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

Abstract

We introduce *effective depth* (D_{eff}), a scalar diagnostic that treats the layer-wise residual stream of a transformer as a discrete-time process, measures how representation similarity decays with layer distance, and aggregates that profile into one number. Across sixteen decoder-only language models, D_{eff} separates a structural consequence of residual accumulation from an empirical one: even maximally diverse orthogonal updates have the closed-form reference $F_L = 2L/(L+1) < 2$, yet fifteen of sixteen default measurements lie below F_L (Qwen3.5: 32–44%, OLMo-2: 40–41%, Pythia: 23–28%). Matched references show that the gap is not caused by the persistent initial state or update-size imbalance, but is largely a calibrated signature of correlated residual updates rather than evidence that depth is unused. Symmetric position-0, token-normalisation, and top-PC controls show the regime is not reducible to BOS or top-PC artefacts: the lone above-reference default outlier joins the same regime, and all sixteen models are sub-reference after token-normalisation or top-1-PC removal. Intermediate checkpoints show that the regime is established early in OLMo-2 and stable through 5T tokens, while Pythia-1.4B follows a distinct decreasing trajectory. A controlled residual-carry intervention supports the mechanism, and D_{eff} is best read as a *global* accumulated-state diagnostic, not as a capability score or pruning method.

Keywords: effective depth; residual stream; transformer interpretability; layer redundancy; centered kernel alignment; large language models

1. Introduction

The depth of a neural network is its most celebrated architectural parameter. Theoretical analyses of deep learning routinely invoke depth as a proxy for expressive power (Telgarsky, 2016; Raghu et al., 2017), and practitioners add layers as a first resort when models underfit. Yet in the transformer era, a growing body of evidence suggests that nominal depth L is a poor surrogate for the number of statistically distinguishable residual states a model actually traverses. Studies of layer pruning (Men et al., 2024; Gromov et al., 2024), representation similarity (Kornblith et al., 2019; Raghu et al., 2021), and skip-layer routing (Elhoushi et al., 2024) collectively demonstrate that large portions of a trained stack can be removed with negligible performance degradation, while Veit et al. (2016) showed that residual networks behave as ensembles of shallow networks whose effective gradient paths are far shorter than the nominal depth. Recent work confirms this pattern in modern LLMs: Sun et al. (2025) show that Pre-LayerNorm causes deeper layers to converge toward identity transformations,

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

and Razzhigaev et al. (2024) find near-perfect linear relationships (Procrustes similarity > 0.99) between adjacent layer representations. This raises a fundamental question:

> *Given a transformer with L layers, how many statistically distinguishable residual states does it traverse, and why?*

Answering this question requires a definition that (i) is grounded in a well-motivated diagnostic framework rather than ad hoc heuristics, (ii) applies uniformly across decoder-only language models, and (iii) connects to interpretable quantities such as layer redundancy.

Our approach. We recast the residual stream $(h_1, \ldots, h_L) \in \mathbb{R}^{L \times d}$ as a *discrete-time process* indexed by layer, measure its similarity autocorrelation $\hat{\varrho}(k)$, and aggregate the full layer-distance profile into a scalar. We measure D_{eff} on the *accumulated* hidden states h_ℓ rather than the per-layer updates $f_\ell = h_\ell - h_{\ell-1}$, because the question we address concerns the distinct information states that downstream computation receives (Section 4):

$$D_{\text{eff}} = \frac{L}{1 + 2\sum_{k=1}^{L-1}\left(1 - \frac{k}{L}\right)\hat{\varrho}(k)}. \tag{1}$$

When layers are independent ($\hat{\varrho}(k) = 0$), $D_{\text{eff}} = L$; when perfectly correlated ($\hat{\varrho}(k) = 1$), $D_{\text{eff}} = 1$. This formulation directly connects the degree of inter-layer redundancy to a single interpretable scalar. The diagnostic has **four concrete uses:** it prevents the mistaken reading that low normalized depth means a model uses only a few layers; gives a global geometry summary for comparing families, checkpoints, and architectural variants; surfaces position-0, norm-spike, and top-PC anomalies when paired with controls; and separates model-level residual-stream behaviour from local layer decisions such as pruning. In this sense, D_{eff} is best read as a calibrated diagnostic for residual-stream geometry: it tells us when a stack follows the familiar residual-accumulation regime and when a new architecture or checkpoint moves off that map.

What this paper does and does not claim. A reader who sees a normalized ratio of $D_{\text{eff}}/L < 0.05$ across 7B+ models might infer that trained transformers "waste" most of their depth. That inference is not what we report. In an idealized residual-accumulation regime with mutually orthogonal per-layer updates, the closed-form value is $D_{\text{eff}} = 2L/(L+1) < 2$ at every L (Section 5, Appendix S2), so any model whose accumulated states approximate this regime will have $D_{\text{eff}}/L = O(1/L)$ by construction. We use this value as an idealized maximally-diverse-update reference, not as a null distribution or a target.

The empirical content is the additional gap below this reference, but this gap should be decomposed rather than overread. We therefore add matched references that retain measured update-size profiles, the persistent initial state, and the measured update similarity profile. These controls show that the sub-F_L gap is not caused by h_0 carry or update-size imbalance; once measured update similarities are retained, the scalar reference is more redundant than the observed accumulated stream. We therefore lead with absolute D_{eff} and gap-to-F_L, but interpret the small ratio as a calibrated signature of correlated residual updates rather than as evidence of learned underuse of layers. CKA also makes D_{eff} insensitive to pure orthogonal rotations of the residual basis, so the diagnostic measures accumulated-state similarity, not every possible form of computation. We restrict empirical

claims to decoder-only autoregressive language models throughout; encoder-only, vision, and multimodal transformers are left to future work.

Contributions.

1. **Effective depth (D_{eff}):** a scalar, architecture-agnostic diagnostic of redundancy in the accumulated residual stream, built from a CKA-based similarity autocorrelation and a weighted full-lag aggregation, both validated independently (Section 4, Appendix S13, Appendix S7).

2. **A calibrated sub-reference operating regime.** Across sixteen decoder-only language models spanning four families, nearly every default measurement shows additional redundancy below the orthogonal-update reference $F_L = 2L/(L+1)$. Matched references show that this gap is not explained by the persistent initial state or measured update-size profile, but is largely accounted for by measured update correlations. The same regime survives position-0, token-normalisation, and top-PC controls applied symmetrically to both outliers and non-outliers; the lone above-reference default outlier joins the regime under every control (Section 5, Figure 1, Figure 2; Appendices S3 and S4).

3. **Training-dynamics arc.** Three OLMo-2 sizes (1B, 7B, 13B) establish a 40–41% gap to F_L by their first landed post-initialisation checkpoints and remain near it through 5T tokens, while Pythia-1.4B follows a distinct monotonically-decreasing trajectory (44% at 2B tokens to 26% at 300B tokens). The sub-reference regime is therefore family-conditioned and early-established rather than solely a function of training progress (Section 5, Figure 3).

4. **Mechanism and global-vs-local distinction.** A controlled residual-carry intervention on a 12-layer nanoGPT shows that mild reductions in residual carry leave $D_{\text{eff}}(h)/L$ unchanged, while the only reductions large enough to raise it substantially also break optimisation (Appendix S17). D_{eff} is a *global* accumulated-state diagnostic, not a capability score or pruning method; Appendix S19 reports pruning boundary checks that keep this scope explicit.

2. Related Work

Layer redundancy and pruning. Men et al. (2024) use Block Influence (BI), the cosine similarity between input and output of each block, to score prunability; Gromov et al. (2024) remove up to 50% of layers with small perplexity loss; Elhoushi et al. (2024), Kim et al. (2024), Raposo et al. (2024), and Fan et al. (2020) leverage the same redundancy via early-exit, structured removal, conditional routing, or stochastic dropout. These works identify redundancy through ablation; D_{eff} formalises it as a single scalar that aggregates the full lag spectrum, of which BI is the lag-1 term.

Depth utilization and residual structure. Sun et al. (2025) identify the "curse of depth" in Pre-LayerNorm architectures, where the Jacobian of deep blocks approaches identity. Razzhigaev et al. (2024) find Procrustes similarity > 0.99 between adjacent layers that drops substantially when the residual component is removed; Csordás et al. (2025) measure update norms in Llama 3.1 and Qwen 3 and find second-half layers contribute

far less than first-half layers; Jastrzebski et al. (2018) distinguish representation learning (early) from iterative inference (later); Lad et al. (2024) argue that decoder-only LMs pass through distinct inference stages; and Veit et al. (2016) show ResNets behave as ensembles of shallow networks. Our null baseline complements these by showing the redundancy is architectural rather than a training pathology, and $D_{\text{eff}}^{\text{update}}$ (Appendix S16) extends the per-layer view into a global scalar.

Mechanistic perspective and massive activations. Elhage et al. (2021) formalise the residual stream as a shared communication channel; Geva et al. (2021, 2022) show feed-forward layers act as key-value memories, and Belrose et al. (2023) show intermediate-layer predictions improve monotonically. Yom Din et al. (2024) use learned linear transformations to expose early-layer predictions, and Skean et al. (2024) study when intermediate-layer representations outperform final-layer features. Queipo-de Llano et al. (2026) show that massive position-0 / BOS activations and within-layer compression valleys are two manifestations of the same mechanism, operating within a single layer and within a single sequence. D_{eff} measures redundancy of the accumulated stream *across depth* and *across samples*; we confirm in Section 5 (Figure 2) that position-0 controls preserve the sub-reference regime rather than explain it away.

Representation similarity and intrinsic dimension. We use CKA (Kornblith et al., 2019) for its orthogonal invariance; alternatives include SVCCA (Raghu et al., 2017), PWCCA (Morcos et al., 2018), and rotation-sensitive shape metrics (Williams et al., 2021). Raghu et al. (2021); Nguyen et al. (2021) apply CKA to vision transformers vs. CNNs and to language models. Our aggregations connect to the intrinsic-dimension literature (Ansuini et al., 2019); the Bartlett aggregation is standard in econometrics (Newey and West, 1987) and MCMC diagnostics (Geyer, 1992) and we use it as a scalar summary of the layer autocorrelation profile, not as an estimator of any specific mixing coefficient.

3. Background

We collect the two ingredients used to define D_{eff}.

Centered Kernel Alignment. For representation matrices $H, H' \in \mathbb{R}^{n \times d}$ with centered columns, the linear CKA similarity (Kornblith et al., 2019) is

$$\text{CKA}(H, H') \;=\; \frac{\|H'^{\top} H\|_F^2}{\|H^{\top} H\|_F \cdot \|H'^{\top} H'\|_F}. \tag{2}$$

$\text{CKA} \in [0, 1]$ is invariant to isotropic scaling and orthogonal transformation, and equals 1 iff $H' = cHQ$ for scalar $c > 0$ and orthogonal Q.

Weighted full-lag aggregation. For a stationary scalar sequence with autocorrelation $\varrho(k) = \text{Corr}(X_t, X_{t+k})$, lag weighting appears in classical long-run-variance and sampling-efficiency estimators (Newey and West, 1987; Geyer, 1992):

$$N_{\text{eff}} \;=\; \frac{N}{1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\varrho(k)}, \tag{3}$$

We borrow only the finite-lag weighting as a scalar summary of the layer-similarity profile. Trained models are not stationary chains, so the strict ESS interpretation does not apply and D_{eff} is not a beta-mixing estimator. Connections to mixing theory under additional norm-process assumptions are deferred to Appendix S8.

4. Method

Effective Depth (D_{eff}). Consider a transformer with L layers; let $H_\ell \in \mathbb{R}^{n\times d}$ denote the matrix of hidden states at layer ℓ over n input sequences. We exclude the embedding layer ($\ell = 0$) and define the *layer-wise process* $(H_1, \ldots, H_L)$. Computing D_{eff} requires a similarity metric between layer representations and an aggregation that converts the resulting profile into a scalar.

Two views of the residual stream. The residual connection $h_\ell = h_{\ell-1} + f_\ell(h_{\ell-1})$ admits two natural objects: the *accumulated representation* h_ℓ and the *layer update* $f_\ell = h_\ell - h_{\ell-1}$. We deliberately measure D_{eff} on h_ℓ because pruning and layer-skipping operate on the accumulated stream, and because the question "how many distinct information states does the network pass through?" concerns h_ℓ. Even maximally diverse updates produce an accumulated-state $D_{\text{eff}} = 2L/(L+1) < 2$ in the idealised construction with zero update cross-covariance (Appendix S2); this is a geometric property of accumulation, not a statement about what each layer computes. A complementary diagnostic on f_ℓ, $D_{\text{eff}}^{\text{update}}$, is reported in Appendix S16.

Definition 1 (Layer similarity autocorrelation) *The* layer similarity autocorrelation at lag k *is*

$$\hat{\varrho}(k) \;=\; \frac{1}{L-k}\sum_{\ell=1}^{L-k} \text{CKA}(H_\ell,\, H_{\ell+k}), \tag{4}$$

with CKA *as in Eq. (2) on column-centered inputs;* $\hat{\varrho}(k) \in [0, 1]$.

CKA is invariant to orthogonal transformation and isotropic scaling; we verify in Appendix S13 that unbiased HSIC-CKA (Song et al., 2012) and cosine yield nearly identical $\hat{\varrho}(k)$. $\hat{\varrho}$ is the analogue of scalar autocorrelation along the layer dimension; we do not claim it estimates any specific mixing coefficient.

Definition 2 (Effective depth) *Given L layer representations and $\hat{\varrho}(k)$,*

$$D_{\text{eff}} \;=\; \frac{L}{1 + 2\sum_{k=1}^{L-1}\left(1 - \frac{k}{L}\right)\hat{\varrho}(k)}. \tag{5}$$

The denominator lies in $[1, L]$, so $D_{\text{eff}} \in [1, L]$ without clipping.

The Bartlett taper $(1-k/L)$ down-weights high-lag terms, so the sum runs over all $L-1$ lags without bandwidth selection. Alternative aggregations, e.g. participation ratio (Roy and Vetterli, 2020) and effective rank (Roy and Vetterli, 2007), give independent summaries of the same data and yield qualitatively identical conclusions (Appendix S7). D_{eff} is a Bartlett-weighted full-lag scalar summary, not a strict sampling-efficiency estimate (trained models are not stationary chains; see Section 5).

Proposition 3 (Orthogonal-update reference) *Let $h_\ell = \sum_{j=1}^{\ell} f_j$ with per-layer updates f_j satisfying common covariance and zero cross-covariance across layers. Then the Bartlett-aggregated effective depth equals*

$$F_L \;:=\; D_{\text{eff}} \;=\; \frac{2L}{L+1} \;<\; 2 \qquad \textit{for all } L \geq 1.$$

We use F_L as an idealized maximally-diverse-update reference throughout Section 5. It is not a null distribution or a target value; the measured gap to F_L quantifies additional accumulated-state redundancy beyond this structural residual-accumulation reference. We further calibrate this gap using matched scalar references that retain measured update-size profiles, h_0 carry, and measured update-similarity profiles (Appendix S3). Proof in Appendix S2.

Proposition 4 (Convergence of CKA-based estimator) *Fix $\ell, m \in \{1, \ldots, L\}$ and feature dimension d. If inputs $x_1, \ldots, x_n$ are i.i.d., the activations satisfy $\mathbb{E}[\|h_j(x_i)\|^4] < \infty$, and the population covariance matrices $\Sigma_{\ell\ell}, \Sigma_{mm}$ are nonzero, then $\hat{\varrho}_{\mathrm{CKA}}(k)$ is a consistent estimator of its population counterpart as $n \to \infty$ (proof in Appendix S5).*

5. Experiments

Models and setup. We evaluate sixteen publicly available decoder-only language models across four architecture families: **Pythia** (Biderman et al., 2023) (70M/160M/410M/1B/1.4B; $L \in \{6, 12, 16, 24, 24\}$, base), **OLMo-2** (Groeneveld et al., 2024) (1B/7B/13B; $L \in \{16, 32, 40\}$, base), **Qwen3.5** (0.8B/2B/4B/9B/27B; $L \in \{24, 24, 32, 32, 64\}$, instruct), and individual base models Llama-3.1-8B (Dubey et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Gemma-3-12b-pt (Gemma Team, 2024) ($L = 32, 32, 48$). All are autoregressive LMs; encoder-only and vision transformers are out of scope. We extract mean-pooled hidden states from N=10,000 FineWeb-Edu (Penedo et al., 2024) passages per model ($N > d$ for every model; finite-sample behaviour is calibrated in Appendices S2 and S13); CKA is computed in float64 with Frobenius normalisation and D_{eff} uses the full autocorrelation ($K = L-1$, no truncation).

Reference value. We compare each model against the *orthogonal-update reference* $F_L = 2L/(L+1) < 2$, the closed-form D_{eff} when per-layer updates have zero pairwise CKA and are accumulated into the residual stream (Appendix S2). This is an idealized maximally-diverse-update reference, not a probabilistic null or a target. The signed gap $\mathrm{gap} = (F_L - D_{\mathrm{eff}})/F_L$ is positive when a model's accumulated states are more similar than this residual-accumulation reference would produce, and negative when they are less so. The evidence below follows the intended use of D_{eff}: establish the regime, calibrate the gap with matched references, test whether it survives pooling and norm controls, trace when it appears during training, probe the residual-carry mechanism, and separate global redundancy from local pruning; we also include a conservative external check against benchmark scaling.

A sub-reference operating regime. Fifteen of the sixteen models in Figure 1 show positive gap below F_L of 20–44%, with gaps clustering by family (Qwen3.5 32–44%, OLMo-2 40–41%, Pythia 23–28%, Llama-3.1-8B and Mistral-7B 21–24%). Within each family the gap is roughly constant in scale; Gemma-3-12b at -17.6% is the lone above-reference default outlier. This pattern suggests that family-level design and training choices, rather than parameter count alone, shape the residual-stream similarity profile. Absolute D_{eff} for 7B+ models stays in $[1.15, 2.30]$ despite nominal depths 32–64; the small D_{eff}/L is a diagnostic signature, not direct evidence of underuse, since residual accumulation alone produces $D_{\mathrm{eff}}/L = O(1/L)$ at every depth (controls below). Passage-bootstrap 95% CIs collapse to the reported point estimate at three-decimal precision (Appendix S18).

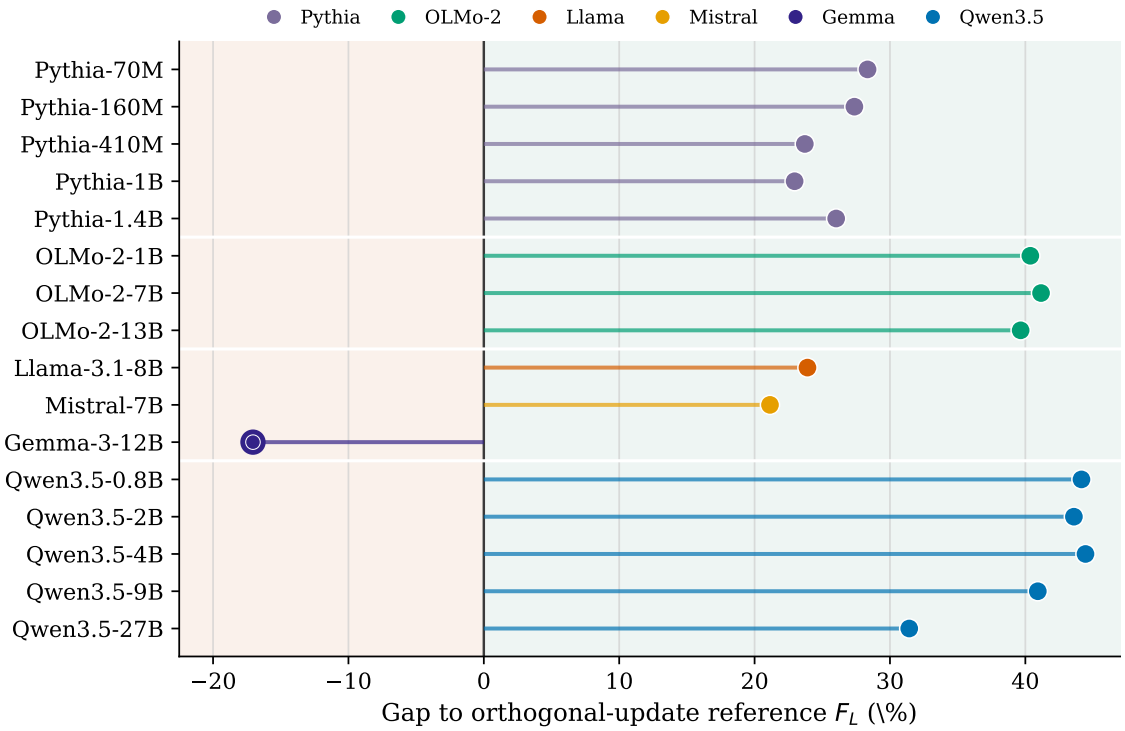


Figure 1: Gap to the orthogonal-update reference $F_L = 2L/(L+1)$ for the sixteen-model sweep (N=10,000, FineWeb-Edu, K=L−1). Positive values indicate $D_{\text{eff}} < F_L$, i.e. accumulated states show additional redundancy beyond the orthogonal-update reference; negative values indicate $D_{\text{eff}} > F_L$. Fifteen of sixteen default measurements are sub-reference, with gaps clustering by model family. Gemma-3-12b is the lone above-reference default outlier; position-0 and norm controls in Figure 2 move it into the sub-reference regime. Full numeric values are reported in Table 2.

Matched references calibrate the gap. The universal F_L reference deliberately removes all model-specific update geometry, so we next ask which model-specific quantities explain the gap. Appendix S3 reports three scalar matched references. Matching the measured update-size profile, with or without retaining the persistent initial state h_0, does not remove the sub-F_L gap: all sixteen models remain below these references. The sign pattern is uniform: gaps to the update-size and h_0-matched references are positive for every row (+17–77%), whereas gaps to the update-correlation-matched reference are negative for every row (−3–−89%). Thus the gap to F_L is not mainly an h_0 or update-norm artifact; it is largely explained by correlated updates. The remaining difference has the opposite sign: real accumulated states are more diverse than a scalar model that preserves only update sizes and average update similarities.

Synthetic constructions verify the diagnostic. On controlled constructions evaluated with the same linear-CKA and weighted aggregation pipeline (Appendix S2), $D_{\text{eff}}/L = 0.258$

for independent states and 0.996 for orthogonal states, so the diagnostic rises when states are genuinely decorrelated. Orthogonal per-layer updates accumulated into states yield the closed-form $D_{\text{eff}} = F_L = 2L/(L+1) < 2$, so even maximally distinct updates produce $D_{\text{eff}}/L = O(1/L)$ in the accumulated stream. Low $D_{\text{eff}}(h)$ is therefore not by itself evidence that specific layers are functionally redundant, and CKA-invisible rotations remain outside the diagnostic's scope.

Position-0 and norm controls: a consistent sub-reference regime. Recent work shows that decoder-only LMs develop massive position-0 / BOS activations and within-layer compression valleys (Queipo-de Llano et al., 2026), which can in principle distort any pooled diagnostic. We re-evaluate D_{eff} on a six-model representative set under four position/norm *control variants*, applying the same controls to default outliers and non-outliers: pooling from position 1 only ("first non-BOS"), ℓ_2-normalising each token before pooling ("token-normed"), and projecting out the top-1 or top-3 cross-sample principal components per layer before CKA. Figure 2 reports the resulting gap to F_L for six representative models; the exact values and full sixteen-model control table are given in Appendix S4.

The pattern is consistent but not invariant. Token-normalisation and top-PC removal leave OLMo-2 and Qwen3.5 sub-reference, while single-position pooling can move Qwen3.5 substantially; this is why we treat the controls as sensitivity analyses rather than as an alternate canonical measurement. Llama-3.1-8B is sub-reference under default (+24%) but its tokenizer auto-prepends BOS and its position-0 norm peaks at 310× at layer 2; once BOS is excluded its gap rises to +41%, the same regime as OLMo-2 and Qwen3.5. The single above-reference model in the default measurement, Gemma-3-12b (11× position-0 norm at layer 7), shifts under *every* control variant from −17% to gaps in $[+27\%, +42\%]$; token normalisation alone is sufficient. Position-0 norm spikes therefore modulate the measured gap, sometimes strongly, but they do not explain the regime away: every model in the panel is sub-reference after first-non-BOS, token-normalisation, or top-PC controls. We retain default pooling as the canonical measurement and use the controls to separate the family-level regime from massive-activation geometry. Across the full sixteen-model sweep, all models are sub-reference after token-normalisation and after top-1-PC removal; first-non-BOS pooling is also sub-reference except for the shallow Pythia-70M boundary case (Appendix S4). Additional pretrained Gemma-3 sizes show that this behaviour is family-level: Gemma-3-4B is near-reference and Gemma-3-27B above-reference by default, but controls move both sub-reference; Gemma-scale interventions show that post-normalisation removal collapses the stream while QK-normalisation removal amplifies the default-pooling anomaly. Initialisation-only and short-pretraining probes do not reproduce the anomaly, so we treat these norm effects as modulators of trained checkpoints rather than as a complete architectural explanation (Appendix S4).

Training dynamics. To test whether the sub-reference regime is a late-training endpoint or an early property of the residual stream, we evaluate D_{eff} across intermediate OLMo-2 checkpoints and a Pythia-1.4B trajectory (Figure 3). Across three OLMo-2 sizes the gap to F_L is established early and stays near it: a $\sim 40\%$ gap by the first landed post-initialisation OLMo-2-1B checkpoint (21B tokens), 40–41% over 412B–3.5T tokens for OLMo-2-7B, and 40–43% from initialisation through 5T tokens for OLMo-2-13B. Pythia-1.4B follows a different trajectory, decreasing monotonically from $\sim 44\%$ at 2B tokens to 26% at 300B tokens. The regime is therefore not solely a function of training progress but a stable family-level

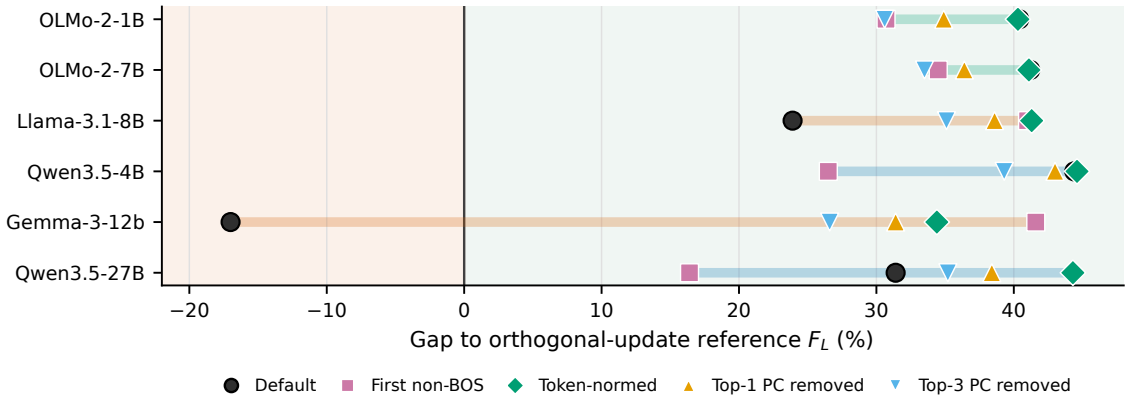


Figure 2: Position-0 and norm control variants for a six-model representative set. Points show gap to F_L in percent; the vertical line marks the reference. "First non-BOS" uses position 1 only, "token-normed" normalises each token before pooling, and top-PC controls remove the leading cross-sample components before CKA. The controls can move the measured gap substantially, especially for Gemma-3-12B and BOS-sensitive Llama/Qwen cases, but all controlled variants in this panel are sub-reference. Exact values and the full sixteen-model table are in Appendix S4.

operating point in OLMo-2, with older architectures following a distinct path; we do not attribute the contrast to any single architectural choice without the controlled probe in Appendix S17.

Null baseline: a strong residual-geometric component. We compute D_{eff} for both trained and **random-weight** versions of the same architectures (Table 1, Appendix S10). Trained and random-weight D_{eff}/L remain in the same low range (maximum absolute difference 0.018), indicating a strong residual-geometric component. A five-seed seed-stability check on seven architectures shows that random-weight D_{eff}/L is seed-stable with standard deviation at most 2×10^{-4}, defining a near-deterministic architectural reference; trained values lie outside this random-seed range in all seven cases, with the direction architecture-dependent (trained values are below random for Qwen3.5 and OLMo-2-1B, but above random for larger OLMo-2 and Llama-3.1-8B; full numbers in Appendix S10). The architecture and initialisation largely determine the order of magnitude of accumulated-state effective depth; training perturbs D_{eff}/L modestly without changing this scale, so D_{eff} is primarily sensitive to residual-stream geometry rather than to the full content of learned computations.

Capability-relevant scaling within Qwen3.5. To test whether D_{eff} is more than an internal measurement, we compare it with external benchmarks. Within Qwen3.5, lower D_{eff}/L co-moves with higher MMLU ($r = -0.896$, $p = 0.040$) and ARC-Challenge ($r = -0.913$, $p = 0.030$); HellaSwag is borderline. Thus D_{eff} tracks a capability-relevant scaling trajectory in this family, although the same axis also changes scale and depth. We therefore do not treat D_{eff} as a capability score, and cross-family panels remain mixed (Appendix S20).

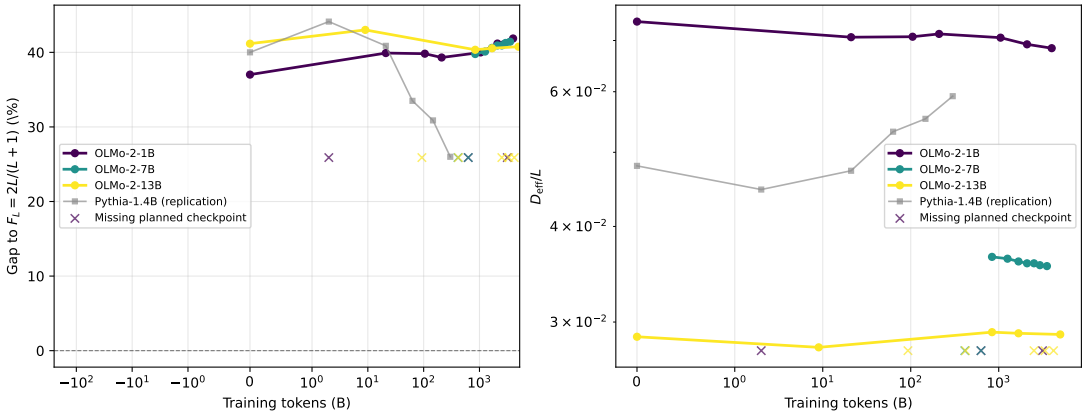


Figure 3: Effective-depth training dynamics. Left: gap to $F_L = 2L/(L+1)$ as a percentage, vs. training tokens. Right: D_{eff}/L vs. training tokens. Three OLMo-2 sizes (1B, 7B, 13B) converge to a stable 40–41% gap by the earliest landed post-initialisation checkpoints and persist through 5T tokens; Pythia-1.4B shows a distinct monotonic decrease. The figure plots all landed checkpoints (25 of 35 planned); missing OLMo-2 points correspond to unavailable or failed revisions and are not interpolated.

Table 1: Multi-family null baseline (N=10,000, FineWeb-Edu). Trained and random-weight D_{eff}/L remain in the same low range across model families for completed pairs (maximum absolute difference 0.018). For OLMo-2-1B/7B/13B, Qwen3.5-4B/9B/27B, and Llama-3.1-8B, the random column reports the mean of five independent seeds; for the other rows it is a single architecture-default initialisation.

Model	L	Trained	Random	$\lvert\Delta\rvert$
OLMo-2-1B	16	0.070	0.074	0.004
Qwen3.5-0.8B	24	0.045	0.060	0.016
Qwen3.5-2B	24	0.045	0.063	0.018
Qwen3.5-4B	32	0.034	0.047	0.013
Llama-3.1-8B	32	0.046	0.035	0.011
OLMo-2-7B	32	0.037	0.036	0.001
Qwen3.5-9B	32	0.036	0.045	0.009
OLMo-2-13B	40	0.029	0.029	0.001
Qwen3.5-27B	64	0.021	0.022	0.001

Residual-carry mechanism. The natural mechanism is the residual skip itself: when each layer adds an update of small relative norm, $h_\ell \approx h_{\ell-1}$ regardless of what the layer computes. Two complementary supports agree. First, the per-layer-update diagnostic $D_{\text{eff}}^{\text{update}}/L$ stays in the same low band as D_{eff}/L across all evaluated models (Appendix S16,

Table S13), so the updates themselves are correlated. Second, a controlled γ-sweep on a 12-layer nanoGPT (Appendix S17) shows that mild reductions in residual carry leave $D_{\text{eff}}(h)/L$ unchanged, while the only reductions that substantially raise it also break optimisation. These results support residual carry as a causal contributor and suggest that escaping the low-D_{eff} regime is tied to stable optimisation in this controlled setting.

D_{eff} vs. BI: global summary vs. local adjacency score. D_{eff} aggregates full-lag accumulated-state similarity via the Bartlett taper; the closest single-lag local diagnostic is the Block-Influence (BI) score (Men et al., 2024), exactly 1 minus the lag-1 cosine similarity between adjacent layers. The two views agree numerically on many deep, lag-1-dominated models, which is expected; Appendix S15 shows that the full-lag aggregation matters in shallow boundary cases such as Pythia-70M, where higher-lag decay would be missed by lag-1 alone. Thus D_{eff} is not a replacement for BI: BI scores individual layers, while D_{eff} is a calibrated full-profile model summary with the F_L reference. We therefore do not present D_{eff} as a pruning method. For completeness, Appendix S19 reports boundary checks showing that local ΔD_{eff} is a poor layer-pruning ranker relative to BI, and that an exploratory model-level pruning-tolerance correlation is not robust enough to support a main claim.

Robustness and within-family. Alternative similarity metrics, aggregations, lag-truncation choices, and tapers all preserve the qualitative ordering (Appendices S13, S7, S6, S9). Appendix S11 gives a path-length companion: geodesic efficiency correlates only moderately with D_{eff}/L ($r = 0.65$), so it is related to but not redundant with the CKA-based summary. Within Qwen3.5, absolute D_{eff} changes only fractionally with scale (1.08–1.15 over the 24/32-layer models, 1.34 at 64 layers), while Llama-3.1-8B and Mistral-7B sit closer to F_L than Qwen3.5 or OLMo-2 at comparable depth; cross-family comparisons should therefore use gap-to-reference rather than D_{eff}/L alone.

6. Discussion

What D_{eff} measures, and what the regime tells us. D_{eff} is a *global* summary of redundancy in the accumulated residual stream: it rises when states are genuinely decorrelated (0.996 on orthogonal, 0.258 on independent) and falls toward F_L when accumulated states approximate maximally diverse residual accumulation. Because $F_L = 2L/(L+1)$ is itself structural, the empirical claim is the additional measured gap and its explanation. Matched references show that the gap survives update-size and h_0 matching but is accounted for by measured update similarity, exposing a correlated-update regime rather than a hidden count of unused layers. Position-0, token-normalisation, top-PC controls, and training dynamics then show that this regime is neither a pooling artifact nor solely a function of training progress; in the evaluated models it behaves like a family-conditioned operating point. The practical value is therefore calibration: D_{eff} tells us when low normalized depth is expected from residual accumulation, when a family sits unusually far below or above that reference, and when a control reveals a position-0 or norm-driven distortion rather than a change in layer computation.

Mechanism: residual carry, interpreted once. A natural mechanism is the residual skip itself: when each layer adds a small relative-norm update, $h_\ell \approx h_{\ell-1}$ regardless of the layer computation. The update-level diagnostic remains in the same low band as D_{eff}/L

(Appendix S16), and the 12-layer nanoGPT γ-sweep shows that only large residual-carry reductions escape the low-D_{eff} regime, while also breaking optimisation (Appendix S17). Residual carry is not the only mechanism: pre-norm, norm-growth, and Procrustes effects (Sun et al., 2025; Razzhigaev et al., 2024) also compound with residual accumulation, and the OLMo-2-vs-Pythia contrast still needs controlled architectural ablation. This is why we frame the result as a geometry of the accumulated stream rather than as a claim that deeper layers do not compute: updates can be meaningful while the states received by downstream layers remain highly correlated.

Global vs. local: D_{eff}, BI, and prunable fraction. $D_{\text{eff}}/L \approx 0.03$–$0.05$ for 7B+ models, yet pruning studies report 25–50% removable (Men et al., 2024; Gromov et al., 2024). The apparent gap is real: D_{eff} counts representational diversity, while pruning measures task-specific functional tolerance. We therefore treat pruning only as a boundary check (Appendix S19): D_{eff} summarises the model, while BI scores local layer adjacency. This division is useful in practice. A new checkpoint, architecture, or compressed model can be compared globally by its gap-to-reference and control variants; local layer removal should still use local influence or ablation scores. Expecting one scalar to answer both questions is precisely the interpretation error that the calibrated reference is meant to prevent.

Limitations and future work. D_{eff} is a corpus-based measurement, not an architectural prediction; CKA is blind to pure residual-stream rotations, so Procrustes or shape metrics (Williams et al., 2021) are natural complements; and trained models are not stationary chains, so D_{eff} is not a literal ESS or beta-mixing estimator (Appendix S8). Empirical claims are restricted to decoder-only LMs on FineWeb-Edu. The next step is a controlled grid varying residual update scale, initialisation, normalisation, and architectures that weaken the unit identity path. Such experiments would turn the family-conditioned patterns observed here into causal architectural statements, and would clarify whether designs such as mixture-of-depths, residual scaling, or parallel depth can deliberately move models away from the standard residual-accumulation regime.

7. Conclusion

Effective depth (D_{eff}) is a global diagnostic of accumulated residual-state redundancy with an explicit structural baseline: orthogonal updates already give $F_L = 2L/(L+1) < 2$. Across sixteen decoder-only LMs, the calibrated gap is mostly sub-reference, is tied by matched references to correlated updates, and survives position-0, token-normalisation, and top-PC controls. Together with the training dynamics, random-weight baselines, and residual-carry probe, this makes D_{eff} a model-level diagnostic of residual-stream geometry rather than a claim that only a few layers matter. Its main use is to give researchers a calibrated map: where a model family sits relative to the residual-accumulation reference, which anomalies are caused by position-0 or norm geometry, and when global state redundancy should be separated from local pruning decisions. In that map, D_{eff} is the global coordinate and BI-like scores are the local companion.

Acknowledgments

We thank the anonymous ACML 2026 reviewers and area chairs for constructive feedback that improved the paper. Computational resources were provided by the Institute of Science and Technology Austria (ISTA) and the Technion Israel Institute of Technology.

References

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

Róbert Csordás, Christopher D. Manning, and Christopher Potts. Do language models use their depth efficiently? *arXiv preprint arXiv:2505.13898*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Kelly Ma, and Carole-Jean Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. *arXiv preprint arXiv:2404.16710*, 2024.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations (ICLR)*, 2020.

Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of*

the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022.

Charles J. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483, 1992.

Dirk Groeneveld, Iz Magnusson, Shayne Han, Yizhong Kim, Kyle Liu, Jiacheng Liu, Kyle Lo, Noah A. Morrison, Niklas Naik, Crystal Nguyen, Jedidiah Ophoff, Chase Reinhart, Roy Schwartz, Saurabh Shen, Emma Sivirsky, Emma Strubell, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, and Noah A. Smith. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Jiawei Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Bo-Kyeong Kim, Geonmin Kim, Tae-Ho Kim, Thibault Castells, Shinkook Shin, Junho Kang, and Simyung Choi. Shortened LLaMA: A simple depth pruning for large language models. *arXiv preprint arXiv:2402.02834*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. The remarkable robustness of LLMs: Stages of inference? *arXiv preprint arXiv:2406.19384*, 2024.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations (ICLR)*, 2021.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Thomas Wolf, Leandro Von Werra, Quentin Lhoest, and Julien Launay. FineWeb: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=c5TFhCJ6fs.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854. PMLR, 2017.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 12116–12128, 2021.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Gerasimenko, Ivan Oseledets, Denis Kuznetsov, and Nikolay Dimitrov. Your transformer is secretly linear. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

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.

Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. In *Journal of Machine Learning Research*, volume 13, pages 1393–1434, 2012.

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.

Matus Telgarsky. Benefits of depth in neural networks. In *29th Annual Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539. PMLR, 2016.

Andreas Veit, Michael J. Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations. In *Proceedings of LREC-COLING*, 2024.

Table 2: Effective depth (N=10,000, FineWeb-Edu, K=L−1). Absolute D_{eff} remains $O(1)$ across scales. All models except Gemma-3-12b sit *below* the orthogonal-update reference $F_L = 2L/(L+1)$ (positive gap), with the gap clustering by family; Gemma-3-12b is the lone above-reference default outlier.

Model	L	d	D_{eff}	D_{eff}/L	gap (%)	$\hat{\varrho}(1)$
Pythia-70M	6	512	1.23	0.205	+28.2	0.862
Pythia-160M	12	768	1.34	0.112	+27.2	0.856
Pythia-410M	24	1024	1.46	0.061	+23.8	0.940
Pythia-1B	16	2048	1.46	0.091	+22.7	0.914
Pythia-1.4B	24	2048	1.42	0.059	+26.2	0.954
OLMo-2-1B	16	2048	1.12	0.070	+40.5	0.972
OLMo-2-7B	32	4096	1.15	0.036	+40.6	0.988
OLMo-2-13B	40	5120	1.18	0.029	+39.7	0.988
Llama-3.1-8B	32	4096	1.47	0.046	+24.1	0.964
Mistral-7B	32	4096	1.54	0.048	+20.8	0.973
Gemma-3-12b	48	3840	2.30	0.048	−17.6	0.910
Qwen3.5-0.8B	24	1024	1.08	0.045	+43.7	0.986
Qwen3.5-2B	24	2048	1.08	0.045	+43.7	0.984
Qwen3.5-4B	32	2560	1.09	0.034	+43.9	0.985
Qwen3.5-9B	32	4096	1.15	0.036	+40.6	0.974
Qwen3.5-27B	64	5120	1.34	0.021	+31.8	0.973

Appendix A. Full Effective-Depth Measurements

Table 2 reports the numeric values underlying Figure 1.

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

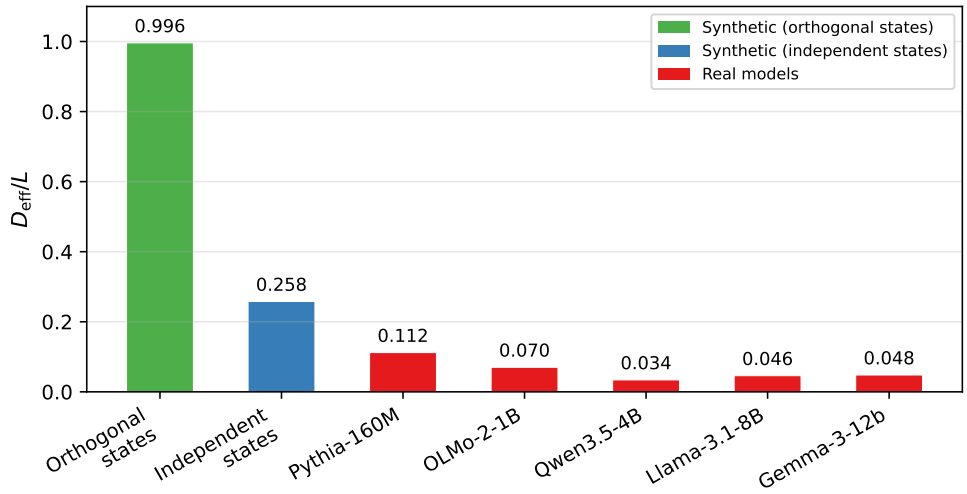


Figure S1: D_{eff}/L for synthetic constructions (green, blue) and real models (red). Orthogonal states recover $D_{\text{eff}}/L = 0.996$; independent states give 0.258; real models fall well below both, confirming that the low values reflect genuine redundancy in the accumulated residual stream.

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 $i < j$,

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

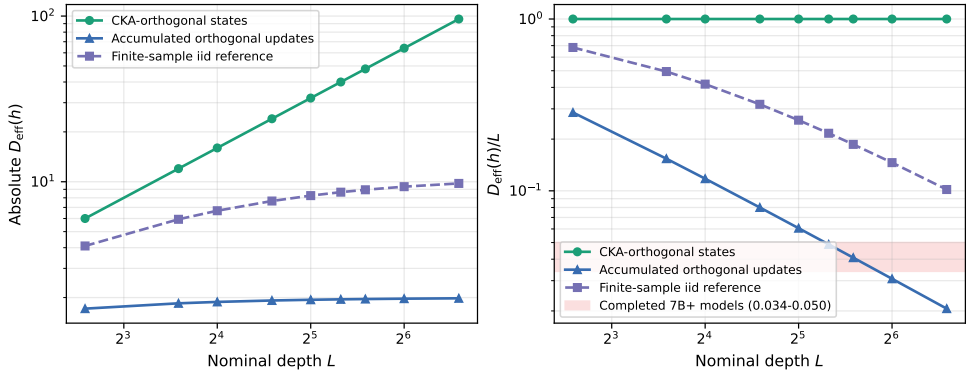


Figure S2: Depth sweep for synthetic residual constructions. Orthogonal states recover $D_{\text{eff}} = L$, whereas accumulated orthogonal updates obey $D_{\text{eff}} = 2L/(L+1) < 2$ and therefore have $D_{\text{eff}}/L = O(1/L)$. The finite-sample iid reference matches the positive CKA bias observed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

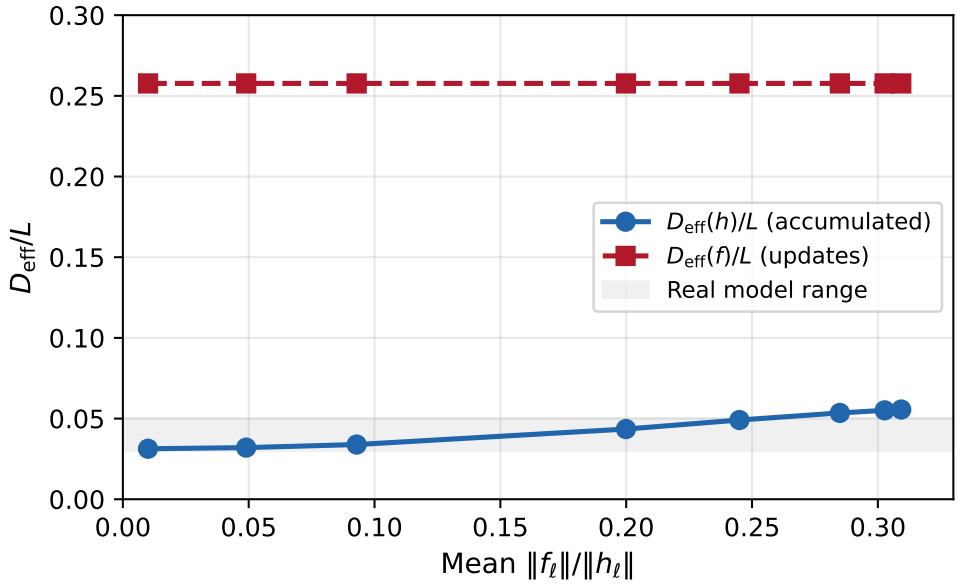


Figure S3: Orthogonal-update sweep (N=10,000, L=32, d=1024). $D_{\text{eff}}(f)/L$ (updates, red) remains constant at 0.258 regardless of scale, while $D_{\text{eff}}(h)/L$ (accumulated, blue) stays in the real-model range (gray band). Diverse updates do not imply diverse accumulated states.

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 $|\varrho^{\mathrm{norm}}(k)|$.

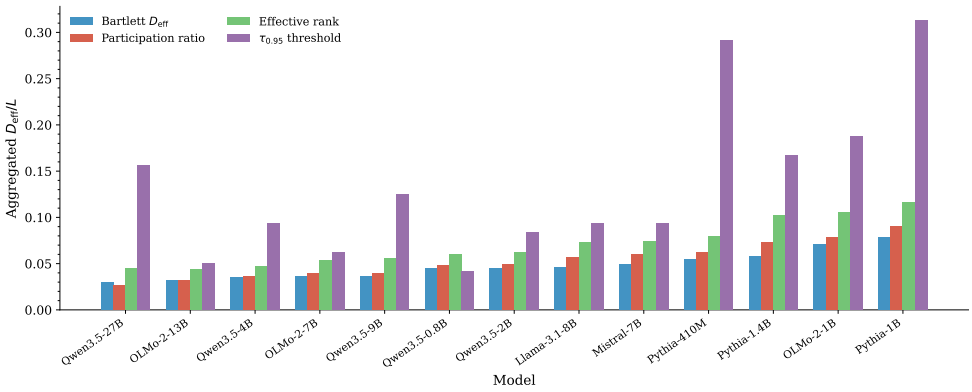


Figure S4: Four independent aggregation methods applied to the same CKA similarity matrices. Despite different mathematical formulations, all methods agree on the model ordering and confirm that deep models ($L \geq 32$) have effective depth below 10% of L.

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.

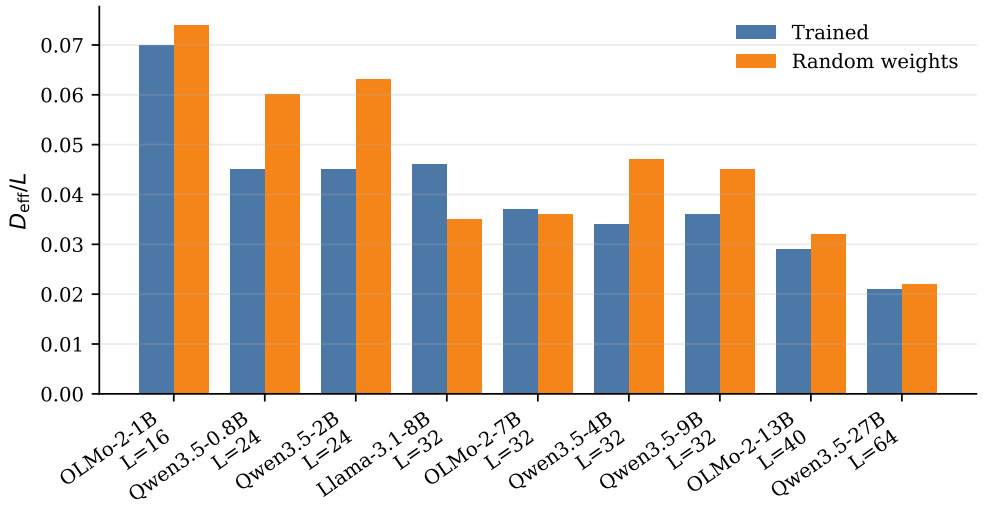


Figure S5: Trained vs. random-weight D_{eff}/L across model families for completed pairs (sorted by depth L). Bar heights remain in the same low range for trained and random, indicating a strong residual-geometric component.

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

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

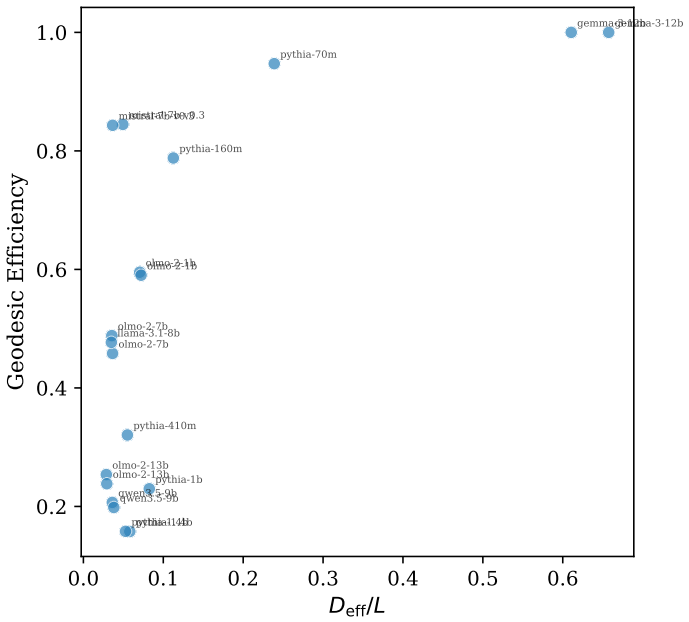


Figure S6: D_{eff}/L vs. geodesic efficiency across all model–dataset pairs. Pearson $r = 0.65$. The moderate correlation confirms that the CKA-based diagnostic and the geometric measure capture related but distinct aspects of layer utilisation.

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.

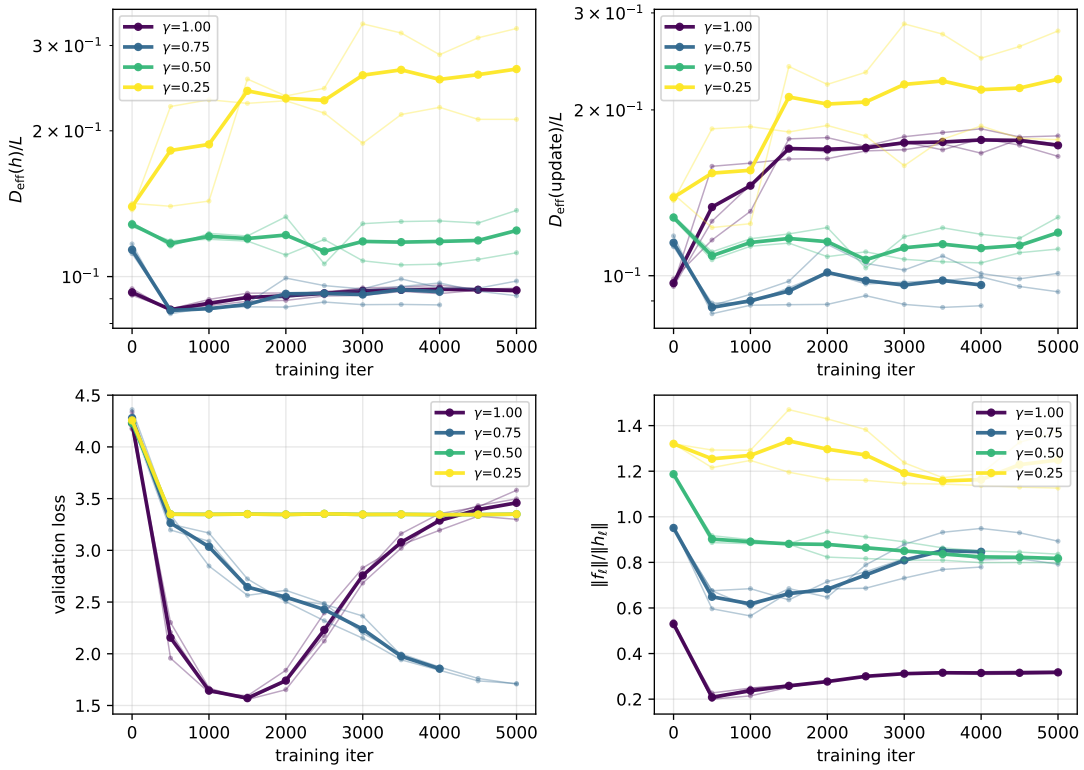


Figure S7: Controlled residual-carry intervention on a 12-layer nanoGPT. Bold lines are the across-seed mean at each γ; faint lines are individual seeds. Top: $D_{\text{eff}}(h)/L$ and $D_{\text{eff}}(\text{update})/L$ trajectories across training, color-coded by γ. Bottom: validation loss and the update-to-state 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 $\mathrm{PPL}_k \leq 1.10 \cdot \mathrm{PPL}_0$; $f_{1.10} = k_{1.10}/L$.

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: $\mathrm{frac}_{1.10}$. Secondary: raw tolerance_$\mathrm{k}_{1.10}$. All p-values are two-sided.

		$\mathrm{frac}_{1.10}$ 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_$\mathrm{k}_{1.10}$ 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.