacts of a particular sample of FineWeb-Edu passages, we re-extract hidden states for five representative architectures and run B=100 passage-level bootstrap resamples (with replacement) from the N=10,000 calibration set, recomputing D_{eff}/L on each resample. Table S15 reports the bootstrap mean and the 95% percentile confidence interval.

Bootstrap CI widths are tight relative to the between-model differences in Table 2 of the main paper, so the model orderings visualised in Figure 1 of the main paper are not affected by sample-level Monte Carlo error. Bootstrap resampling is at the passage level, so these CIs reflect uncertainty over the calibration corpus only; uncertainty over alternative model initialisations is treated separately by the seed-robustness analysis in Section 5 of the main paper.

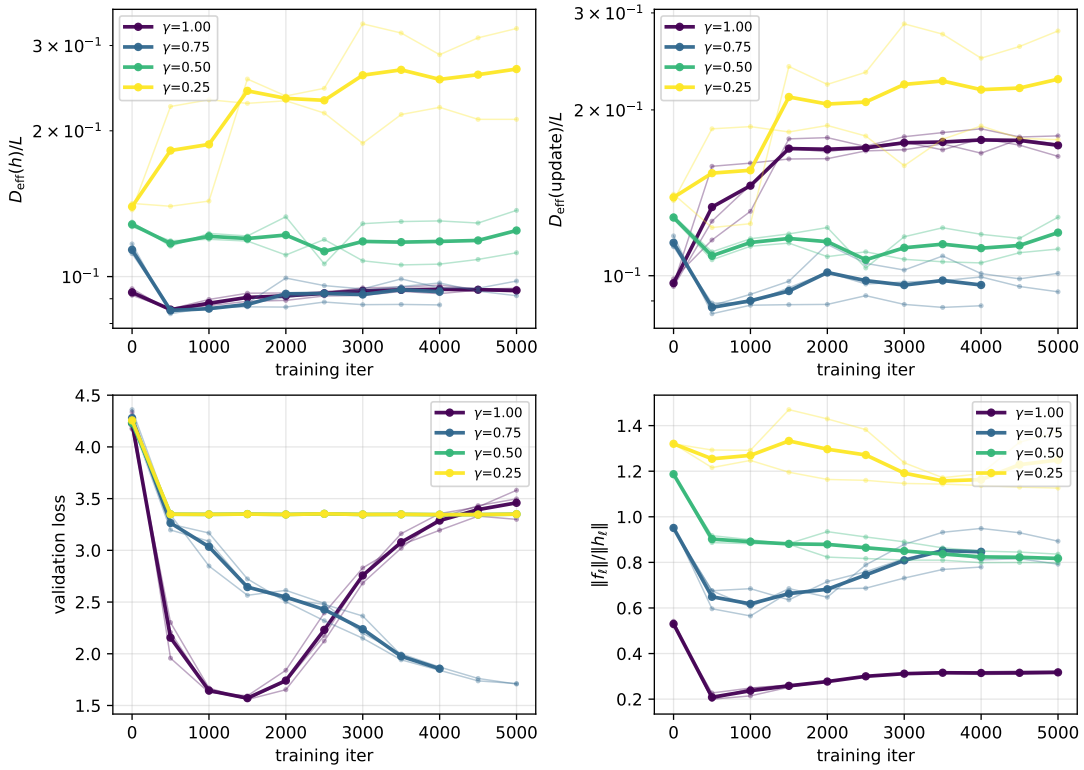


Figure S7: Controlled residual-carry intervention on a 12-layer nanoGPT. Bold lines are the across-seed mean at each γ; faint lines are individual seeds. Top: $D_{\text{eff}}(h)/L$ and $D_{\text{eff}}(\text{update})/L$ trajectories across training, color-coded by γ. Bottom: validation loss and the update-to-state norm ratio. Substantially reducing γ ($\gamma \leq 0.5$) raises $D_{\text{eff}}(h)/L$ from initialisation onward in every seed, but training fails; γ=0.75 leaves $D_{\text{eff}}(h)/L$ statistically indistinguishable from γ=1.0 while training slightly better.

Appendix S19. Pruning Boundary Checks

We do not propose D_{eff} as a pruning method. It is a global accumulated-state diagnostic, whereas layer pruning is a local, task-specific intervention. As a boundary check, we compare a per-layer leave-one-out diagnostic $\Delta_\ell = D_{\text{eff}}(\text{full}) - D_{\text{eff}}(\text{remove } \ell)$ against BI by progressively skipping layers in each ranking and measuring perplexity on a held-out FineWeb-Edu split. Table S16 shows that BI is the correct local tool for this question.

A separate, exploratory question is whether *model-level* D_{eff}/L correlates with how many BI-ranked layers a model can lose before its perplexity degrades substantially. We test this on eight decoder-only models. We frame this as a diagnostic robustness check, not a predictive-validity claim.

Table S15: Passage-bootstrap D_{eff}/L on five representative architectures (N=10,000, B=100). The bootstrap mean matches the point estimate to four decimal places, and the 95% CI width is at most 3×10^{-4} in every case. At the three-decimal precision used in Table 2 of the main paper, the interval collapses to the reported point estimate, so we omit it there.

Model	L	Point D_{eff}/L	Bootstrap mean	95% CI	CI width
Llama-3.1-8B	32	0.0461	0.0461	[0.0459, 0.0462]	0.0003
OLMo-2-7B	32	0.0357	0.0356	[0.0356, 0.0357]	0.0001
Qwen3.5-4B	32	0.0337	0.0337	[0.0336, 0.0337]	0.0001
Qwen3.5-9B	32	0.0358	0.0358	[0.0357, 0.0359]	0.0002
Qwen3.5-27B	64	0.0211	0.0211	[0.0210, 0.0212]	0.0002

Table S16: Boundary check: perplexity after k skipped layers, ranked by BI vs. by local ΔD_{eff}. BI is substantially better at predicting safe removals on both models, consistent with D_{eff} being a global accumulated-state diagnostic rather than a per-layer importance signal. Spearman correlations between the two rankings are $\hat{\varrho}_s = -0.27$ ($p = 0.31$) for OLMo-2-1B and $\hat{\varrho}_s = -0.19$ ($p = 0.30$) for Llama-3.1-8B.

Model	Method	k=0	k=1	k=2	k=4	k=6	k=8
OLMo-2-1B	BI	6.5	9.1	17.9	85.0	2,021	16,091
OLMo-2-1B	Local D_{eff}	6.5	31.0	1,322	87,606	190,276	2,768,762
Llama-3.1-8B	BI	4.6	5.0	5.8	10.0	18.5	56.1
Llama-3.1-8B	Local D_{eff}	4.6	431.1	662.0	34,625	89,183	487,295

For each model we measure perplexity on a held-out FineWeb-Edu split under BI-ranked progressive layer skipping at $k \in \{0, 1, 2, 4, 6, 8\}$ and define

$$\text{tolerance_k}_{1.10}(\text{model}) = \max\big\{\, k \in \{0, 1, 2, 4, 6, 8\} \;:\; \text{PPL}_k \le 1.10 \cdot \text{PPL}_0 \,\big\}.$$

We pre-register the depth-normalized fraction $\text{tolerance_frac}_{1.10} = \text{tolerance_k}_{1.10}/L$ as the primary metric and the raw count as the secondary. Table S17 reports per-model values; the Qwen3.5-27B row is from a BI-only run that omits the local-D_{eff} leave-one-out computation (the full-grid leave-one-out CKA loop did not finish within wall-clock budget at L=64).

Reading. At the full n=8 panel both tolerance metrics correlate negatively with D_{eff}/L: models with lower D_{eff}/L tolerate more BI-pruned layers. The pre-registered primary metric, the depth-normalized fraction $\text{frac}_{1.10}$, passes the strong-correlation threshold ($|r|, |\varrho| \ge 0.7$, $p \le 0.05$) at n=8 but does *not* survive the leave-one-out check that drops OLMo-2-1B ($|r| = 0.59$, $p = 0.16$). The single-point sensitivity reflects a measurement-granularity issue: the BI-pruning grid samples only $k \in \{0, 1, 2, 4, 6, 8\}$, so the integer k values are tightly

Table S17: Per-model BI pruning tolerance (N=10,000 held-out FineWeb-Edu). PPL_0 is the baseline; $k_{1.10}$ is the largest $k \in \{0, 1, 2, 4, 6, 8\}$ at which $\text{PPL}_k \leq 1.10 \cdot \text{PPL}_0$; $f_{1.10} = k_{1.10}/L$.

Model	L	D_{eff}/L	gap (%)	PPL_0	$k_{1.10}$	$f_{1.10}$
OLMo-2-1B	16	0.070	+40.5	6.54	0	0.000
Pythia-1.4B	24	0.059	+26.2	8.11	0	0.000
Mistral-7B	32	0.048	+20.8	3.94	1	0.031
Llama-3.1-8B	32	0.046	+24.1	4.60	1	0.031
OLMo-2-7B	32	0.036	+40.6	5.08	1	0.031
Qwen3.5-9B	32	0.036	+40.6	6.37	1	0.031
Qwen3.5-4B	32	0.034	+43.9	7.17	2	0.062
Qwen3.5-27B	64	0.021	+31.8	6.01	2	0.031

Table S18: Correlations of BI pruning tolerance with D_{eff}/L at the full n=8 panel and three leave-one-out subsets that drop the largest- and smallest-D_{eff}/L endpoints. Pre-registered primary metric: $\text{frac}_{1.10}$. Secondary: raw tolerance_$\text{k}_{1.10}$. All p-values are two-sided.

		$\text{frac}_{1.10}$ vs D_{eff}/L		$k_{1.10}$ vs D_{eff}/L	
Subset	n	Pearson r (p)	Spearman ϱ (p)	Pearson r (p)	Spearman ϱ (p)
Full panel	8	−0.74 (0.04)	−0.79 (0.02)	−0.90 (0.002)	−0.93 (0.001)
Drop OLMo-2-1B	7	−0.59 (0.16)	−0.67 (0.10)	−0.87 (0.01)	−0.91 (0.005)
Drop Qwen3.5-4B	7	−0.82 (0.03)	−0.80 (0.03)	−0.94 (0.002)	−0.91 (0.005)
Drop both	6	−0.68 (0.14)	−0.66 (0.15)	−0.92 (0.009)	−0.86 (0.03)

clustered while L varies 16–64, and dividing by L amplifies the L-dependent noise. The raw secondary metric tolerance_$\text{k}_{1.10}$ is robust to all leave-one-out subsets ($|r| \geq 0.87$, $|\varrho| \geq 0.86$, $p \leq 0.03$ in every case), but because the fraction was the pre-registered primary, we do not promote this to a main-text claim.

We treat the model-level pruning-tolerance check as suggestive rather than conclusive: raw skipped-layer tolerance correlates strongly with D_{eff}/L, but the depth-normalized tolerance fraction is less robust. A finer k-grid (e.g. $k \in \{0, 1, 2, 3, 4, 5, 6, 8\}$) would reduce the granularity-induced noise in the fraction metric and is left as future work. The paper's main pruning claim (Section 5 of the main paper, Table S16) remains that local ΔD_{eff} is a worse per-layer pruning ranker than BI, consistent with D_{eff} being a global accumulated-state diagnostic rather than a per-layer importance score.

Table S19: Capability-evaluation coverage. MMLU, ARC-Challenge, and HellaSwag are available for all 16 models. GSM8K-CoT is evaluated only for models with at least 4B parameters; the Qwen3.5-27B and Gemma-3-12b GSM8K values exhibit suspicious answer-extraction failures and are not used for claims.

Family	Models	MMLU	ARC	HellaSwag	GSM8K-CoT
Pythia	70M, 160M, 410M, 1B, 1.4B	5/5	5/5	5/5	0/5
OLMo-2	1B, 7B, 13B	3/3	3/3	3/3	2/3
Qwen3.5	0.8B, 2B, 4B, 9B, 27B	5/5	5/5	5/5	3/5
Llama	3.1-8B	1/1	1/1	1/1	1/1
Mistral	7B	1/1	1/1	1/1	1/1
Gemma	3-12b	1/1	1/1	1/1	1/1
Total	16 models	16/16	16/16	16/16	8/8 eligible

Other predictors. For completeness, neither the absolute D_{eff} nor the gap to F_L are useful predictors of BI pruning tolerance at any subset (all $p > 0.5$). This is consistent with the role we assign these quantities throughout the paper: D_{eff}/L is the relevant *redundancy* summary, while gap-to-F_L is a regime indicator that does not directly encode tolerance.

Appendix S20. Capability-Relevant Scaling Check

We include a deliberately conservative benchmark comparison because one natural question is whether the global redundancy diagnostic is also a capability predictor. The answer from this small panel is: *not as a main claim.* D_{eff} is not intended to be a capability score, and benchmark comparisons are confounded by model scale, training data, architecture, tokenizer, and evaluation details. We therefore report the results only to delimit the diagnostic's scope: all sixteen models, a modern-only subset excluding Pythia, and the within-family Qwen3.5 panel. OLMo-2 has only three evaluated sizes, so within-family correlations are underpowered and omitted. Even the within-Qwen3.5 correlations should be read as co-movement along a shared scaling axis, not as evidence that D_{eff}/L predicts capability beyond parameter count or depth.

Reading. The capability check is mainly a negative scope result. Within the Qwen3.5 family, lower D_{eff}/L co-moves with higher MMLU and ARC-Challenge performance ($r = -0.896$ and $r = -0.913$), with HellaSwag just below the pre-specified significance threshold. But the same within-family axis also changes parameter count and depth, so the correlation does not establish predictive information beyond scale. Cross-family modern-only correlations are mixed, HellaSwag is partly driven by the Pythia-vs-modern era gap, and GSM8K-CoT is unreliable in this run because several generated-answer extraction scores are anomalous. We therefore do not use capability prediction as evidence for D_{eff}. The main role of D_{eff} remains the one used throughout the paper: diagnosing global residual-stream redundancy and comparing residual-stream operating regimes.

Table S20: Exploratory correlation between benchmark score and D_{eff}/L. Values report Pearson r and Spearman ϱ with two-sided p-values. All-model correlations are informative but confounded by family and era; the within-Qwen3.5 panel is the cleanest within-family scaling check, but remains collinear with scale and depth and is not a predictive-validity claim.

Panel	Task	Pearson r	p	Spearman ϱ	p
All 16	MMLU	-0.640	0.008	-0.931	< 0.001
All 16	ARC	-0.664	0.005	-0.842	< 0.001
All 16	HellaSwag	-0.718	0.002	-0.794	< 0.001
All 16	GSM8K-CoT	-0.291	0.484	-0.349	0.396
Modern only	MMLU	-0.749	0.008	–	–
Modern only	ARC	-0.600	0.051	–	–
Modern only	HellaSwag	-0.426	0.192	–	–
Modern only	GSM8K-CoT	-0.291	0.484	–	–
Qwen3.5	MMLU	-0.896	0.040	-0.872	0.054
Qwen3.5	ARC	-0.913	0.030	-0.872	0.054
Qwen3.5	HellaSwag	-0.864	0.059	-0.872	0.054
Qwen3.5	GSM8K-CoT	$n = 3$, skipped			
OLMo-2	all tasks	$n = 3$, skipped			

Table S21: Leave-one-family-out robustness for all-16 correlations between score and D_{eff}/L. Entries are Pearson r with two-sided p-values in parentheses. MMLU and ARC remain negative under every family deletion; HellaSwag weakens after dropping Pythia, indicating an era/family confound.

Dropped family	MMLU	ARC	HellaSwag	GSM8K-CoT
Gemma	-0.696 (0.004)	-0.730 (0.002)	-0.785 (0.001)	-0.021 (0.964)
Llama	-0.636 (0.011)	-0.663 (0.007)	-0.719 (0.003)	-0.305 (0.505)
Mistral	-0.637 (0.011)	-0.663 (0.007)	-0.724 (0.002)	-0.190 (0.683)
OLMo-2	-0.617 (0.025)	-0.638 (0.019)	-0.708 (0.007)	-0.197 (0.709)
Pythia	-0.749 (0.008)	-0.600 (0.051)	-0.426 (0.192)	–
Qwen3.5	-0.564 (0.071)	-0.606 (0.048)	-0.699 (0.017)	-0.848 (0.070)

Appendix S21. Alignment-free similarity check: a Gromov-Wasserstein pilot

Motivation. CKA measures second-order geometric similarity on the same samples, and its invariances are a known limitation. As an alignment-free cross-check, we compare the layer-pair structure induced by the Gromov-Wasserstein (GW) distance, which first aligns the metric structure of two point clouds by optimal transport and therefore does not assume a shared coordinate system, against the CKA structure used in the paper.

Setup. For each model we subsample 256 evaluation passages (identical indices across layers). For each layer we form the intra-cloud Euclidean distance matrix of the 256 pooled states, normalized by its mean so only relative metric structure matters. For every layer pair we compute the squared-loss GW distance with the exact conditional-gradient solver of the Python Optimal Transport (POT) library. We then correlate the resulting layer-pair GW matrix with the CKA matrix over off-diagonal pairs (embedding row dropped).

Table S22: Gromov-Wasserstein pilot. GW distance is strongly anti-correlated with CKA over layer pairs, and increases with layer distance, so the regime described in the paper is visible under an alignment-free optimal transport lens as well.

Model	$\varrho_s(\mathrm{GW}, \mathrm{CKA})$	$\varrho_s(\lvert i-j\rvert, \mathrm{GW})$	$\varrho_s(\lvert i-j\rvert, \mathrm{CKA})$
Pythia-410M	-0.86	$+0.45$	-0.48
Llama-3.1-8B	-0.68	$+0.54$	-0.89
OLMo-2-7B	-0.47	$+0.44$	-0.82

Reading. The qualitative picture is concordant: where CKA reports high cross-layer similarity, GW reports small metric distortion, and GW grows with layer distance. The quantitative differences confirm that the two views capture different aspects of representational geometry. We nevertheless retain linear CKA for the diagnostic itself because the orthogonal-update reference $F_L = 2L/(L+1)$ exists in closed form for it; to our knowledge no analogous closed-form reference under residual accumulation is available for GW, so a GW-based effective depth would currently be uncalibrated. Extending the calibration theory to transport-based similarities is an interesting direction for future work.

References

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Stanisław Jastrzebski, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio. Residual connections encourage iterative inference. In *International Conference on Learning Representations (ICLR)*, 2018.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pages 606–610, 2007.

Subhaneil Roy and Martin Vetterli. On the effective measure of dimension in the analysis cosparse model. *IEEE Signal Processing Letters*, 27:1045–1049, 2020.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. *arXiv preprint arXiv:2502.05795*, 2025.