PERMUTATION ALIGNMENT AND DEPTH IN WEIGHT-TIED LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Averaging the weights of two independently trained language models can yield a far worse model, even after their hidden units are permuted into correspondence. We ask how much of this damage permutation alignment removes, and how to interpret the remaining damage in weight-tied (looped) models, whose depth can vary at runtime. We measure fifteen pairs of 35M-parameter looped models at 16 and 192 loops, before and after a permutation alignment that respects the shared weights, recording the quality of the averaged (midpoint) model and both endpoints. Before alignment, the median barrier (the ratio of midpoint to better-endpoint perplexity) is 1.8 dex (base-10 log units) higher for eight pairs trained mostly with random loop counts than for six trained at a fixed 16 loops; after alignment the gap is under 0.05 dex. A shared residual remains: with the best of four alignment methods, midpoint perplexity is at least 12 times the better endpoint's on every pair. Equal-norm random perturbations reproduce its typical size; perturbations along the aligned weight difference's leading directions cause more damage than orthogonal ones on every pair tested. At 192 loops, endpoint quality sets the barrier: it falls where endpoints become unusable and persists between depth-stable ones. Continuing training from the aligned midpoint for 30% of the original budget removes over 90% of its residual on two depth-stable pairs, and distilling a probability-mixture teacher keeps deep-runtime degradation below 0.05 dex, versus 0.40 dex for plain continuation. Relative barriers should thus be reported with absolute endpoint quality and runtime depth.

1 INTRODUCTION

Averaging weights is a basic tool for combining neural networks: trajectory averaging, model soups, and task-vector merging all build a new model from the parameters of existing ones (Izmailov et al., 2018; Wortsman et al., 2022; Ilharco et al., 2022). Yet two language models can each make useful predictions and still fail when their weights are averaged (Entezari et al., 2021; Ainsworth et al., 2022). A large loss at the midpoint of the linear path between two models is often described as a barrier between basins (Garipov et al., 2018; Draxler et al., 2018), but that description combines three questions. How much of the barrier comes from misaligned hidden coordinates? How much of that misalignment does a given alignment method, or *matcher*, remove? And are the endpoints themselves good models under the computation used to evaluate the midpoint?

Existing measurements rarely separate these questions. Permutation matching and repair methods report the barrier left by a particular search procedure (Entezari et al., 2021; Ainsworth et al., 2022; Jordan et al., 2022). Many reported merging successes involve fine-tuned descendants of a shared run, so independently trained pairs are less studied. A relative barrier also does not show whether its endpoints are good models.

Weight-tied, or looped, language models make the third question unavoidable. Such a model applies one shared transformer block R times, so identical weights can be evaluated with different amounts of computation, which we call runtime depth (Dehghani et al., 2018; Giannou et al., 2023; Geiping et al., 2025). A smaller relative barrier at greater depth may then reflect a worse reference endpoint rather than a better midpoint.

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

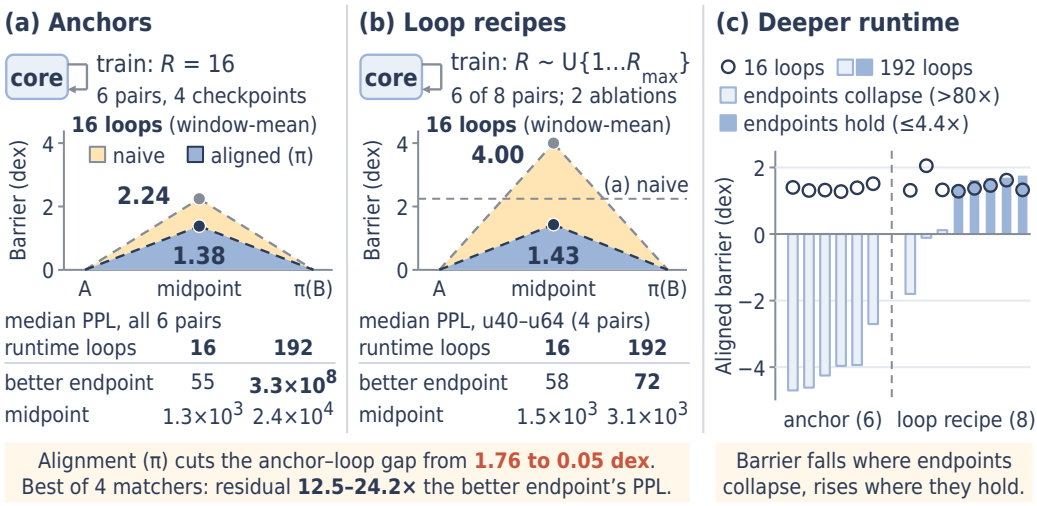


Figure 1: **Alignment removes the recipe contrast and isolates a shared residual; at deeper runtime, endpoint quality sets the relative barrier.** (a,b) Independently trained 35 M weight-tied models whose shared core is unrolled R times: six anchor pairs from four checkpoints trained at $R = 16$ (a), and eight loop-recipe pairs, six trained with $R \sim U\{1, \ldots, R_{\max}\}$ and two loop ablations (b). Each apex is the class-median midpoint barrier at 16 loops in dex (base-10 log of midpoint over better-endpoint perplexity), for naive interpolation (cream) and after the original activation-Hungarian permutation π of B onto A (blue); dashed sides are schematic (full paths in Fig. 2). Alignment cuts the anchor–loop gap from 1.76 to 0.05 dex, and the best of four matchers lowers the residual to 12.5–24.2× the better endpoint's perplexity on the 14 pairs other than the normalization ablation. Inset tables: median perplexity (PPL) of the better endpoint and the aligned midpoint (in (b), pairs `u40`–`u64`). (c) Aligned barrier of each pair at 192 loops (bars) and 16 loops (open circles). It falls to about zero or below on the nine pairs whose better-endpoint perplexity grows over 80-fold from 16 to 192 loops (pale bars) and rises on the five whose growth is at most 4.4-fold (solid bars). Apexes use window-mean perplexity; inset tables and (c) use corpus perplexity (Section 2).

We separate the three questions in fifteen pairs of 35 M-parameter looped language models that share one tokenizer, data stream, and training budget of 0.25 B tokens. Six *anchor* pairs are trained at a fixed 16 loops. Six of the eight *loop-recipe* pairs draw a random loop count for each batch, the other two are loop ablations, and a fifteenth pair is a normalization ablation. The key design choice is to measure every pair along three axes at once: before and after a permutation alignment that respects the shared weights, at the trained depth of 16 loops and at 192 loops, and with midpoint and endpoint perplexities recorded separately. The barrier is the ratio of midpoint perplexity to the better endpoint's perplexity, expressed in dex, its base-10 logarithm (1 dex is a tenfold ratio). Two controls probe the *aligned displacement*, the weight difference after alignment: a random displacement of equal norm tests its magnitude, and equal-norm perturbations inside and outside its leading subspace test its direction. Same-initialization descendants and public Pythia pairs (Biderman et al., 2023) complement the independently trained primary pairs.

Figure 1 summarizes the main findings. Alignment removes most of the large naive (unaligned) contrast between anchors and loop recipes and isolates a midpoint residual shared by both (panels a and b). Deeper evaluation separates the recipes again, largely through their endpoints: the aligned barrier falls where endpoints collapse and persists where they stay usable (panel c). We therefore recommend reporting relative barriers alongside absolute endpoint quality and runtime depth. Our contributions are:

- **A joint alignment–quality–depth protocol.** Each barrier is reported with the achieved alignment, the absolute quality of both endpoints, and the runtime depth, together with norm- and direction-controlled perturbation baselines (Section 2).
- **A decomposition of the recipe contrast.** Alignment reduces the naive contrast between anchors and loop recipes from 1.76 dex (bootstrap 95% interval $[0.07, 2.77]$) to below 0.05 dex, and the

best of four matchers isolates a residual of 12.5–24.2× the better endpoint's perplexity on the fourteen pairs other than a normalization ablation. Norm-matched random perturbations reproduce this scale, and equal-norm perturbations in the leading subspace of the aligned displacement cause more damage than orthogonal ones on all fourteen pairs (median $+1.1$ dex; Section 3).

- **A depth analysis of the reference endpoints.** At 192 loops, the aligned barrier falls to about zero or below on the nine pairs whose better-endpoint perplexity grows more than 80-fold, including all six anchors, whose median better-endpoint perplexity reaches 3.3×10^8. Between depth-stable endpoints the aligned residual persists and grows (on `u56`, from 1.32 dex at 16 loops to 1.49 dex at 192 loops on a ten-point depth grid, window-mean perplexity; Section 4).
- **Measured repairs.** Brief continued training from the aligned midpoint removes 90.4–92.0% of the aligned barrier on the two depth-stable pairs, and representation stitching recovers 90% and 95% of midpoint damage on the 160M and 410M Pythia pairs (Section 5.2). A probability-mixture teacher lowers held-out perplexity by 7.3%, and distilling it keeps deep-runtime degradation at 0.017–0.048 dex across three seeds, versus 0.40 dex for plain continuation (Section 6).

2 Models and comparison quantities

Architecture and training. The primary 35 M models use width 512, eight heads, one pre-block, a shared attention–MLP core, and one post-block. The core repeats with $z = \mathrm{LN}_{\mathrm{inj}}(h + e)$ and $h \leftarrow h + \mathrm{core}(z)$ (e: pre-block output; LN: layer normalization); the normalization ablation uses $h \leftarrow \mathrm{RMSNorm}(h + \mathrm{core}(z))$. Fixed-depth models train at $R = 16$. Sampled-depth recipes draw a batch-level count uniformly from $\{1, \ldots, R_{\max}\}$, with $R_{\max} = 24$–64, including the evaluation depth 16. Both are evaluated with their weights frozen at runtime depths up to 192 (schematic in Appendix A).

Training uses a FineWeb-Edu sample (Penedo et al., 2024), the GPT-2 tokenizer (vocabulary 50,304), context 1024, 49,152 tokens per update, and 5,000 updates (0.246 B tokens). AdamW uses $(0.9, 0.95)$, matrix weight decay 0.1, and cosine decay after 200 warmup updates. Width and budget controls vary only the stated factors; Appendices N and O list run configurations and detailed results.

Pairs, classes, and uncertainty units. The fifteen-pair population comprises three classes. Six fixed-depth pairs formed from four checkpoints make up the *anchor* class (`anchor`ij pairs checkpoints i and j). Eight *loop-recipe* pairs comprise the sampled-depth pairs `u`N, trained with $R_{\max} = N$, and two loop ablations: `mb2`, whose shared core has two blocks, and `k32f`, trained at a fixed depth of 32. One pair, `dnlnrms`, is the normalization ablation. Unless stated otherwise, ranges over "the fourteen pairs" exclude only the normalization ablation; configurations are listed in Table 3 (Appendix N). The six anchors reuse endpoints, so we report pair-level values and name the resampling unit for each interval, treating evaluation windows and training replicates as distinct units. Width, public-model, and positional-profile studies have their own pair sets; Pythia comparisons use original-versus-deduplicated checkpoints (Biderman et al., 2023).

Two perplexity estimators. Each evaluation uses sixteen windows, with 1024 scored positions per window, drawn from the last 4 M tokens of a fixed held-out shard with random seed 0. If ℓ_j is a window's mean token loss, *corpus* perplexity is $P_{\mathrm{geom}} = \exp(\frac{1}{16}\sum_j \ell_j)$. The trained-depth interpolation measurements instead use $P_{\mathrm{arith}} = \frac{1}{16}\sum_j \exp(\ell_j)$, the *window-mean* perplexity. Every displayed comparison states its basis: the trained-depth columns and the rod$_a$ column of Table 1, Fig. 2, the best-of-four ranges, and the ten-point depth grid of Section 4 use P_{arith}, whereas the deep log-barrier column of Table 1, Fig. 3, and the inset tables and panel (c) of Fig. 1 use P_{geom}, recomputed from the retained per-window values.

Weight interpolation and relative damage. For a valid alignment π, form the float32 weight path $\theta_t = (1-t)\theta_A + t\pi(\theta_B)$; with $\pi = I$ this is *naive* (unaligned) interpolation. The main statistic is

$$b_R(\pi) = \log_{10} \frac{P_R(\theta_{1/2})}{\min(P_R(\theta_A), P_R(\theta_B))}. \tag{1}$$

With corpus perplexity this equals a midpoint-versus-best-endpoint cross-entropy difference divided by $\ln 10$; it compares the midpoint with the better endpoint, which differs from the maximum excess

above the line joining endpoint losses along the full path. We report b_R in dex and use 0.5 dex, a 3.16-fold ratio, as a descriptive comparison cutoff. Trained-depth grids use $t = 0, 1/8, \ldots, 1$ or $0, 1/4, \ldots, 1$; the deepest probe uses $t \in \{0, 1/2, 1\}$.

Achieved alignment and runtime drift. The original activation-Hungarian procedure matches residual-stream coordinates, whole attention heads, and MLP-hidden units under the model's weight-tying constraints. Four further matchers are compared separately: activation-free weight matching, optimal-transport (OT-Sinkhorn) activation matching, and two variants of the original matcher that keep the heads unpermuted or match them by weights. The *best of four* is their per-pair minimum, which OT-Sinkhorn attains on every pair. We call $b_R(I) - b_R(\pi)$ the removed component, or *shell* R_{sh}, and $b_R(\pi)$ the achieved aligned *residual*; shells are reported with their sign. A norm-matched *twin* replaces the aligned displacement $\pi(\theta_B) - \theta_A$, the weight difference after alignment, with a random displacement of the same per-tensor norm. To interpret depth, we also record $D_i = P_{192}(\theta_i)/P_{16}(\theta_i)$ for both endpoints and the midpoint, and the midpoint drift ratio $\mathrm{rod} = D_{\mathrm{mid}}/\max(D_A, D_B)$. These ratios are reported alongside absolute losses, so that every relative barrier can be read against the quality of the models it compares.

3 Alignment explains the recipe contrast and isolates a shared residual

3.1 Naive interpolation differs across training recipes

How large is the naive midpoint damage, and does it differ between anchors and loop recipes? Every primary pair has a trained-depth midpoint above the 0.5 dex comparison cutoff, across all 240 recorded pair–window cells (median of the per-pair window-median barriers 2.38 dex). Uniformly averaging the four anchor checkpoints gives a soup perplexity about 1.2×10^3 times the mean member perplexity. Figure 2 shows the full paths behind these midpoints.

The naive contrast is large: the loop class has median $b_{16} = 4.00$ dex, versus 2.24 dex for the fixed-depth anchors, a 1.76 dex difference, with a two-stage pair-cluster bootstrap 95% interval of $[0.07, 2.77]$ dex (Appendix Q). A separately trained dense 18-layer, 79 M-parameter control also has substantial naive damage ($b_{16} = 1.83$ dex) despite stronger endpoints (PPL 42.0/41.9), so the damage also occurs without weight tying. By contrast, two continuations of one endpoint on disjoint data streams, i.e., descendants of a shared initialization as in model soups, show no naive barrier (at most 0.0072 dex at 16 loops), and their midpoint beats both endpoints (Appendix F.5).

3.2 Alignment isolates a residual shared by both recipes

How much of the recipe contrast is coordinate misalignment that permutation matching can remove? We permute endpoint B toward A under the tied residual-stream, whole-head, and MLP-hidden symmetries. The permuted endpoints reproduce the original endpoint measurements on all fifteen pairs. Fig. 2 shows the original activation-Hungarian curves; Table 1 adds the best of four matchers.

A large recipe contrast becomes small after matching. After matching, the best-of-four residual is 1.20 dex on the loop recipes and 1.24 dex on the anchor pairs (class medians), so the removed shell measures 2.80 and 1.00 dex per class (Table 1; Fig. 7 in the appendix). The 1.76 dex class difference becomes $+0.048$ dex for the original activation-Hungarian matcher and -0.043 dex for the best of four matchers, both on the window-mean basis (two-stage bootstrap 95% interval for the former: $[0.003, 0.31]$ dex, Appendix Q). The latter removes 0.41–3.91 dex per pair, with median 1.15 across all fifteen pairs (1.22 over the fourteen pairs). Most of the reduction is carried by residual-stream and MLP-hidden matching; removing the whole-head permutation class changes the shell modestly (Appendix B.1). This reduction supports a coordinate-dependent explanation of much of the *between-recipe contrast*. The six anchor pairs share four checkpoints, whereas each loop pair consists of two seeds of one recipe; checkpoint-aware resampling and leave-out estimates preserve the small aligned contrast (Appendix H).

The residual is consistent across function-informed matchers. Across the fourteen pairs, the best of four matcher families attains a midpoint ratio of 12.5–24.2× (median 16.9×, window-mean

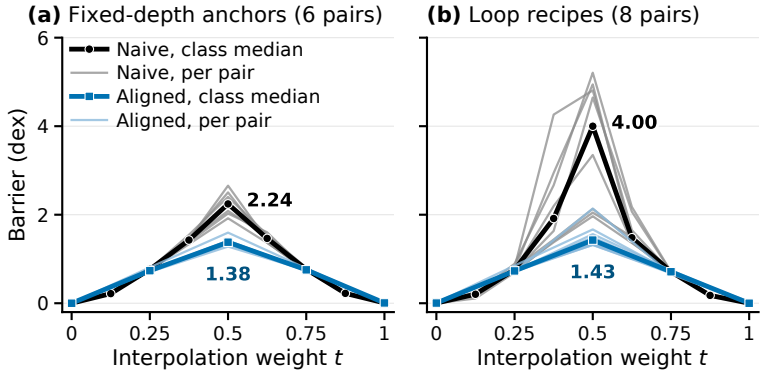


Figure 2: Naive and aligned interpolation at sixteen loops. The vertical axis is the barrier $\log_{10}[P_t/\min(P_0, P_1)]$ in dex, using window-mean perplexity. Thin lines are individual pairs; heavy lines are class medians. Gray/black: naive interpolation; blue: original activation-Hungarian alignment. (a) Six anchor pairs from four checkpoints. (b) Eight loop-recipe pairs including two ablations. Markers on the class medians mark the evaluated weights (naive: steps of $1/8$; aligned: steps of $1/4$); segments between them are drawn for display only. Numbers are class-median midpoint barriers in dex, naive (black) and aligned (blue). All cells use the same sixteen windows; per-pair traces show the spread.

perplexity). Activation-free weight matching, by contrast, leaves a midpoint perplexity 1.8×10^{11} times that of the better endpoint on `u56`, so function-informed matching drives the reduction (Appendix I). With corpus perplexity, the original activation-Hungarian matcher gives ratios of 18.8–112.2×. The matched dense-18 control has a comparable 1.29 dex best-family residual. Every tested matcher leaves every pair above the 0.5 dex cutoff. A disjoint-window evaluation shifts the activation-Hungarian residual by a median -0.006 dex (median absolute shift 0.035 dex; Appendix H).

Random perturbations reproduce the magnitude, and direction adds damage. Norm-matched random twins reach a median 1.29 times the aligned log-barrier on the first window draw (distribution-free order-statistic interval $[1.13, 1.43]$, exact coverage 98.7% for $n = 14$), and 1.42 on a second draw. Perturbations of comparable norm thus produce the observed scale of damage. Pairwise, the twin tracks the aligned residual on six pairs, the naive barrier on three, both on three, and is ambiguous on two (Appendix I). At identical tensorwise dose, perturbations in the displacement's top-eight output subspace cause more damage than those in its orthogonal complement on all 14 pairs (median $+1.14$ dex; paired bootstrap 95% interval $[+0.66, +2.16]$). Perturbing a random eight-dimensional subspace at the same dose, instead of the displacement's leading one, causes a median 1.27 dex less damage. Thus both norm budget and direction shape the residual.

The residual lies on a shared norm-dose curve. Scaling the twin displacement to $\lambda \in \{0.25, 0.5, 1, 2\}$ times the matched per-tensor norm raises the population-median twin barrier superlinearly, from 0.08 to 0.40, 1.83, and 4.48 dex. Anchor and loop-recipe pairs follow the same growth ratios (×4.6–4.9 per doubling up to the matched norm, then ×2.4–2.6), so the measured residual lies on a norm-dose curve shared by both recipes; a per-tensor norm-rescaled aligned path keeps the residual (median 1.48 versus 1.41 dex; Appendix F.3).

The decomposition replicates on a second corpus. A fresh `u56`-recipe pair trained on WikiText-103 (Merity et al., 2016) with the same architecture, vocabulary, and token budget (endpoint PPL 45.8/48.4) reproduces the pattern: the naive barrier exceeds the 0.5 dex cutoff on all sixteen windows (median 2.22 dex), alignment leaves a residual of 1.21 dex (16.1× the better endpoint), norm-matched twins again reach at least the residual's scale (1.70 times its log-barrier), and both endpoints remain depth-stable at 192 loops (1.17–1.22×; Appendix O).

Table 1: All fifteen pairs. Trained-depth columns (window-mean perplexity): the naive midpoint log-barrier, the shell R_{sh} removed by alignment, and the best-of-four post-alignment residual B_a^* (as a log and as a ratio to the better endpoint). R=192 columns (original activation-Hungarian alignment): the deep aligned log-barrier (corpus perplexity) and rod_a (window-mean perplexity), the aligned midpoint's depth drift divided by the larger endpoint drift (values below 0.01 shown as <0.01). Anchor: fixed-depth anchors; Loop recipe: sampled-depth recipes and two loop ablations; Norm abl.: normalization ablation (excluded from fourteen-pair ranges). Logarithms are base ten.

		Trained depth R=16 (window-mean PPL)				Runtime depth R=192	
						corpus PPL	window-mean PPL
Class	Pair	Naive $\log B$ (dex)	Shell R_{sh} (dex)	$\log B_a^*$ (dex)	B_a^* (×)	$\log B_{a,192}$ (dex)	rod_a (ratio)
Anchor	anchor01	2.66	1.38	1.27	18.7	−2.71	<0.01
	anchor02	2.03	0.81	1.22	16.4	−4.70	<0.01
	anchor03	2.40	1.15	1.25	17.8	−4.25	<0.01
	anchor12	2.50	1.29	1.22	16.4	−4.62	<0.01
	anchor13	2.09	0.84	1.24	17.5	−3.94	<0.01
	anchor23	1.92	0.68	1.24	17.3	−3.96	<0.01
Loop recipe	u24	1.97	0.78	1.18	15.3	−1.80	<0.01
	u32	3.35	1.97	1.38	24.2	−0.06	<0.01
	u40	5.21	3.91	1.30	19.9	1.69	1.02
	u48	4.81	3.60	1.21	16.3	1.62	0.07
	u56	4.94	3.78	1.16	14.5	1.44	1.41
	u64	4.65	3.50	1.15	14.0	1.68	1.91
	mb2	2.13	0.87	1.27	18.4	0.01	<0.01
	k32f	2.05	0.95	1.10	12.5	1.75	2.61
Norm abl.	dnlnrms	1.89	0.41	1.49	30.6	1.72	1.48

4 Depth changes the reference as well as the midpoint

Does a measured barrier stay meaningful when the same aligned weights run at greater depth?

Endpoint quality determines the deep relative barrier. The same aligned weights are evaluated at sixteen and 192 loops in Figure 3. On all six anchor pairs the endpoints deteriorate much faster than the midpoint ($\mathrm{rod}_a < 0.01$, Table 1). Their median aligned relative barrier therefore falls to -4.11 dex at 192 loops, while their absolute perplexity becomes unusable (median better-endpoint perplexity 3.3×10^8; Fig. 1a). The absolute measurements thus identify this drop as a relative improvement against damaged references. Loop-recipe pairs instead retain a median aligned barrier of 1.53 dex. Across all fourteen pairs, the aligned barrier falls to about zero or below on the nine pairs whose better endpoint degrades more than 80-fold, and rises on the five whose better endpoint changes at most 4.4-fold (Fig. 1c). Appendices F and K report drift ratios and positional traces.

Depth-stable pairs retain a measurable residual. On two depth-stable loop pairs (u40 and u56), endpoint PPL remains 60.2–80.1 over a ten-point depth grid from 16 to 192 loops. Their naive relative barriers decrease with depth, while the aligned residual grows: on u56 from 1.32 dex at 16 loops to 1.49 dex at 192 (window-mean perplexity; about 31 times the better endpoint; Appendix K), and on u40 to 1.69 dex at 192 loops (Table 1, corpus perplexity). Thus the residual is also present between usable endpoints. Alignment lowers the fraction of token positions where the midpoint's cross-entropy exceeds twice that of both endpoints from 0.60–0.72 (naive) to 0.36, and the residual is carried by the remaining affected positions. Training at a range of depths moderates deep degradation: the fixed-depth recipe with its sampler changed to mild depth jitter, $R \sim U\{12, \ldots, 20\}$, falls between the anchor and loop classes at 192 loops (Appendix B.2).

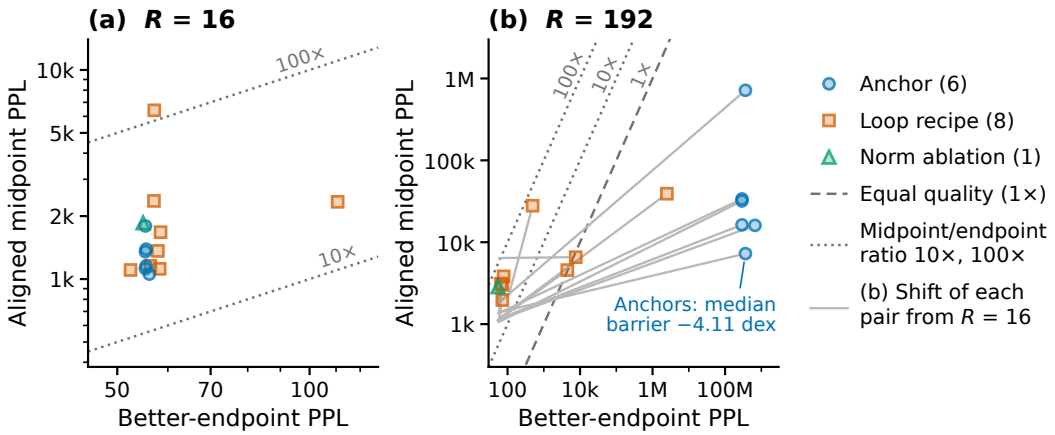


Figure 3: Absolute quality explains the relative barrier at depth. Better-endpoint PPL (horizontal) against aligned-midpoint PPL (vertical) for each pair, with the original activation-Hungarian alignment and corpus perplexity; log axes, k $= 10^3$, M $= 10^6$. Color and marker give the class, as in Fig. 8a: anchors (blue circles), loop recipes (vermillion squares), normalization ablation (green triangle). Dashed line: equal quality; dotted lines: midpoint $10\times$ and $100\times$ the better endpoint. (a) R=16; (b) R=192, with gray lines from each pair's 16-loop position (different axis ranges). At deep runtime, lower relative damage accompanies poorer endpoint quality: the anchors reach a median better-endpoint PPL of 3.3×10^8 and a median aligned barrier of -4.11 dex.

Table 2: Width ladder and budget sibling (naive interpolation, before alignment). $\log B_{16}$ and $\log B_{192}$ are midpoint log-barriers at 16 and 192 loops; D_A, D_{mid}, and D_B are depth drifts P_{192}/P_{16} of endpoint A, the midpoint, and endpoint B; the drift ratio rod is $D_{\text{mid}}/\max(D_A, D_B)$. Ratios are given to three significant figures, except the budget row's D_{mid} and rod, which retain the one decimal place quoted in the text. The budget row compares same-initialization checkpoints: its trained-depth barrier is 1.54 dex lower, and its rod is the largest measured. Widths 1280 and 2560 are in Appendix E.

	Barrier (dex)		Depth drift P_{192}/P_{16} (ratio)			
Width d (seeds 0 vs 2)	$\log B_{16}$	$\log B_{192}$	D_A	D_{mid}	D_B	rod (ratio)
1024	2.48	2.81	1.47	2.73	1.28	1.86
1536	2.43	3.15	1.03	5.30	1.02	5.15
2048	2.46	3.04	1.03	3.93	1.02	3.82
3072	2.04	3.99	1.06	95.1	1.05	89.7
2048, seed 0 vs 2×-budget sibling	0.92	3.89	1.03	959.5	1.04	923.6

5 Width, Training Budget, and Measured Repairs

5.1 Width and Training Budget

How does width change the measured damage at a fixed token count, and what changes when the training budget is doubled? Table 2 reports the fixed-token width ladder. Across the four main widths 1024–3072, the trained-depth barrier is near 2.45 dex through 2048 and falls by 0.42 dex at 3072. The rank correlation between width and barrier is $\rho_S = -0.8$ over these four widths ($n = 4$) and -0.6 over all six widths ($n = 6$). At 192 loops the midpoint drift ratio rod is not monotone in width: it is about 1.2 at width 2560 but 89.7 at 3072 (Appendix E, Fig. 5).

A doubled-budget width 2048 sibling, which shares initialization and data order with its partner, reduces the trained-depth barrier from 2.46 to 0.92 dex and has the largest deep drift ratio, 923.6. The two depths thus respond separately to budget, so the protocol reports both.

5.2 WHAT THE MEASURED REPAIRS ACCOMPLISH

What do the tested repairs recover?

Brief continued training closes the trained-depth gap. Interpolated midpoints are effective starting points for training. On the two depth-stable loop pairs (`u40`, `u56`), 1500 steps of continued training from the midpoint (30% of the training budget) remove 96.5–96.7% of the naive and 90.4–92.0% of the aligned barrier (Appendix F.1). The naive midpoint catches up quickly: after 500 steps it is only about 0.06 dex above the aligned one on both pairs, and on `u56` both recovered midpoints are within 1.4–1.7× of the endpoints' perplexity at 192 loops.

Comparison with channel-moment renormalization. We also compared two REPAIR-style variants (Jordan et al., 2022), which rescale each activation channel of the aligned midpoint toward the average of the endpoints' first and second moments, either in the residual stream alone or also in the attention- and MLP-internal channels. On three representative pairs (a fixed-depth anchor at the depth where its endpoints begin to degrade, and two loop pairs whose deep barrier persists), the plain aligned midpoint is 1.23–1.84 dex better than both variants on every pair, including settings where the rescaled midpoint does not collapse to a uniform output distribution (Appendix M).

Representation stitching repairs the 160M and 410M Pythia pairs. For original-versus-deduplicated Pythia pairs, the naive barriers are 4.40, 3.17, and 2.05 dex at 70M, 160M, and 410M. At 160M and 410M, representation stitching recovers 90% and 95% of midpoint damage (Appendix R). Section 6 turns to probability mixing and distillation.

6 PROBABILITY MIXING AND CONSOLIDATION

Can probability mixing and distillation combine the endpoints' information, and under which budgets? Weight interpolation and prediction combination are different operations. On one stronger loop pair, trained with five times the standard budget (Appendix C), each endpoint is the more accurate model at about half of the token positions (the better endpoint at 50.1%), so their predictions are complementary; this motivates a probability mixture.

The probability teacher improves held-out loss. For endpoint predictions p_A, p_B and $q_\alpha = (1-\alpha)p_A + \alpha p_B$, the measured midpoint teacher has corpus PPL 34.375 versus 37.088 for the better endpoint, a 7.3% reduction. The best measured weights lie in $[0.40, 0.60]$ across the reported calibration measurements; scoring new weights reuses cached endpoint probabilities.

The teacher's gain is uniform and transfers across corpora. The midpoint teacher improves on endpoint A on all sixteen windows (margins 0.043–0.152 nats) and is itself depth-stable: its perplexity at 192 loops is 1.02× its 16-loop value, within the endpoints' own range of 1.02–1.04. On WikiText-103 windows never used in training, the same teacher beats the stronger endpoint by 11.2%, with a positive margin on every window (Appendix O).

Distillation preserves deep-runtime quality. The student p_s starts from intact endpoint A and minimizes the cross-entropy (CE) and KL-divergence objective $(1-\lambda)\mathrm{CE} + \lambda\mathrm{KL}(q_{0.5}\|p_s)$ with $\lambda = 0.5$. The control has identical initialization, token stream, and depth draws but $\lambda = 0$. After 12.5 k updates, the student captures 67.6% of the teacher–endpoint gap. The control has similar trained-depth quality but worse deep-runtime PPL (91.85, window-mean perplexity), a 2.53-fold depth ratio versus about 1.08 for the student, so the student's advantage lies in deep-runtime quality. Capture grows with budget without a plateau (38%, 67.6%, and 88.8% at 5 k, 12.5 k, and 25 k updates), and on WikiText-103 windows the 12.5 k student remains depth-stable (depth ratio 1.03).

At the full continuation budget, continued training from endpoint A improves corpus PPL to 33.741, surpassing the 34.375 teacher. On a second pair, the `u56` pair of the main population (endpoint PPL 61.2 and 60.2), both methods surpass their teacher after 2,500 steps, half the pair's training budget, and both remain depth-stable. Appendix L gives all budget and seed results.

Seeds and controls. Across three seeds at $\lambda = 0.5$, the student captures 0.67–0.69 of the teacher gap while its deep-runtime degradation stays at 0.017–0.048 dex, versus 0.40 dex for the continued-training control. The capture requires both the combined objective and the endpoint initialization, as KL-only and from-scratch controls show (Appendix F.5). Including the teacher's two forward passes, a distillation step costs about $5/3$ plain steps; at matched compute, distillation contributes deep-runtime stability (Appendix G.1).

7 Related work

Connectivity and coordinate choices. Loss-path studies distinguish linear slices from curved connections (Goodfellow et al., 2014; Garipov et al., 2018; Draxler et al., 2018; Kuditipudi et al., 2019; Frankle et al., 2019; Vlaar & Frankle, 2021). Permutation invariance, weight matching, and activation repair directly address coordinate misalignment (Entezari et al., 2021; Ainsworth et al., 2022; Jordan et al., 2022). Optimal-transport fusion and feature matching broaden the alignment family (Singh & Jaggi, 2019; Cuturi, 2013; Stoica et al., 2023). Transformer alignment extends these methods to attention and residual-stream symmetries via optimal transport (Imfeld et al., 2024) and permutation-based merging of independently initialized text models (Verma & Elbayad, 2024). We impose weight-tying constraints and report residuals with endpoint quality across runtime depths.

Averaging, merging, and prediction combination. Trajectory averaging and fine-tuned model soups share optimization history (Izmailov et al., 2018; Wortsman et al., 2022; 2021). Task arithmetic, interference reduction, sparsified merging, Fisher weighting, and regression-based fusion provide additional methods (Ilharco et al., 2022; Yadav et al., 2023; Yu et al., 2023; Matena & Raffel, 2021; Jin et al., 2022); our pairs are independently trained rather than fine-tuned descendants of a shared model. Probability mixing and distillation (Hinton et al., 2015) combine predictions rather than remove weight-space barriers; we compare them with continuation at the measured additional budgets.

Representations, width, and repeated computation. Representation comparisons and stitching test distinct forms of compatibility (Raghu et al., 2017; Kornblith et al., 2019; Bansal et al., 2021; Csiszárik et al., 2021). Sufficiently wide classifiers can achieve low-loss linear connectivity without permutation after suitable softmax-temperature calibration (Ito et al., 2026); our fixed-token width ladder measures interpolation at the default softmax temperature. Weight sharing adds runtime depth as an axis (Dehghani et al., 2018; Lan et al., 2019; Giannou et al., 2023; Geiping et al., 2025; Zhu et al., 2025); we evaluate midpoints and endpoints at several unroll counts.

8 Conclusion and limitations

Conclusion. Jointly measuring permutation alignment, runtime depth, and endpoint quality clarifies weight interpolation. In the measured 35 M loop population, alignment explains the recipe contrast and isolates a shared residual. Norm-matched perturbations reproduce its scale; perturbations along the aligned displacement add damage. Deeper evaluation distinguishes relative valleys caused by endpoint collapse from residuals between usable endpoints. Continued training and probability-mixture distillation provide repairs, with the latter preserving deep-runtime quality. We recommend reporting achieved alignment, absolute endpoint quality, and runtime depth together.

Limitations and future work. Four directions would extend this study. *Scale and replication:* controlled measurements cover one small weight-tied family with few independent runs per configuration; larger models and more seeds would test generality. The tested matchers remove at most 5% of the naive barrier on public Pythia pairs (Appendix R), indicating family-dependent alignment benefits. *Alignment coverage:* the reported residuals are upper bounds for the tested matchers and estimators, which searches over further symmetries could reduce. *Width and budget:* a compute-matched ladder with softmax-temperature calibration would support a scaling study of width and budget. *Repairs:* consolidation comparisons cover few pairs per setting. Repairs for recurrent computation, including representation- and circuit-level methods, remain to be compared. Appendix U gives details.

Ethics Statement

This work uses publicly available corpora, benchmarks, and model checkpoints. We did not conduct human-subject experiments. The work shares the risks common to language-model research, including bias and potentially harmful content in public corpora and model outputs.

Reproducibility Statement

The supplementary ZIP archive provides code, configurations, launch and analysis scripts, and recorded results, together with reproduction instructions. Experimental settings are specified in the main text and appendix.

AI Use Statement

During the project, vibe coding tools (Codex and Claude Code) were used to help develop the code for the methods and experiments. All such work was conducted under human supervision and underwent human review. During manuscript preparation, AI tools were used to polish and revise the text and figures.

References

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022. URL https://arxiv.org/abs/2209.04836.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *arXiv preprint arXiv:2106.07682*, 2021. URL https://arxiv.org/abs/2106.07682.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*, 2023. URL https://arxiv.org/abs/2304.01373.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. *arXiv preprint arXiv:2110.14633*, 2021. URL https://arxiv.org/abs/2110.14633.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *arXiv preprint arXiv:1306.0895*, 2013. URL https://arxiv.org/abs/1306.0895.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal Transformers. *arXiv preprint arXiv:1807.03819*, 2018. URL https://arxiv.org/abs/1807.03819.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. *arXiv preprint arXiv:1803.00885*, 2018. URL https://arxiv.org/abs/1803.00885.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021. URL https://arxiv.org/abs/2110.06296.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. *arXiv preprint arXiv:1912.05671*, 2019. URL https://arxiv.org/abs/1912.05671.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. *arXiv preprint arXiv:1802.10026*, 2018. URL https://arxiv.org/abs/1802.10026.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025. URL https://arxiv.org/abs/2502.05171.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023. URL https://arxiv.org/abs/2301.13196.

Ian J. Goodfellow, Oriol Vinyals, and Andrew M. Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014. URL https://arxiv.org/abs/1412.6544.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. URL https://arxiv.org/abs/2212.04089.

Moritz Imfeld, Jacopo Graldi, Marco Giordano, Thomas Hofmann, Sotiris Anagnostidis, and Sidak Pal Singh. Transformer fusion with optimal transport. In *International Conference on Learning Representations*, 2024.

Akira Ito, Masanori Yamada, Daiki Chijiwa, and Atsutoshi Kumagai. Do we really need permutations? Impact of model width on linear mode connectivity. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5eceaacf571380e70c1052ef6bd58cb3-Abstract-Conference.html.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. URL https://arxiv.org/abs/1803.05407.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022. URL https://arxiv.org/abs/2212.09849.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. *arXiv preprint arXiv:2211.08403*, 2022. URL https://arxiv.org/abs/2211.08403.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019. URL https://arxiv.org/abs/1905.00414.

Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Sanjeev Arora, and Rong Ge. Explaining landscape connectivity of low-cost solutions for multilayer nets. *arXiv preprint arXiv:1906.06247*, 2019. URL https://arxiv.org/abs/1906.06247.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. URL https://arxiv.org/abs/1909.11942.

Michael Matena and Colin Raffel. Merging models with Fisher-weighted averaging. *arXiv preprint arXiv:2111.09832*, 2021. URL https://arxiv.org/abs/2111.09832.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024. URL https://arxiv.org/abs/2406.17557.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv preprint arXiv:1706.05806*, 2017. URL https://arxiv.org/abs/1706.05806.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *arXiv preprint arXiv:1910.05653*, 2019. URL https://arxiv.org/abs/1910.05653.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. ZipIt! merging models from different tasks without training. *arXiv preprint arXiv:2305.03053*, 2023. URL https://arxiv.org/abs/2305.03053.

Neha Verma and Maha Elbayad. Merging text transformer models from different initializations. *Transactions on Machine Learning Research*, 2024.

Tiffany Vlaar and Jonathan Frankle. What can linear interpolation of neural network loss landscapes tell us? *arXiv preprint arXiv:2106.16004*, 2021. URL https://arxiv.org/abs/2106.16004.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. URL https://arxiv.org/abs/2109.01903.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022. URL https://arxiv.org/abs/2203.05482.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023. URL https://arxiv.org/abs/2306.01708.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. *arXiv preprint arXiv:2311.03099*, 2023. URL https://arxiv.org/abs/2311.03099.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025. URL https://arxiv.org/abs/2510.25741.

APPENDIX GUIDE

Appendix A gives the protocol schematic, Appendix B the permutation-class and training-support results summarized in Sections 3 and 4, and Appendix C the supplementary comparisons and the terms used in the detailed results. Appendices D–G report the anchor depth band (D), the width and budget ladder (E), positional tracers together with recovery, dose, and robustness results (F), and the capture–depth ladder of the consolidation (G). Appendices H–L give detailed alignment (H), decomposition (I), runtime-depth (J and K), and consolidation (L) measurements. Appendix M discusses the tested repairs and their scope, Appendix N lists the training configurations and the artifact index, Appendix O collects the per-campaign result records, Appendix P maps headline claims to their evidence, and Appendix Q gives the uncertainty estimates. Appendices R–T cover the public Pythia pairs (R), random-permutation and cross-fit checks (S), and the relation between alignment benefit and endpoint distance (T); Appendix U extends the discussion of limitations.

Per-run scientific outcomes, including negative results, are reported. Experiment labels (for example MB17 or RE3) name the measurement campaign a result comes from and do not define new statistical units. Most campaigns were run under a pre-registered specification, which is included in the supplementary archive (folder `projects/lmc_basin/versions/`). Words in small capitals (for example RAMP-DEAD) are the verdict labels defined in those specifications; the surrounding text states the measured quantity each label summarizes, and Appendix P explains the labels in plain language.

A PROTOCOL SCHEMATIC

Figure 4 schematizes the protocol of Section 2; the measured overview is Fig. 1.

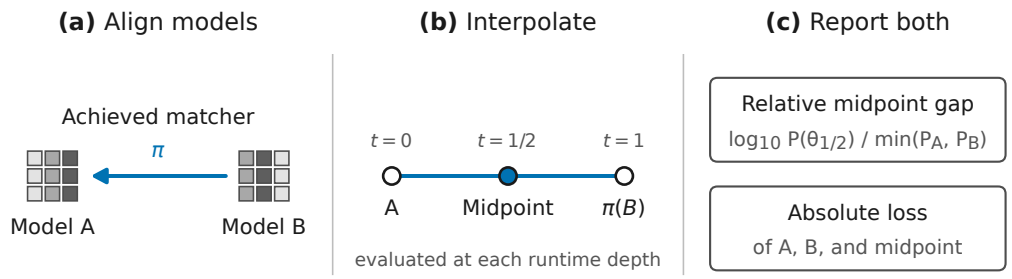


Figure 4: **Report relative and absolute interpolation quality together.** An implemented tying-aware matcher aligns one endpoint to the other's coordinates; it is an achieved heuristic, not a global optimum or a transformation of one model's function into the other's. The midpoint and both endpoints are evaluated at each runtime depth. The relative measure is the log ratio of midpoint perplexity to the better endpoint, not a maximum-path barrier. Absolute endpoint and midpoint losses reveal when a smaller relative gap accompanies poor models. The matcher π permutes B onto A; the aligned path $\theta_t = (1-t)\theta_A + t\pi(\theta_B)$ is drawn in blue, as in Fig. 2. Schematic only; measured results are shown in the main text.

B PERMUTATION CLASSES AND TRAINING SUPPORT

This appendix gives the full results behind two one-sentence summaries in the main text (Sections 3 and 4).

B.1 WHICH PERMUTATION CLASSES CONTRIBUTE?

Removing the whole-head class on the four largest-shell hosts reduces the shell by a median 0.25 dex (4.8–7.2%). Most of the measured reduction is therefore carried jointly by residual-stream and MLP-hidden matching, rather than by head permutation alone. Per host (`u40`,

u48, u56, u64), the shell without head permutation (matcher M3A) is 0.26, 0.24, 0.18, and 0.25 dex (6.8%, 6.7%, 4.8%, and 7.2%) smaller than the OT-Sinkhorn (M2) shell; the record is versions/mb31_class_carrier.txt in the supplementary archive, and the related head-class attribution ladder is in Appendix O (MB15/MB19).

B.2 Training support moderates the depth failure

Changing only the clean fixed-depth recipe's sampler to $R \sim U\{12, \ldots, 20\}$ gives an intermediate outcome on two seeds. Endpoint drift is 1.03/1.05 at 32 loops and 1.32/1.74 at 64, but 22.9/17.2 at 192. The raw midpoint barrier falls from 2.19 to 1.64 dex, while the aligned residual is 1.51 dex. This supports a graded dependence on training support in the measured recipe; it does not establish a universal support-to-stability law. Detailed factor and position probes (Appendices J and K) show host-specific onsets and reject a universal ramp shape.

C Supplementary comparisons and terminology

The supplement expands the main comparisons with a dense-18 fixed-depth control, four matcher families, additional depth evaluations, checkpoint- and window-resampling analyses, channel-moment repair, and norm-matched random-averaging and model-soup controls. These comparisons address different potential explanations and are reported with their own units.

Training duration and corpus. Two fresh u56 runs use the original recipe at five times the training budget: 25,000 updates and 1.229×10^9 tokens, seeds 3 and 4. Endpoint perplexities on the common windows are $38.4/38.4$ (training-development reads $37.6/38.5$), compared with $61.2/60.2$ for the original pair. A separate five-times-budget pair, seeds 7 and 8, has endpoints $38.3/38.7$. A WikiText-103 replication uses fresh seeds 101 and 102, 2.458×10^8 tokens per endpoint, and reaches perplexity $45.8/48.4$ at $R = 16$. Its functional comparison, depth trajectory and matcher-temperature measurements are reported with that pair. Width-1280 and width-2560 pairs, per-position harm measurements, and additional matching on the four original width pairs further extend the tested configurations (Appendix O). Stronger endpoints, another corpus, and larger width are separate replications rather than a single pooled sample.

Terms used in the detailed results. A *breach* is the first position where cumulative midpoint excess crosses the stated threshold; it is not the onset of a causal process. *Ignition onset* denotes the smallest tested depth-ramp factor at which either endpoint's drift reaches $10\times$. Labels R-floor, R-saturation and R-acceleration describe observed ramp shapes. *Repair* means channel-moment renormalization of an interpolated point; soup q is soup perplexity divided by mean member perplexity. The terms "barrier" and "residual" in the detailed tables refer to their stated measured path or midpoint quantities, not a proven minimum over all permutations or paths.

Notation and recurring terms. ρ_n and ρ_a denote the naive and aligned midpoint log-barriers b_R (Eq. 1), by default at trained depth; ρ_N is the same statistic for a norm-matched twin, and $\log B$ is used interchangeably with ρ. In the appendices $R = \rho_n - \rho_a$ also denotes the *alignment benefit* (the shell) when it is not a loop count; K denotes a fixed training loop count (a fixed-$K16$ model trains at $R = 16$). A *host* is a trained configuration or pair (for example the u56 host), and an *arm* is one measurement condition (naive or aligned; one matcher; a distillation or control run). The *gauge* is the permutation symmetry of the weight-tied model, and the *gauge shell* is the part of the barrier that alignment removes. The *census* is the set of per-pair values over the fifteen principal (main-population) pairs, and a *census band* is their range. *Certified windows* are the sixteen fixed evaluation windows of Section 2, on which endpoint values are checked against stored values before a result is used; a *certificate* is the stored record of such a check. A *pre-specified* band, rule, or specification was fixed before measurement; a result is *scored* when it is evaluated once under that rule, recorded in a write-once score log; a *gate* is a pre-specified validity check; and *registered* means recorded in the campaign's pre-registered record. Depth-*tame* (depth-stable) endpoints change little in perplexity between 16 and 192 loops. The *usable band* T_{usable} of a depth campaign is its pre-specified perplexity ceiling, $4\times$ the larger $R{=}16$ endpoint perplexity; a cell inside it is SAFE, and a cell outside it is BREACHED. A *rung* is one budget level of a ladder; the consolidation arms are named by objective and budget (d12, d25: distillation at 12.5k and 25k steps; c12, c25: the

plain cross-entropy control at the same budgets). Matchers: M0 is the original activation-Hungarian matcher, M1 activation-free weight matching, M2 OT-Sinkhorn activation matching, and M3A/M3B the variants with the head class dropped or weight-replaced. G1, G2, and G3 denote the three permutation classes (residual-stream basis, whole attention heads, MLP hidden units) and, as match diagnostics, the mean absolute activation correlation after matching within each class. TR, TN, and TB are the twin classes defined in Appendix O. F2 is the recipe class gap (loop-class minus anchor-class median barrier) of Section 3, with subscripts n and a for the naïve and aligned paths. The *poles* of an interpolation path are its two endpoints; a *deep pole* is a large perplexity increase of a model at deep runtime.

D ANCHOR ALIGNED MIDPOINTS HAVE NO USABLE DEPTH BAND

For the anchor family there is no depth at which the aligned midpoint is usable: across $R \in \{16, 20, 24, 28, 32, 48\}$ its best case is R=16 at $\geq$19$\times$ endpoint-R=16 perplexity, worsening monotonically to $\sim 2.4\times10^{3}\times$ by R=48, and the naïve midpoint likewise worsens monotonically through R=20–28. Combined with the dense-radius probe for $R\geq$32 (a separate protocol, Appendix O, MB22-D), which finds no usable anchored mixture there, the anchor aligned midpoint is unusable at every probed $R \geq 16$ — the deep-valley dissolution of aligned residuals is specific to the depth-tame loop hosts, not the fixed-K anchor class.

E WIDTH: A FOUR-POINT RELIEF TREND AT TRAINED DEPTH AND GROWING DEEP-RUNTIME INSTABILITY

MB2 scales the fixed recipe across width $d \in \{1024, 1536, 2048, 3072\}$ (four seed pairs) and adds one budget cross-pair, with the law read by a pre-registered Spearman statistic on $(\log_{10} d, \log B_{16})$ and relief declared at $\rho \leq -0.8$; MB23 extends the ladder to n=6 with identical-recipe pairs at $d\in\{1280, 2560\}$, evaluated on the same certified windows and metric, and judged against the extended statistic's pre-registered band.

Trained-depth widths: relief, but barely. The barriers sit at $\log B_{16} = 2.476, 2.432, 2.459, 2.040$ across the four widths – a plateau near 2.45 dex from d=1024 to d=2048 followed by a single -0.42 dex step at d=3072 (Fig. 5). The statistic equals the pre-registered threshold exactly, $\rho = -0.8000$. The shape is a plateau plus one step, not a monotone slope, and a four-point rank statistic is an observation, not a law. Tripling width yields a factor of $\sim$2.7 in trained-depth barrier – small compared with the 10^{14} scale of the depth phenomena (Section 4). The n=6 extension places both new rows inside the plateau – $\log B_{16} = 2.2948$ at d=1280 and 2.3172 at d=2560 (plateau 2.29–2.48, $\pm$0.16 dex of pair-level variance, the same variance class as the budget arm's seed-pair split, Appendix M) – and puts the six-point statistic at $\rho = -0.6000$: the pre-registered band (negative sign, $|\rho| \geq 0.7$) is missed on magnitude while the sign holds; on six pairs the shape is again a plateau plus one step at d=3072, and the width lever on trained-depth connectivity is weaker than the four-point rank suggested. The aligned arm of the ladder, completed at n=6 by the same campaign (alignment added on the four original pairs; Appendix O, MB23-A), oscillates at ρ_a=2.16–2.68 – above the census top 2.1391 on $5/6$ rungs, with the alignment benefit R below the two-sided $+0.3$ floor on $4/6$ (negative on two) – and the exception is again d=3072 (ρ_a=1.5335, R= + 0.5067 in-band, and the only width pair whose aligned deep profile also improves): the relief pair and the alignment-paying pair are the same seed pair.

Sensitivity to the budget clause. The pre-registered $\pm$0.5 dex budget criterion is met: doubling training budget at d=2048 moves the trained-depth barrier by 1.54 dex, from 2.459 to 0.922. The 2$\times$-budget checkpoint is the closest trained neighbor in the whole capacity study – and the same checkpoint produces the worst deep mixture in the census (rod 923.6). With a single budget pair we report the two facts and make no causal claim; the comparison nevertheless shows that improved trained-depth connectivity can accompany reduced deep stability within one recipe. The archived run arguments show that both siblings share seed= 0 – shared initialization and identical data-order streams – so the 0.92 proximity is an *along-trajectory* distance between two checkpoints on one optimization path, not a distance between independent modes: the closest-neighbor reading is therefore a descriptive same-trajectory observation (the deep-side rod 923.6 is unchanged as a same-trajectory measurement).

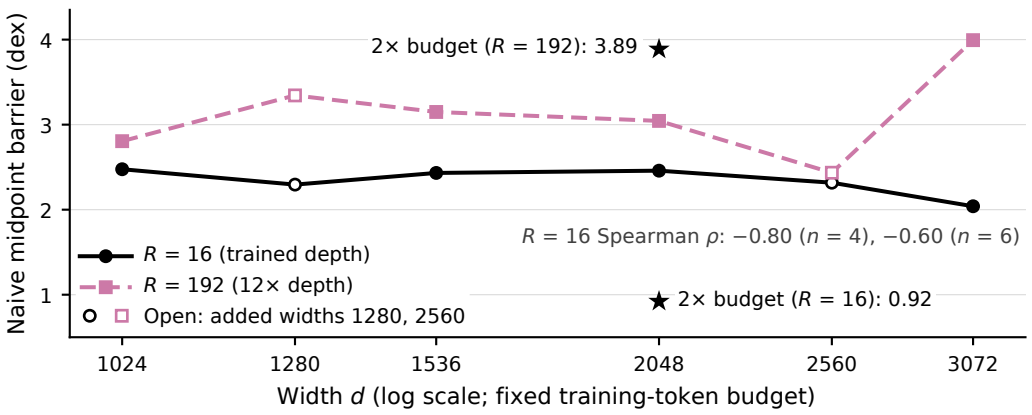


Figure 5: Width ladder at a fixed training-token budget (naive interpolation). Naive midpoint barrier at trained depth (R=16, black, solid) and at 12× depth (R=192, reddish purple, dashed). Open markers are the two added widths (d=1280, 2560). At R=16 the barrier stays on a plateau (2.29–2.48 dex) up to d=2560 and drops once, at d=3072 (Spearman $\rho = -0.80$ on the four main widths, -0.60 on all six). At R=192 the barrier is not monotone in width. Stars: the 2×-budget sibling at d=2048, which shares seed 0 with its partner; it reaches 0.92 dex at R=16 but 3.89 dex at R=192 (its drift ratio, rod 923.6, is the largest measured; Fig. 8b). Points are single pairs; no uncertainty is shown.

Deep instability widens with width (one inversion). Table 2 (Section 5.1) gives the deep profile (rod $1.86 \to 5.15 \to 3.82 \to 89.7$ across the ladder, Fig. 8b – a rising four-point trend, dominated by d=3072; we do not claim monotonicity or superlinearity from four noisy points; the n=6 rungs reorder the deep ladder outright – d=1280 reads logB192 = 3.3438 mid-ladder, while d=2560 is the *most* stable deep mixture measured on any ladder host, rod = 1.245 with logB192 = 2.4345 below d=1024's 2.8057); note that these are the *same* checkpoints whose trained-depth barriers plateau. The six-pair ladder does not establish a general width law: its trained- and deep-runtime profiles are non-monotone, and the shape also depends on matcher choice. Any use of wide loop-LM mixtures (model soups, checkpoint averaging for robustness) must budget the deep axis separately – the trained-depth statistic alone does not establish how the averaged model behaves at greater depth.

The certified values of the four main width pairs and the budget sibling (trained-depth and 12×-depth barriers, endpoint/midpoint depth drift, and the drift ratio rod; budget clause $\Delta = 1.5365$ dex at trained depth with the census-worst deep rod) are given in Table 2 in the main text (Section 5.1).

Matcher dependence. The aligned-arm values above use the census matcher family (G1/G2/G3 Hungarian on $1-|\mathrm{corr}|$); a same-window, same-checkpoint re-measurement of all six rungs under the OT-Sinkhorn arm (Appendix O) instead reads a *monotone* shell contraction ρ_a=2.2638 $\to$ 1.6275 from d=1024 to 3072 (Spearman -1.0000, slope -0.342 dex/doubling, all-positive benefit), against the G-family oscillation quoted above; on identical windows per-rung arm deltas reach $+0.10$ dex, so both shapes are measured upper bounds *at their specified arm* and neither contradicts the other. The ladder's contraction separates from the same-arm census reference ($I_{c,\mathrm{OT}}$=$(1.1916, 1.2642)$ vs width band $(1.9467, 2.3854)$, overlap 0.00), so the trend claims are arm-internal and cross-family transfer of width claims is not assumed (full results in Appendix O).

Scope of the capacity and budget ladder. The ladder spans d=1024–3072 within one recipe family (six widths, one 2×-budget sibling, one matched-seed control), and the budget statement is scoped identically: the 5× result is pair-specific, not a budget law (Appendix M). Two extensions are outside the present scope: (i) re-running the ladder on a second recipe family, which would separate a family-level width law from within-family pair variance; and (ii) mapping the capture–budget curve of the consolidation arm beyond the short-to-mid horizon — the scratch-init trajectory control shows its quality gap still closing at half budget with an unflattened late-window slope (Appendix F.5), so additional budget is a quantitative continuation, not a qualitative regime change. Widening the

ladder is a resource question; it would not decide any separation mechanism beyond those already addressed above.

F WHAT THE DEEP MIDPOINT BECOMES: POSITION-LEVEL TRACERS

Sections 3–4 and Appendix E localize the separation *in weight space and along depth*. MB4 asks the complementary function-space question with three pre-registered probes on the certified windows, each with a pre-specified decision rule (concentration P1, disagreement-tracking P2, ignition onset P3). We examine three hosts: u56 (persistent-deep loop host, aligned midpoint primary), d=3072 (deep-rod extreme, naïve midpoint – the MB4 probes were run before the width pairs were aligned; the aligned width-pair measurements come from MB23 and are reported in Appendix E), anchor01 (aligned deep-valley counterpoint). Aligned measurements re-apply the fixed MB3 permutations, and the permuted endpoints are re-checked for bit-identical predictions in the same run.

P1 – diffuse damage everywhere; no event structure. On all three hosts the worst 1% of positions carry a median share of total midpoint excess CE of 6.1%, 5.8%, and 2.8% (u56, anchor01, d=3072) – every median below the 10% uniform-damage floor, the 25% event threshold never approached in any window population. The residual wall of Section 3 is not the footprint of a handful of catastrophic positions; it is a mild surcharge spread over the whole context (still 3–6$\times$ the exact-uniform $\sim 1\%$: "diffuse" is a threshold judgment, not a uniformity claim). An exploratory read of the full positional profile sharpens the shape (Fig. 6): the excess is position-stationary (participation ratio 0.93–0.99 of the 1024 positions carries the mass), position 0 is below the running median on all 10 midpoint-depth cells, and the level tilts mildly *upward* toward late context – diffuse, slightly late-tilted damage with no onset lesion anywhere.

P2 – damage does not track parent disagreement. The per-position correlation between parent top-64 JSD and midpoint excess is 0.080 (u56, noise level by the pre-specified ≤ 0.1 rule) and -0.143 (d=3072, noise level); anchor01 lands at 0.1014 against the pre-specified 0.1000 noise ceiling and is reported as undecided. Where parents' predictive distributions diverge most, the midpoint is *not* preferentially broken: two seeds are not two algorithms whose interpolation fails at the seams of disagreement; their average is generically noise-damaged even where they agree.

P3 – depth failure is onset-immediate, family-resolved. At twelvefold depth the cumulative-excess breach fires in $16/16$ windows at median position 2 (u56) and 0.5 (d=3072) – the first one to two tokens – versus $3/16$ windows on the aligned anchor (naïve midpoint $0/16$). There is no drift phase to catch: the deep mixture is damaged from its first prediction and stays damaged, on exactly the hosts whose deep barrier persists; and it is quiescent on the host whose aligned residual dissolves into the deep valley.

Synthesis. The three probes characterize where the *measured* separation lives. Trained-depth damage: diffuse (P1), disagreement-free (P2), class-flat after gauge removal (0.048–-0.004 dex gap, basis-dependent sign, Sec. 3). Deep damage: family-resolved (Sections 3 and 4) and onset-immediate (P3). These positional probes do not expose a localized repair target. The failure of the tested repairs is a separate empirical result (Appendix M); diffuse CE damage alone does not rule out a repair family (Appendix M.1, item (d); the direction probe of Section 3.2 shows the damage is not a pure-norm effect – it is elevated in the displacement's own subspace — so a structurally targeted repair remains a live hypothesis, tested for one such family, representation stitching, on the public Pythia pairs, Appendix R).

F.1 RECOVERABILITY UNDER CONTINUED TRAINING: BRIEF TRAINING ERASES THE SEPARATION

Does the separation survive continued training? Brief continued training (AdamW, as in the original runs) from the interpolation midpoints (1500 steps, 30% of the original budget, constant terminal learning rate, host recipe unchanged) erases it on both depth-tame hosts tested: the naïve midpoint ($9.5\,\mathrm{M}/5.3\,\mathrm{M}$ midpoint perplexity at R=16) loses 96.5–96.7% of its log-barrier, and the aligned midpoint loses 90.4–92.0%, with the naïve arm trailing the aligned arm by only ~ 0.06 dex after 500 steps — the gauge shell closes almost immediately under continued training. Endpoint continuation under the same continued-training protocol is drift-neutral (-0.012 dex). For this pair class the catastrophic barrier and the aligned residual are therefore *not* distinct modes but *contact-*

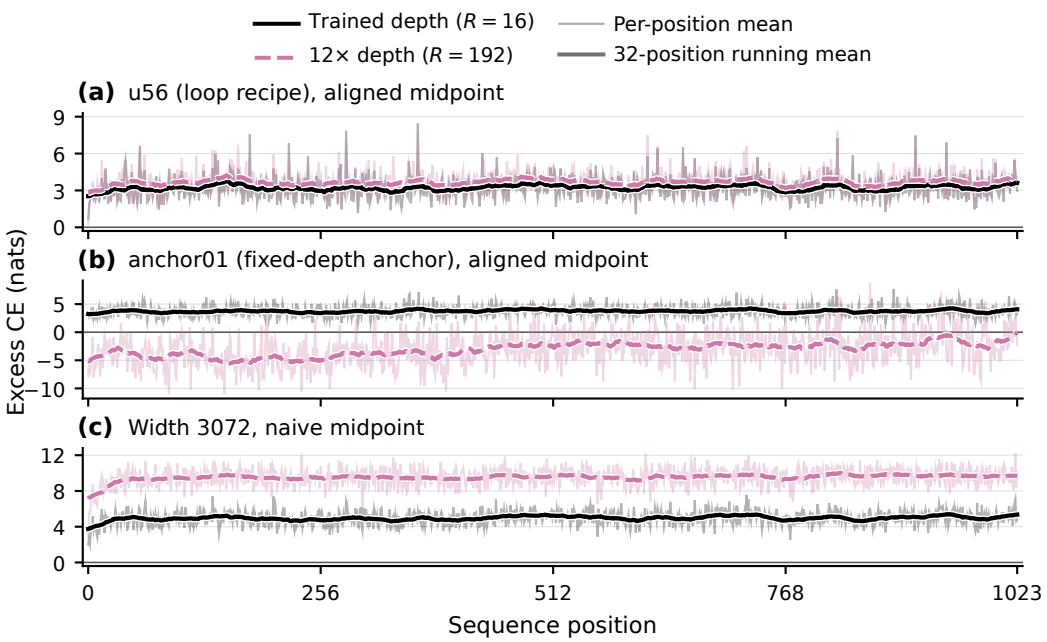


Figure 6: Positional profile of the midpoint's excess CE in nats over the better endpoint (window-mean; each panel has its own vertical scale) on three pairs: (a) the aligned midpoint of `u56`, (b) the aligned midpoint of `anchor01`, and (c) the naive midpoint of the d=3072 width pair, at trained depth (black, solid) and 12× depth (reddish purple, dashed). Thin lines: per-position mean over the 16 windows; thick lines: 32-position centered running mean (a smoothed summary only; the statistics in the text use the unsmoothed profile). The profile is stationary with a mild late tilt, position 0 never the worst; the aligned anchor deep trace is predominantly *negative* (the deep valley) with three late-context positions spiking positive under heavy window scatter, and sits far below its trained-depth profile. No uncertainty band is shown.

erasable norm damage: the zero-training "no gluing" results stand as zero-training facts, while trainable reachability places both endpoints in one basin. The depth axis is what outlives the collapse: recovered midpoints stay near endpoint class at R=192 (1.4–1.7× on u56), and on u40 the aligned-recovered arm retains residual deep damage (4.4–5.4×) that the naïve-recovered arm does not (1.8–2.2×) — host identity persists on the depth axis after the trained-depth separation is gone (*scope: u40/u56 depth-tame lattice pairs, single continuation seed, 30% budget; the anchor family at R=20–28 remains untested*). That deep-state erasure lags the trained-depth recovery host-dependently: on u56 both arms converge to near-endpoint deep behavior by 30% budget, while on u40 the aligned-recovered arm's deep gap peaks after 500 steps and is still ∼5× the endpoint band after 1500 steps — the depth axis completes its recovery last.

F.2 Erasure speed depends on the twin class

Erasure is not one law across pairs. On the TR-class hosts (twins track the aligned residual; classes defined in Appendix O) the naïve midpoint erases immediately: at 4% of the original budget (200 steps) ∼90% of the log-barrier is already gone (remaining fraction $\rho(200) \approx 0.10$). On the TN hosts, whose twin tracks the naïve barrier rather than the residual, the same continued-training protocol erases the barrier more slowly ($\rho(200)$ 0.21–0.26 at the anchor-family hosts, above the pre-registered +0.05 call line against the TR median) — consistent with the pair-specific retained structure that defines TN-class resisting brief continued training — while still converging by 30% (median ρ 0.076). The deep axis does not follow the trained-depth rescue: TN midpoints keep deep deltas of 1.02–1.58 dex at 30% budget, an order beyond the TR hosts (+0.14–+0.30 dex). (k32f, the weaker K=32 host, overshoots its own endpoints from 10% budget while keeping the deepest burden; reported descriptively.)

F.3 THE TWIN BARRIER IS A DOSE LAW, AND NORM IS NOT ITS CARRIER

The twin account supports a dose law over the measured range, and only there: the measurements cover $\lambda \in \{0.25, 0.5, 1, 2\}$ and extrapolation past these edges is not claimed. The construction is exact: the twin at dose λ and noise seed s is $A + D(\lambda, s)$, where D is a per-tensor isotropic draw scaled to $\lambda \|B' - A\|_F$ per tensor (B' the aligned endpoint, norm-matched to 10^{-6}), and the barrier is evaluated at the single midpoint position $A + \frac{1}{2}D(\lambda, s)$ — at λ=1 the displacement magnitude equals that of the aligned A–B' midpoint whose twin-share it calibrates against (a pre-specified replication check ties this λ=1 row to the corresponding census measurement). Varying $\lambda \in \{0.25, 0.5, 1, 2\}\times$ the matched per-tensor displacement budget, the population-median twin barrier grows steeply super-linearly: $+0.082, +0.401, +1.832, +4.480$ dex — super-quadratic through the matched point (share 1.30 at λ=1, replicating the single-matched-point census on the same construction; the full three-regime reading is in Appendix F.4). The magnitude of the residual therefore sits on a norm-dose curve rather than at a coincidental point. Norm predicts magnitude but is not the causal carrier: a per-tensor norm-rescaled (norm-preserving) aligned path keeps a residual statistically indistinguishable from the plain aligned one (median $+1.48$ vs $+1.41$ dex; per-pair share 0.66–1.30) and still grows a deep residual at R=192 on every pair. The two-level reading is therefore: population-wide, any random direction at matched norm reproduces this magnitude (the dose law is content-free); pair-wide, the specific content of the B endpoint still shows through — the trained midpoint retains modest structure its norm-matched twin cannot reproduce, and at low doses the twin barrier is nearly pair-deterministic. The dose law is shared across recipe families (anchors and sampled-U hosts track the same doubling ratios, $\times$4.6–4.9 then $\times$2.4–2.6, with matched-point shares $1.352/1.207$); at low doses the twin barrier is nearly pair-deterministic (0.01–0.03 dex spread), so the trained-depth pair variance observed for ρ_a lies in pair-specific match structure, not in the norm-noise response.

F.4 IS THE DOSE LAW THE ENDPOINT'S LOCAL QUADRATIC? THREE REGIMES

The dose law is not merely phenomenological curve-fitting: whether the endpoint CE Hessian alone predicts it is answered in three regimes. Hutchinson probes of the twin-class direction space — a per-tensor isotropic Gaussian probe at the matched norm budget — yield curvature that is positive on all 224 probed windows (14 pairs, 16 windows each; relative standard deviation $T_{\mathrm{sd}}/|T|$ of the curvature estimate T, median 0.055). At small dose the local quadratic approximates the measured barrier within the stated tolerance: for $\lambda \leq 0.5$ the pre-specified corridor $|\mathrm{pred} - \mathrm{meas}| \leq 0.10$ dex is met (0.042 dex; ratios $0.92/0.78$). At matched dose (λ=1) the pure quadratic underestimates the measured barrier by 0.58 dex (a 0.67 ratio): the true surface grows *faster* than the local expansion — a super-quadratic remainder in the middle regime. At λ=2 the growth slows again: the measured dose ratio sits at 2.50 — the median over pairs of the per-pair ratio of measured barriers at λ=2 and λ=1 (the ratio of the two population medians reads 2.45; §F.3) — well below the pure-quadratic 4.00; the prediction-to-measurement ratios at λ=2 spread 0.70–1.58 across pairs with median 1.06 (sub-quadratic saturation). The endpoint curvature therefore explains the dose law's *onset*, not its full dose response: the small-dose measurements are consistent with a local quadratic approximation and super-quadratic in the matched-dose middle, before saturating.

F.5 ROBUSTNESS: SEEDS, LOSS CONTROLS, INITIALIZATION, AND COMPUTE

Across three seeds at λ=0.5 capture spans only 0.017 dex (0.67–0.69 of the teacher gap) with depth safety preserved (deep-runtime degradation Δ_{deep} $+0.017$–$+0.048$ dex vs $+0.4025$ for the plain continuation); KL alone captures nothing; annealed λ captures more but sharply increases deep-runtime degradation ($+0.5312$ dex); initialization from scratch gives negative capture ($c \approx -1.37$): the deficit is budget, not basin.

Deployed-averaging regime (shared-init bridge). Deployed averaging (WiSE-FT/soups) works on *descendants of one shared initialization*, outside the *from-scratch* setting of the other measurements. Two 2,500-step continuations of endpoint A of the `u56` pair under disjoint data streams bridge the regimes (campaign RE3; record `versions/re3_bridge_verdict.txt`): the naive barrier vanishes on all three induced pairs ($\rho_n \leq 0.0072$ at R=16 vs the census band $[1.28, 1.54]$), the descendant midpoint beats both endpoints, the endpoint–descendant mixture improves by 2.27 ppl, matching is the identity at every site, and depth drift is flat.

Compute accounting. A distillation step costs $5/3$ plain steps: student forward–backward ≈ 3 forward-equivalents plus the mixture teacher's two forwards — the pre-specified FLOPs rule used in Appendix G.1; wall-clock readings are in Appendix O (MB20).

G THE CAPTURE–DEPTH LADDER AND THE PARETO VERTEX

Holding the endpoint-init recipe fixed (constant λ=0.5), capture climbs monotonically with budget: $c = 0.378/0.676/0.888$ (arith) at 5k/12.5k/25k steps, while depth tameness holds only through 12.5k ($\Delta = +0.007/+0.035$ dex) and breaks at 25k ($\Delta = +0.719$ dex). The anneal schedule keeps climbing the capture axis ($c = 1.09$ at 25k) while its depth penalty grows with budget ($\Delta = +0.53 \to +1.73$ dex), and unstabilized CE out-captures the distill arms at the 12.5k and 25k budgets ($c = 0.729/1.203$ against $0.676/0.888$ there; the 5k budget still favors distillation, $+0.378$ vs $+0.222$) while exploding the deep axis earlier ($\Delta = +0.40/+0.58$). Within this ladder the constant-λ 12.5k arm is the unique probed vertex that holds full depth tameness at near-CE capture — the schedule-controlled depth safety of Appendix F.5 is what makes such a vertex exist at all.

G.1 AT MATCHED COMPUTE, PLAIN CONTINUATION LEADS AT TRAINED DEPTH

Under the compute accounting of Appendix F.5 ($5/3\times$ distill steps), plain CE continuation outperforms every measured distillation rung at trained depth under the same compute budget (distillation's trained-depth perplexity advantage over CE, $g_{\text{arith}} = \text{ppl}_{\text{CE}} - \text{ppl}_{\text{distill}} = -0.36/-1.13$ ppl at the 8.3k/20.8k CE-equivalent points; the longer rungs are extrapolation-excluded) — while CE's own deep-delta crosses the tame band before the first matched point ($\Delta_{CE} = +0.23/+0.54$ versus $+0.007/+0.035$ for the distill arms). Endpoint-init distillation therefore buys depth-priority, not quality: its value region is a window where the trained axis is near-unharmed and the deep axis stays tame simultaneously, at a modest trained-depth discount, and the capture ladder above shows how narrow the budget interval is that realizes it.

H ADDITIONAL ALIGNMENT MEASUREMENTS

The following records retain the window, family, and dependence details supporting Section 3. The shell is the reduction achieved by a specified matcher, not an estimate of the globally removable symmetry component.

A shell exists everywhere, and it is the minority dex share on most pairs. Removing the gauge shrinks the barrier on all fifteen pairs, by $R = 0.41$–3.91 dex (population median 1.15; 1.22 on the fourteen discriminating pairs; per-class medians of the per-pair shells 1.00/2.73 anchor/loop; Section 3.2 instead reports the naive minus best-of-four aligned class-median barrier, 1.00/2.80 dex) — a median 48% of a pair's barrier in dex, but the larger share ($\sim 75\%$) on the four deep-loop hosts where the naïve recipe split lives. No pair becomes connected or mild: the classification after alignment is partial everywhere at trained depth. Per-pair magnitudes are window-construction-dependent (second draw: $13/15$ pairs with at least one of the naïve or aligned cells outside its draw-1 band; the aligned census band moves $[1.13, 1.52]$ vs $[1.28, 2.14]$ draw-1); census-level claims are the reported object (Appendix O, window-draw re-measure). To assess reuse of the matching windows, the draw-1 families applied to a disjoint second draw shift the trained-depth residual by a median -0.0058 dex over 14 pairs (median magnitude 0.0346 dex; per-class $+0.0043/-0.0246$ dex anchor/loop), so the same-window selection-bias check reads NEGLIGIBLE (cross-fit cells, Appendix S, Table 8).

The recipe split collapses. On the MB1 class assignment, the trained-depth class gap falls from 1.7556 dex (naïve) to $+0.048$ dex on the single aligned arm, flips sign to -0.043 dex on the best-of-four arm (window-mean basis; Fig. 7, Table 1; per-arm results in Appendix O) — and reads -0.004 dex under the corpus-level geometric-mean basis: $\sim$97% of the recipe-conditioned split lived in the symmetry shell, and the residual's class gap sign-flips jointly with arm and basis choice. The mechanism is visible inside the loop class: shell size is bimodal – four sampled-U hosts carry $R \in [3.50, 3.91]$, exactly the high naïve barriers ($\log B \approx 4.6$–5.2); the remaining four loop hosts span $R \in [0.78, 1.97]$, three of four inside the anchor band $[0.68, 1.38]$ (u32 above it). The observed recipe contrast is strongly reduced by matching; this does not identify the complete permutation quotient or a causal effect of recipe. Counting the two loop ablations in the loop class is conservative:

recounting with only the six sampled-depth pairs *grows* the naïve gap to 2.4849 dex (cf. Section 3.1). Pairing design is confounded with recipe by construction (anchors pair four shared checkpoints, loop pairs are within-host seed pairs); the dependence audits below bound both, so these corrections do not by themselves identify a causal recipe effect.

Dependence audit of the anchor class. The six anchor pairs share four trained checkpoints ($a_0 \dots a_3$); every dependence-aware correction preserves the reduction in the class gap (leave-one-out $[+0.0151, +0.1047]$; matchings $[-0.0107, +0.0912]$; census band $(1.278, 1.544)$ unchanged: clustered $(1.228, 1.546)$, leave-out $\leq (1.292, 1.622)$). The reading is "gone at any scale this population resolves", not "exactly zero".

Four matcher families, one tight band; weight-only matching does not control function. Under four matcher classes (same pairs, windows, endpoint certificates), the component tests pass on every arm ($15/15$ pairs keep residual $\geq 10\times$; split ≤ 0.5 dex), and the activation-free arm produces catastrophic function mismatch (up to $1.8 \times 10^{11}\times$ on u56) – the split lives along the function-informed axis, not in weight space. The weight-only family M1 is degenerate, not a peer: residuals 1.84–11.25 dex, on several pairs *exceeding the naïve barrier*; the meaningful family set is the function-informed trio, whose per-pair minima form the $\log B_a^*$ column of Table 1. Arm choice moves the census level, not its content: the older activation-Hungarian arm's band re-measures at $[1.0972, 1.3840]$ under the best family, every class-level read direction-insensitive to the arm. The set-level verdict stays undecided (stored-precision flags; arm-level stability and the M1 outlier are the content), the head-attribution ladder moves ρ_a by $\leq 0.19/0.42$ dex sign-mixed, and the dense-18 re-test is arm-stable.

Population closure and positioning. Two population functionals close it: the truncation-audited nested-bootstrap collapse share gives *at least* 83% *of the naïve recipe split is gauge at* 97.5% *confidence* (the interval excludes the negligible region), and the norm-matched twins reproduce a median share 1.2884 of the aligned residual (98.7% interval $[1.1336, 1.4310]$, a population-level explanation) – stable across the two-stage bootstrap and every dependence-corrected layer (independent subset, leave-one-checkpoint-out, checkpoint-cluster; all intervals exclude 1, full results in Appendix O; draw 1 passes $13/14$ (exact sign test $p{=}9.2{\times}10^{-4}$, sole sub-1 u32 0.863)), and regraded to median 1.4199 with all fourteen shares > 1 on a second window draw. This is a quantitative sharpening of the perturbation-instability thesis for barriers (Kuditipudi et al., 2019): that account predicts such twins exist; ours are calibrated to the exact per-tensor norm budget of the aligned displacement, reproduce the residual's *magnitude* at population level, and carry a measured dose response, a matcher-stable residual census, and a depth-axis family split no fixed-depth framing expresses (Section 4). The bound is pair-level: the $6/3/3/2$ pair-level split means a third of the discriminating pairs retain structure a pure norm account does not buy – most visibly in the anchor family. Direction matters beyond isotropy: concentrating the *identical* per-tensor dose into the real displacement's top-8 left-singular output subspace raises midpoint damage by a population median $+1.14$ dex over its orthogonal complement ($14/14$ pairs; two-stage bootstrap median-of-medians $+1.10$, CI95 $[+0.66, +2.16]$), while the complement ($+0.006$ dex) and a paired random-frame control at identical concentration and dose ($+1.27$ dex below the real subspace) both land isotropic-like: the increase in damage occurs for dose allocated along the displacement's own low-rank output subspace, not dose concentration in general.

I Stronger matchers and trained-depth twin controls on the decomposition

The two paragraphs below give the full results of the matcher battery (MB15) and the trained-depth twin control (MB12) summarized in Section 3.

Stronger matchers reproduce the split; weight-only matching does not control function. The decomposition is re-measured under four matcher classes on the same fifteen pairs, windows, and endpoint certificates (MB15 campaign: M1 activation-free weight matching à la re-basin Hungarian on concatenated unit matrices; M2 OT-Sinkhorn activation matching, $\varepsilon{=}0.05 \times 200$ fixed iterations, hard projection; M3A the G2 head-class dropped; M3B the G2 class weight-replaced). The pre-specified component tests pass on *every* arm: on all four matchers $15/15$ pairs keep post-alignment residual $\geq 10\times$ (shell-size gate likewise, collapse share ≥ 0.5, split ≤ 0.5 dex: M1 $+0.459/0.739$, M2 $-0.043/1.025$, M3A $+0.037/0.979$, M3B $+0.035/0.980$). Under the function-informed matchers

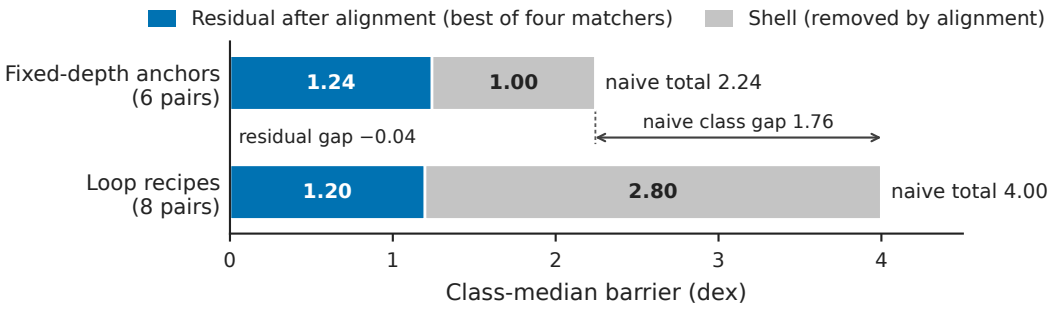


Figure 7: Trained-depth barrier decomposition as differences of class medians of per-pair log-barriers (best of four matcher families; shell median 1.22 dex on the fourteen pairs). Rows: fixed-depth anchors (top) and loop recipes (bottom), as in panels (a) and (b) of Fig. 2. The naïve barrier (full bar length) splits into the measured post-alignment residual $\log B_a^*$ (blue; best of four matchers) and the shell R_{sh} removed by alignment (gray). Annotations give the naïve class gap (1.76 dex) and the residual class gap (-0.04 dex), both as differences of class medians; on the window-mean basis the class gap falls from 1.76 dex to -0.043 dex under the best-of-four matcher. Bar segments are differences of class medians, not medians of per-pair shell sizes; no uncertainty is shown.

the residual census reproduces at the M0 magnitudes ($\rho_a \in [1.10, 1.95]$, residual 12–89×, shell positive, post-match G1/G3 correlation $\sim 0.30/\sim 0.36$, and the weak-G2 signature 0.02–0.04 reappears unchanged); the activation-free matcher by contrast produces catastrophic function mismatch (residual up to $1.8 \times 10^{11}\times$ on u56, shell *negative* on $11/15$ pairs, ρ_a up to 11.25): alignment quality is the controller, and the split lives along the function-informed axis, not in weight space alone. The pre-specified set-level verdict is AMBIGUOUS (stored-precision flags: one M2 Sinkhorn plan row/column error above 10^{-6} at the 200-iteration convergence floor, and the dense-leg M3 index arrays not bit-identical to the archived control permutations) – the arm-level stability above and the M1 outlier are the substantive results; the set-level label does not change either. The G2-head-attribution ladder answers its own question: dropping or weight-replacing the weakest-matched class moves ρ_a by ≤ 0.19 and ≤ 0.42 dex respectively across all fifteen pairs, sign-mixed – the residual is not a head-class artifact. The dense-18 control re-test is arm-stable (all four arms DENSE-RESIDUAL-LOOP-LIKE, $\rho_a \in [1.29, 1.86]$).

Does the residual survive a norm-matched random twin at trained depth? Partially (verdict MIXED-VERBATIM: no pre-specified majority class). Extending the deep-drift twin control to the trained-depth midpoint (MB12; two noise seeds per pair; every twin norm-matched to its pair's aligned endpoint budget on all fifteen pairs), the pre-specified class split lands at $6/14$ TRACKS-RESIDUAL, $3/14$ TRACKS-NAIVE, $3/14$ TRACKS-BOTH (the corridor class; all at host gaps < 1.0) and $2/14$ AMBIGUOUS (seed splits) – neither pre-specified $\geq 13/14$ majority is reached, so the residual is uniformly accounted for by neither the norm budget nor retained structure. The twins' own residual concentrates at 1.62–2.00 dex on every pair (residual census $[1.28, 2.14]$) and falls inside the pre-specified band of the aligned residual on $20/28$ discriminating-pair seeds; they track the aligned residual in both seeds exactly at the six largest naïve-vs-aligned host gaps (1.06–3.63 dex). Reading: where the naïve path is catastrophically worse than the aligned one, the trained-depth residual is *largely a norm-budget account*; across anchor-family gaps the residual retains structure the budget does not buy. The full fifteen-pair per-seed results are in Table 4 (Appendix O).

J FACTOR-RAMP CENSUS AND A QUANTITATIVE DRIFT MODEL (MB21)

The two paragraphs below give the full results summarized in Section 4.

The factor ramp: component-wise ignition, breach-immediacy factor-invariant. The 12× slice (R=192) of Section 4 is one certified factor of a five-point ramp bracket $R \in \{64, 96, 128, 160, 192\}$ – the three intermediate factors measured for this ramp, the two end factors taken from the existing R=64 and R=192 measurements. Three facts extend the single-factor picture. (i) *Immediacy is not a* 12× *artifact.* The first-breach-immediacy statistic (Appendix F) is factor-invariant across the full

ramp on every host where depth damage exists ($3/3$ certified-ramp hosts factor-stable, max factor ratio 0.368 in $\log_4$ units): the deep mixture is first-token damaged at R=64 exactly as at R=192. (ii) *Ignition onset is host-specific*: over the eleven ramp census cells the six anchor cells ignite in $(64, 96]$, u24 in $(96, 128]$, u32 is late in $(160, 192]$, and u40, u56 and the capacity hosts do not ignite by R=192 – an onset multiset $\{6 \times (64, 96],\ 1 \times (96, 128],\ 1 \times (160, 192],\ 3 \times \text{absent}\}$ that pins, per host, which ramped-average schedules can exist at all (Appendix M). (iii) *The ramp shapes split into distinct regions rather than one law*: over the thirty host-factor shapes (10 hosts $\times$ the three fresh factors), $26/30$ are bounded (R-floor – flat within the certified window protocol – or R-saturation on already-breached shapes) while $4/30$ are R-acceleration, u24-led, where breach onset lands *inside* the ramp. The pre-registered monotone-concavity hypothesis for the ramp (a smooth, universal onset curve) is not supported by this census; any depth-contraction statement is region-conditioned, not ramp-universal. The single-factor $12\times$ results above hold at that factor, and the ramp extends them to the other factors.

A quantitative drift model (MB21): host-constant self-alignment and a radius-ignition band. The deep drift above admits a compact descriptive account in two host-level quantities, the self-alignment α and a relative-radius ignition band. Instrumenting a certified window through every core unroll (radius r_k, core-output scale $F_k = \mathrm{rms}(f_k)$, and self-alignment $\alpha_k = \cos\langle h_k, f_k\rangle$ for five hosts $\times$ {parents, naive midpoint}): (i) α is a *host constant*, not a depth switch – the pre-registered ignition-switch hypothesis is falsified ($|\alpha| > 0.3$ from the first unrolls whenever positive at all); (ii) its sign sets the class: endpoints of depth-tame families sit at $\alpha \approx 0$ ($-0.17 \ldots + 0.09$, a self-neutralizing recurrence) and the d=1024 anchor parents at $\alpha < 0$, while deep-runaway parents live at $\alpha \in [0.52, 0.77]$ and every naïve midpoint is born locked at $\alpha \in [0.54, 0.87]$; (iii) the slope law $r_{k+1} = \sqrt{r_k^2 + 2\alpha F r_k + F^2}$ agrees closely with the measurement: the u56 naïve midpoint's measured slope 0.1419/unroll matches $\alpha F = 0.144$ within 1.46% (single-host measurement; the law class holds across the tame hosts, with host-specific slopes; the next paragraph states what this agreement does and does not establish); (iv) dense $\mathrm{CE}(R)$ curves complete the description with a *relative-radius ignition band*: the CE ignition of all four deep-runaway parents co-locates (within $\Delta R \approx$ 10–25 unrolls) with the radius crossing $\rho^* \in (1.65, 2.3) \times r_{16}$, while $\alpha \leq 0.1$ parents never cross it – the host-specific onset classes of the ramp census coincide with crossings of this radius band. Open questions: the ignition criterion's sensitivity, the mild upward curvature of u24/u32 trajectories (α is only approximately depth-constant there), and the inverted–decreasing–CE-versus-radius trajectory of born-catastrophic midpoints; the onset-class bands are therefore a radius-threshold account of what ignites, not yet a first-principles predictor of the census numbers. A dense-radius probe on the anchor01 parents (Appendix O, MB22-D) shows that the anchored family has no usable depth region: both parents ignite early (R_{ig}=$62/72$) and their aligned midpoint ignites at R_{ig}=32 – its radius crossing ρ=1.826 sits inside the $(1.65, 2.3)$ band, with the early-arc calibrated closed form landing within ± 2.3 unrolls on the parents and the aligned midpoint alike (this confirms the early-arc calibration: the post-ignition arc inflates the mean-αF model); no anchored mixture reaches the usable-quality region before crossing the ignition band.

The depth-drift and width measurements are plotted in Fig. 8 (Appendix K); absolute endpoint and midpoint quality at depth is in Fig. 3 in the main text.

Radial-flow descriptive model (a parallel micro-measurement). For a residual update $h_{k+1,p} = h_{k,p} + f_{k,p}$ without internal normalization, positionwise RMS radii $r_{k,p} = \|h_{k,p}\|_2/\sqrt{d}$, update norms $F_{k,p} = \|f_{k,p}\|_2/\sqrt{d}$, and angles $\alpha_{k,p} = \cos\langle h_{k,p}, f_{k,p}\rangle$ satisfy exactly $r_{k+1,p}^2 = r_{k,p}^2 + 2\alpha_{k,p} r_{k,p} F_{k,p} + F_{k,p}^2$. Recorded trajectories average r, F, α over positions; substituting those means into the nonlinear recurrence is an approximation, not the same identity. On u56 the measured mean-radius slope is 0.1419/unroll versus the descriptive $\bar{\alpha}\bar{F}$=0.144: this near agreement does not identify a constant-gain mechanism or justify interpreting an RMS measurement as an L2 norm. The alignment ranges and radius-associated CE transitions here are host-conditioned associations; thresholds fitted to them are not independent onset predictions. For the anchored pair the aligned midpoint already deteriorates at R=32, so relative improvement against exploding parents does not establish usable deep models.

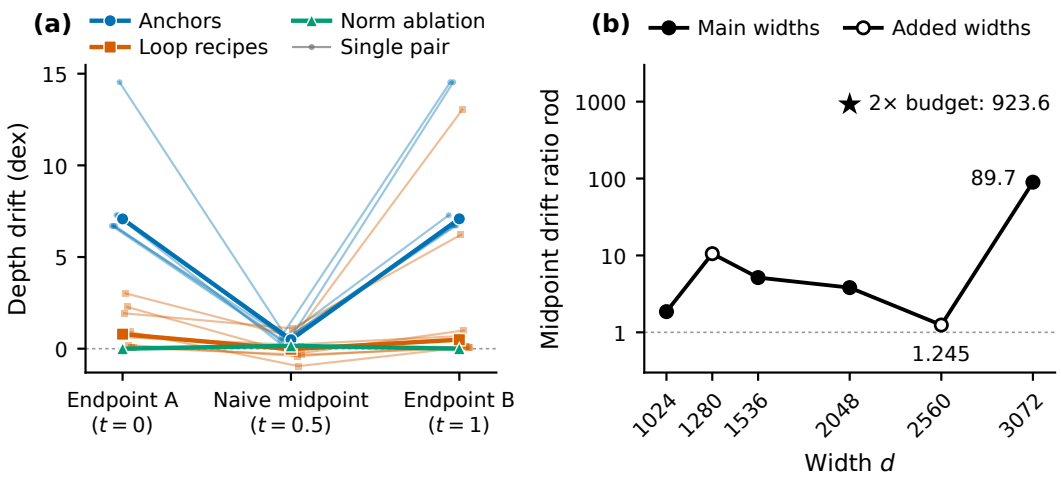


Figure 8: Runtime-depth diagnostics. **(a)** Depth drift $\log_{10}(P_{192}/P_{16})$ in dex of endpoint A, the naïve midpoint, and endpoint B (0 dex means no change); points are window-median perplexity ratios at R=192 versus R=16. Thin traces are single 35 M pairs; bold traces are class medians over them (summaries, not uncertainty bands). Colors: anchors (blue), loop recipes (vermillion), norm ablation (green). Anchor endpoints can deteriorate far more than their midpoint, so a low relative barrier against them does not establish usable absolute perplexity. **(b)** Midpoint drift divided by the larger endpoint drift (rod; logarithmic axis) across all six widths; filled circles are the four main widths and open circles the two added widths (d=1280, 2560). Their non-monotone ordering includes d=2560 at 1.245, compared with 89.7 at d=3072. The black star (923.6) is the 2×-budget sibling at d=2048; it compares same-initialization checkpoints at different training budgets, rather than independent runs. Dotted gray lines mark no drift (0 dex in (a), a ratio of 1 in (b)).

K Detailed runtime-depth and positional measurements

Trained depth measures the midpoint as a sixteen-iteration model; the deep probe unrolls the same three weight points twelvefold.

Drift dominance (exploratory depth atlas). For anchor pairs the *parents* explode under twelvefold unrolling (median endpoint drift 4.97×10^6–3.43×10^{14}) while their naïve fifty-fifty midpoint drifts only 1.2–5.2× (rod $\approx$ 0; Fig. 8a): the midpoint is orders of magnitude *more* depth-stable than either independent training run it averages. The claim is qualified: norm-matched twins (Sec. 3) reproduce drift sign and magnitude class on both families (not discriminated by these controls), and every R=192 anchor model is unusable in absolute perplexity (the "valley" is the ratio to exploding endpoints). The depth-tame loop hosts show the opposite degeneracy: endpoints are already depth-tame, no discount to harvest, and the barrier persists depth-invariantly (corr 0.77–0.97 across the five tame ladder hosts); the norm ablation is exactly depth-invariant (D_A=D_B=1.000).

Alignment sharpens, not erases, the split. With the gauge shell removed (Sec. 3), the anchor residual dissolves below the mild line at depth (corpus-geometric median $\log B_{a,192} = -4.11$): aligned anchor midpoints have low relative barriers against deteriorating parents, consistent with drift dominance. The loop class's residual persists at depth (corpus-geometric median +1.53). The depth axis re-instantiates a family distinction, in the measured post-alignment residual.

Capacity inverts the 35 M prior. At width the atlas flips on all four capacity hosts (Table 2): deep barriers exceed trained-depth barriers ($\log B_{192} = 2.81$–3.99 vs $\log B_{16}$), endpoints stay depth-tame ($D_A/D_B \in [1.02, 1.47]$), and the midpoint outlier ratio rises to 89.7× at d=3072 (rod 1.86, 5.15, 3.82, 89.7); the 2×-budget sibling is at once the closest trained-depth neighbor and the worst deep mixture (rod = 923.6; Sec. 5.1).

Where the deep failure begins. The MB4 first-breach probe localizes the deep failure in the sequence: on persistent-deep hosts the cumulative excess of the deep midpoint crosses its threshold at median position 0.5/2 (d=3072/u56) in all sixteen windows – deep damage is present from the first

token, no ignition phase (the crossing is magnitude, not event). On the aligned anchor the breach almost never fires ($3/16$ windows) – the tracer-level shadow of the deep valley, positionally diffuse.

The factor ramp, summarized. A five-point ramp bracket $R \in \{64, 96, 128, 160, 192\}$ shows: breach-immediacy is factor-invariant (max factor ratio 0.368); ignition onset is host-specific (anchors $(64, 96]$, u24 $(96, 128]$, u32 $(160, 192]$); the thirty host-factor shapes split $26/4$ bounded/R-acceleration, refuting the pre-registered universal-onset concavity; the $12\times$ claims apply at that factor (Appendix J).

The usable deep axis on depth-tame hosts: the barrier weakens, the *aligned* component persists. On the two depth-tame loop families the deep axis is measured at *usable* quality: a ten-factor additive grid ($R \in \{16, \ldots, 192\}$) on the mixture grid, every grid endpoint inside the pre-specified usable band out to R=192 (range 60.2–80.1) – exploitable models, not 10^4–10^{16} catastrophes. Three facts. (i) *The naïve barrier weakens with usable depth*: $\rho_{192} - \rho_{16} = -0.395$ (u56), -0.919 (u40), with all additive small-step deltas $R16 \to R32$ negative and monotone per host – the smaller increments show the same direction as the wider depth comparison. (ii) *What survives is the aligned, non-gauge component*: the residual grows monotonically from $+1.318$ (R=16) to $+1.493$ dex (R=192, $\approx$ 31–32$\times$ excess midpoint damage) even as the naïve barrier compresses. (iii) *Breach geometry is depth-stable*: threshold crossing immediate, the aligned arm's damaged-position share (0.36) below the naïve arm's (0.60–0.72) – gauge removal protects position coverage, not residual inflation. Depth choice and the tested alignment change different measured aspects of this comparison; the measured residual persists after the tested alignment.

Support is graded, not binary. Training new hosts tests the role of the training-depth support: cloning the clean fixed-$K16$ anchor recipe with only the depth sampler changed to a mild jitter ($K \sim U\{12\ldots20\}$, two seeds, identical budget) moves the hosts most of the way but not into the tame class. Endpoints stay tame in-support and just past the support edge (D_{32}=1.03/1.05, D_{64}=1.32/1.74) and degrade gradually at the $12\times$ probe (D_{192}=22.9/17.2) – an intermediate regime: an order above the depth-tame band and orders below the anchor explosions. The deep midpoint still reads anchor-style, its naïve barrier dropping 2.19→1.64 dex at depth, while the shell-resolved residual is already loop-class (aligned ρ_a=1.51 dex). Usable depth is a graded function of the recipe's depth support – and the fixed-K explosion is an observed recipe-class difference, not an artifact of midpoint construction.

L CAPTURE CURVE, PAIRED CONTROL, AND REPLICATION DETAIL FOR THE CONSOLIDATION

The paragraphs below give the full results summarized in Section 6; `d12`/`c12` and `d25`/`c25` denote the distillation and plain cross-entropy (CE) arms at 12.5k and 25k steps.

Capture curve and the depth dual account. Quality capture toward the teacher, measured on the certified windows (R16, arithmetic / geometric): 1k steps: $38.553/37.317$ (capture -6% – the warm-restart transient dominates); 5k: $37.330/36.110$ ($+37.8\%$ / $+36.8\%$, already 2.7% below A); 12.5k: $36.498/35.235$ ($+67.6\%$ / $+68.7\%$), on a pre-registered prior band $[0.44, 0.65]$ whose log-linear prediction (0.628) the measurement exceeds by 4.8 points – the curve rises concavely with no plateau to 12.5k. The paired control exposes the second account plainly: at 12.5k the pure-CE arm reaches $36.351/35.046$ – it matches the student at R16 while its depth behavior *collapses*: R192 $91.85/50.52$ (sd 156.5, a $2.53\times$ trained-depth pole) against the student's $39.52/37.50$ ($1.08\times$, the tame 1.02–1.04 endpoint class). The probability teacher therefore has a dual effect: it transfers part of the ensemble advantage at sub-endpoint budget *and* suppresses the long-horizon depth runaway that plain continued training develops – the same pole geometry this paper isolates in Section 4. At 25k the curve closes at $+88.8\%$ / $+89.3\%$ ($35.908/34.669$) against a pre-registered prior band $[0.82, 0.90]$ – the student reaches the teacher's neighborhood exactly at the pre-registered budget. The paired control shows where this advantage ends, and the boundary is sharp: at full budget the pure-CE arm ($35.030/33.741$) passes *through* the teacher (its 33.741 geometric sits below the teacher's 34.375), so the fixed-λ KL term anchors the student to a target the arm would otherwise surpass, and the control wins on every axis, depth included – both arms leave the endpoint tame class at 25k (R192 $134.08/53.56$ vs $188.06/59.30$), making the 12.5k depth protection a bounded, mid-budget phenomenon. These results delimit the useful regime of probability-teacher distillation: it is effective at short-to-mid adaptation budgets, capturing 38%–68% of the ensemble advantage

on the 5–12.5k rungs while preserving endpoint-class depth behavior — at those rungs the matched pure-CE control reaches the same or higher trained-depth shares ($22.2\%/72.9\%$) yet develops a $2.53\times$ deep pole from 12.5k — and plain continuation leads in the full-budget regime where fixed-λ distillation pays an anchoring penalty – a reading converted into a tested claim by a falsifier arm: annealing λ linearly to zero over the same horizon recovers 65% of the fixed-λ deficit (35.334 against control 35.030, arithmetic basis; 66% on the geometric basis, 34.052 against 33.741), localizing the anchor tax as the majority account, and no annealed variant repairs the full-budget depth regime ($1889/154$ whole-NLL at R=192, the family's worst); the deployment claim is a single 35 M model at adaptation cost far below an endpoint training run (teacher forwards are a one-time training-side cost, their share stable at $45/47/47/47\%$ of student compute across the four rungs). On a held-out corpus (WikiText-103 windows, never in training), the same protocol replicates the orderings with an out-of-distribution discount: the teacher beats the stronger endpoint by 11.2% (stronger than in-corpus), student captures run 40–76%, the mid-budget KL arm stays depth-tame ($1.03\times$), and the full-budget continuation eruption replicates while the mid-budget control's two-window poisoning does not – the latter is corpus-specific, the former is model-intrinsic (App. O). A function-space control beyond CE answers how much of that recovery is memory of the norm budget: on the u24 pair the trained aligned midpoint still beats its norm-matched twin by $+1.16$ nats of CE with $1.46\times$ smaller symmetric-KL distance to the endpoint – real but modest retained structure on top of the norm budget (App. O, MB22-C).

MB38 second-pair replication. This campaign follows the pre-registered specification `versions/MB38_second_pair_consolidation_spec.md`. Endpoints: the two u56 models (runs `e69_u56` and `dense_u56_s2` in Table 3), A at 61.2086 (relative deviation from the stored value 1.15×10^{-5}) and B at 60.1861 (1.31×10^{-5}), re-checked against the MB1 measurements; teacher $q_{0.5}$ $56.4659/54.8133$ (arith/geo, advantage $7.75\%/8.02\%$ vs A; at $R192$ teacher $66.11/64.40$ below both tame endpoints; WikiText-103 $133.27/127.03$ vs $153.86/146.28$, 13.4%), Jensen pointwise gaps ≥0 with minimum $+0.043$ nats (all-position dominance). Arms at 2,500 steps ($0.5\times$ the pair's own budget), λ=α=0.5, seed 3, paired stream: distill $54.6771/53.1547$ R16 (captures $+1.3772/+1.3471$), plain-CE control $52.5996/51.1376$ ($+1.8152/+1.7693$); depth ratios $R192/R16$ $1.1905/1.1931$ vs $1.3011/1.3016$ (both tame, no pole). WikiText-103 transfer: $128.9487/122.9449$ vs $121.8126/116.3970$ R16; ratios $1.3577/1.3555$ vs $1.3975/1.3996$. Verdict PAIR2-INTERMEDIATE (`logs/mb38_verdict.txt`): both arms overshoot the much weaker teacher, so the paired control leads from the same rung on R16 on both corpora; the train-time teacher keeps only a ratio-level depth edge. The short-to-mid distillation advantage is thereby bounded to the large-gap strong-pair regime; plain continuation's lead domain extends to close weaker pairs from the vertex rung.

MB39 vertex release. A fresh 25,000-step run, under the pre-registered specification `versions/MB39_teacher_release_spec.md`, shares the d25/c25 learning-rate schedule, data stream sampler and per-step R draws, with the KL weight stepping λ $0.5\to0$ at the 12.5k vertex (post-switch teacher forwards skipped: teacher cost 2965.4 s, half of d25's full 5987.6 s). The released run at 25k steps evaluates to: FineWeb-Edu R16 $35.3894/34.0971$ (arith/geo), R64 $35.3981/34.1023$, R192 $5181.3108/223.7885$ (ratios $146.41/6.56$); WikiText-103 R16 $66.6147/64.3296$, R192 arith 1079.42 med 108.36 p90 3804.06 max 4892.27. Against the pre-specified comparators (c25 R16 $35.0299/33.7407$ R192 $134.08/53.56$; d25 R16 $35.9082/34.6692$ R192 $188.06/59.30$; d12-vertex R192 ratios $1.083/1.064$): quality strictly between d25 and c25 at R16 (quality predicates: better than d25, true; better than c25, false), depth tameness absent at R192 (depth predicates against c25 and d25 both false). Verdict RELEASE-FAILS (`logs/mb39_verdict.txt`): the vertex release halves the teacher cost but provides neither the continuation's R192 tameness nor the distill arm's endpoint quality; the full-budget boundary stands as measured.

M TESTED REPAIRS, IMPLICATIONS, AND SCOPE

Scope of the measured repairs. The following measurements delimit the tested interventions. (i) The naive four-member soup is catastrophically damaged ($q = 1179.55$) and permutation alignment does not rescue it ($q_a = 1058.70$): multi-member uniform averaging fails both with and without alignment. (ii) The two-member residual after alignment is class-flat: aligned class gap $+0.048/-0.004$ dex (window-mean/geometric), point-estimate-zero, basis-dependent sign against

the 12.5–$24.2\times$ band, so this contrast does not identify a recipe-specific repair target. (iii) Trained-depth damage is positionally diffuse ($< 10\%$ share in the worst 1% of positions), which provides no evidence for a small set of damaged positions, but does not exclude function-preserving repair. (iv) The damage does not concentrate where parents disagree, so disagreement-localized stitching – the natural candidate gluing principle – is not supported by these positional tracers; richer representation-level stitching remains possible. (v) Channel-moment renormalization – the natural repair baseline, which rescales each activation channel of the aligned midpoint towards the average-parent first/second moments – does not restore connectivity on any of three representative hosts (a deep-valley anchor at ignition depth, and two persistent-deep loop hosts): on both repair arms (residual-stream only, and residual-stream plus attention/MLP-internal channels) the repaired midpoint is *worse* than the unrepaired one by 1.23–1.84 dex on every host (verdict REPAIR-DOES-NOT-CONNECT under the pre-specified band), and the same failure persists with the standardization-saturation artifact absent. The diffuse deficit is compatible with distributed perturbation damage; it does not establish that every repair method must fail. Scope note: these results reject the measured configurations – uniform souping and activation-matched permutation alignment (plus disagreement-targeted stitching, which has no measured site, and channel-rescaling repair, which is measured and harmful here). Representation- or circuit-level repairs are not bounded by our probes (Appendix M.1, item (d)).

The depth-schedule hypothesis. The different runtime-depth profiles motivate testing depth-ramped interpolation, but do not imply that all repairs must control depth. The measured onset bands are anchors $(64, 96]$, u24 $(96, 128]$, and u32 $(160, 192]$; u40, u56, and the capacity hosts do not ignite by $R{=}192$. These bands defined the tested schedules. Their failure on two persistent hosts, reported below, limits those schedules rather than all possible interpolation paths. Norm-matched twins also reproduce the deep-drift sign and magnitude classes, so drift alone does not identify a loop-specific repair mechanism.

Budget does not rescue connectivity either. The pre-registered stronger-endpoints test re-measures the decomposition on fresh u56-recipe pairs at $2\times$ and $5\times$ budget ($10{,}000/25{,}000$ steps; fresh seeds $5/6$ and $3/4$, disjoint sampler and validation streams; identical windows, grids, and scoring rules; Section 2), giving a three-rung dose ladder at fixed architecture. Endpoint quality improves monotonically ($61.2/60.2 \to 47.8/47.0 \to 38.4/38.4$) and is uninformative in both directions: the naïve trained-depth barrier is *non-monotone* ($4.94 \to 4.27 \to 7.99$ dex – the $2\times$ rung shrinks the shell share), while the gauge-resolved residual *amplifies at every rung* ($1.32 \to 1.78 \to 9.51$ dex; at $5\times$ the aligned midpoint even exceeds the naïve). Onset-immediacy is class-invariant (breach medians 3.0–3.5 naïve / 5.0–5.5 aligned at $2\times$); the $5\times$ verdict is V1-TRANSFER-CONFIRMED and the $2\times$ rung is a descriptive middle rung (Appendix O). An independent-seed control at the same $5\times$ rung (fresh seeds $7/8$, disjoint streams; endpoints $38.7/38.3$) instead reproduces the principal family: naïve ρ 4.03 dex, aligned ρ_a 2.15 dex, alignment benefit $R{=}+1.88$ dex inside the two-sided band, and an aligned midpoint $75\times$ *better* than the naïve one – so the 9.51 amplification and the aligned-above-naïve inversion are pair-specific outcomes of that ladder's $5\times$ member, not budget-axis laws (Appendix O, MB23). Budget therefore moves both axes and not in one direction: at $d{=}2048$ dense anchors a $2\times$ budget buys 1.54 dex of trained-depth relief (the same-trajectory sibling, Appendix E), while at 35 M sampled-U the gauge-resolved separation grows at every measured rung. Endpoint quality is not a proxy for connectivity: the axis is pair-variable, not a deterministic law.

Depth-ramped averaging does not restore connectivity. On both depth-stable pairs (u56, u40), every tested depth-ramped average – a linear or quadratic runtime-depth ramp $R(t) = 16{+}(\mathrm{Rmax}{-}16) \cdot (2|t - \frac{1}{2}|)^p$ with $\mathrm{Rmax} \in \{64, 192\}$, hosted drift-free by construction and onset-capped per the measured census – leaves the pre-specified usable-window band on the mid-window cells, while *every ramp endpoint at* Rmax stays inside the band (window-max ≤ 146.5 against bands $245/241$ and the $\times 100$ spike guards). The failure is not a too-shallow effect a ramp can repair: the mid-window perplexity *worsens monotonically with runtime depth* – at $t{=}0.25$ the window-mean perplexity rises from 353 at $R{=}16$ to 814 at $R{=}192$ on u56, and from 462 to 1679 on u40 ($t{=}0.75$ mirrors). These runtime-depth schedules therefore do not restore connectivity on the depth-stable pairs at 35 M (MB11). In the onset-capped anchor control, the $t{=}0$ endpoint cell under its $\mathrm{Rmax}{=}64$ ramp also leaves its band (254.4 vs 229), so ramp traversal on the anchor family is not uniform in t; this control is descriptive and does not change the result above (Appendix O, MB11).

Scope. The scope of these measurements is stated item by item in Appendix M.1, and Appendix U discusses the limitations of the study as a whole.

Related strands. Mode connectivity and its loss-barrier phenomenology (Garipov et al., 2018; Izmailov et al., 2018; Kuditipudi et al., 2019; Frankle et al., 2019); permutation-alignment methods for weight-space matching (Entezari et al., 2021; Ainsworth et al., 2022; Jordan et al., 2022) – the shell we resolve is the loop-architecture analogue, with tying-specific constraints; model-soup and weight-averaging practice (Wortsman et al., 2022; Ilharco et al., 2022) – the naive and aligned soups above bound where uniform averaging can work on this family; weight-tied looped/universal transformers (Dehghani et al., 2018; Giannou et al., 2023) – our depth axis is the native complexity axis of this family, of which trained-depth numbers are single slices; and compute-scaling analyses of barriers – our width trend is boundary-exact at $n{=}4$ and misses its own pre-registered band at $n{=}6$, paired with a deep-axis cost that scaling studies typically do not instrument.

M.1 Scope and limitations in detail

This subsection states the scope of the individual measurements; Appendix U discusses the limitations of the study as a whole.

Scope by measurement. (a) Population: fifteen pairs at 35 M, six width (capacity) pairs, three tracer hosts; single-seed cross pairs at each width; one budget pair (no causal budget claim). (b) The gauge resolution covers three pre-specified permutation classes; the decomposition was also re-measured against stronger matcher classes on the identical census (Section 3): the function-informed matchers reproduce the shell/residual magnitudes on every pair, while the activation-free weight matcher does not control function (residual up to $1.8 \times 10^{11}\times$) – alignment *quality* is the controller. The set-level verdict of that campaign is AMBIGUOUS (stored-precision flags); permutations outside the measured classes remain unmeasured, and a broader search could reduce the best-achieved residual; the random-twin controls do not bound that optimum. This insensitivity replicates at $\approx 2.5\times$ scale: at Pythia-410M the naïve barrier narrows to 2.0542 dex and no named family removes any of it (shell share 0.0000, verdict REPLICATE, Appendix R). (c) The breach geometry reported here sits at extension factor $12\times$; the dense grid plus a four-factor ramp ($64 \ldots 192$) is measured in Section 4 and Appendix J: breach immediacy is factor-invariant on the persistent hosts (max factor ratio 0.368 in $\log_4$ units), ignition onset is host-specific, and any contraction claim is region-conditioned. In absolute perplexity every $R{=}192$ anchor-family model is catastrophic (Section 4), so the deep "valley" is a relative barrier statement against exploding parents, not a claim of usable deep models; u24's extreme aligned value on the arithmetic-window basis (-8.85 dex; Table 1 reports -1.80 dex on the corpus basis) is aggregation-sensitive (window-median $+0.21$) and is reported as measured; on the same arithmetic-window basis, excluding u24 shifts the loop-family deep median from $+1.62$ to $+1.75$ ($n{=}7$; the corpus-basis median over all eight is 1.53 dex) – the anchor-vs-loop deep split does not rely on the outlier. The deep drift itself is controlled, not loop-idiosyncratic: same-per-tensor-norm random averaging twins reproduce both its sign and its magnitude class on the persistent and valley hosts alike (drift compression on anchors, inflation on tame hosts; residual calibration 0.2–0.8 dex, AMBIGUOUS under the pre-specified band) – the deep-unroll drift is predominantly a weights-space smoothing/norm-budget account, not a loop-recurrent-specific one. Channel renormalization at deep factors saturates to uniform output ($\mathrm{ppl}{=}$ vocab) on two of three hosts – a protocol consequence of simultaneous per-slot marginal standardization, not a measured confound. (d) Tracer probes measure CE damage and parent-JSD conditioning on sixteen windows; richer functionals are complementary, not covered. (e) The depth-schedule hypothesis is tested and dead on the persistent family (RAMP-DEAD, above; the onset-capped control is CONTROL-INCONSISTENT); the function-space branch lands: probability-teacher consolidation around the $5\times$ endpoints transfers 68% of the ensemble advantage into a single student at half budget with endpoint-class depth intact, bounded at full budget where plain continuation overtakes (Section 6); remaining: budget replication of the residual and stronger alignment families. (f) All perplexity/CE claims are tied to the fixed window construction; extrapolation is measured once out of corpus (WikiText-103, unseen): the probability teacher beats the stronger endpoint by 11.2% with all-positive margins, capture orderings transfer out of distribution (student captures 40–76% on WikiText-103), and the 25 k continuation eruption replicates while the 12.5 k control's two-window eruption does not – separating corpus-specific poisoning from model-intrinsic deep instability (Section 6; single corpus, the cross-data map remains open).

N TRAINING CONFIGURATION, DATA DISCLOSURE, AND ARTIFACT INDEX

All runs in Table 3 share one recipe family (Section 2): FineWeb-Edu sample/10BT, the GPT-2 byte-pair-encoding (BPE) tokenizer (vocab $50{,}304$), validation split = last 4×10^6 tokens of FineWeb-Edu shard 004 (the fixed evaluation windows are drawn from exactly that tail), fused AdamW $(0.9, 0.95)$, weight decay 0.1 on matrix parameters only, warmup 200 then cosine to a 10% floor, context 1024. Per-run seeds drive the data sampler streams; tokens seen = steps $\times$ batch size $\times$ context $\times$ gradient accumulation.

Table 3: Per-run training configuration, grouped by population (from the archived run arguments). Every run has `n_layers=18` in its arguments. All runs except the dense control are weight-tied looped models (`arch=loop`, mode `dense`; the normalization ablation uses mode `dense_ln`); the dense control is a standard untied 18-layer transformer (`arch=tpp`), for which the pre/post and loop arguments are inert. Pre/post: non-looped blocks before/after the looped block. Loop arguments use their archived names: `hi` is `r_sample_hi` (the uN recipes use `hi`=N), `blocks` is `loop_blocks`, and `fixed` is `r_fixed`. Tokens = steps $\times$ batch $\times$ context $\times$ accumulation. A dash (–) marks an argument that does not apply.

Run ID(s)	d	Heads	Pre/post	Seed(s)	Steps	Tokens (B)	Loop args
Fixed-depth anchors							
`sdc_anchor_dense(/_s1)`	512	8	1/1	0/1	5000	0.246	`hi`=0
`sdc_anchor_dense_s2/s3`	512	8	1/1	2/3	5000	0.246	`hi`=0
Loop recipes and loop ablations							
`dense_u24_s1/s2`	512	8	1/1	1/2	5000	0.246	`hi`=24
`dome_dense_u32(/_s2)`	512	8	1/1	0/2	5000	0.246	`hi`=32
`dense_u40_s0/s2`	512	8	1/1	0/2	5000	0.246	`hi`=40
`dense_u48_s0/s2`	512	8	1/1	0/2	5000	0.246	`hi`=48
`e69_u56 / dense_u56_s2`	512	8	1/1	0/2	5000	0.246	`hi`=56
`dense_u64_s0/s2`	512	8	1/1	0/2	5000	0.246	`hi`=64
`e65_mb2 / mb2_s2`	512	8	1/1	0/2	5000	0.246	`hi`=0, `blocks`=2
`e65_k32f / k32f_s2`	512	8	1/1	0/2	5000	0.246	`fixed`=32
Normalization ablation							
`main_dnln_rms(/_s2)`	512	8	1/1	0/2	5000	0.246	`hi`=0
5×*-budget* `u56` *recipe*							
`mb16_x5_s3/s4`	512	8	1/1	3/4	25000	1.229	`hi`=56
`mb23_b5x_s7/s8`	512	8	1/1	7/8	25000	1.229	`hi`=56
Width ladder and budget sibling							
`big_dense_d1024(/_s2)`	1024	16	3/3	0/2	5000	0.246	`hi`=0
`big_dense_d1280(/_s2)`	1280	20	3/3	0/2	5000	0.246	`hi`=0
`big_dense_d1536(/_s2)`	1536	24	3/3	0/2	5000	0.246	`hi`=0
`big_dense_d2048(/_s2)`	2048	32	3/3	0/2	5000	0.246	`hi`=0
`big_dense_d2048_2x`	2048	32	3/3	0	10000	0.492	`hi`=0
`big_dense_d2560(/_s2)`	2560	40	3/3	0/2	5000	0.246	`hi`=0
`big_dense_d3072(/_s2)`	3072	48	3/3	0/2	5000	0.246	`hi`=0
Dense (non-looped) control							
`mb7_fixed18_s0/s1`	512	8	1/1	0/1	5000	0.246	–

Pre-registration and path conventions. Each study fixed its population, grids, two-sided bands, falsifiers, and a scoring rule in a specification before execution; results between bands are reported as undecided, and boundary-exact statistics are reported as such. All claims use the post-scoring records; exploratory analyses of scored results are labeled exploratory where they first appear. Printed paths beginning with `logs/`, `versions/`, `results_mb40b/`, `results_mb43/`, `results_mb44/`, or `results_mb45/` refer to the folder `projects/lmc_basin/` of the supplementary archive; the two `analysis/` paths refer to the top level of the archive.

Artifact index. The supplementary archive contains, per campaign, the evaluation code (evaluators, validity gates, scorers, rehearsal checks, and launch scripts); the pre-registered specifications,

fixed before measurement (one per study, with attempt logs and OUTCOME sections); the archived evaluation windows (per-window vectors from the MB1 evaluation; window tokens from the tail of FineWeb-Edu shard 004, `default_rng(0)`, 16 windows of length 1025); the scored result rows and job logs; and the archived training arguments of every checkpoint. A determinism check recomputes every derived quantity from the archived cells exactly.

O PER-CAMPAIGN RESULT RECORDS

This appendix collects the per-campaign records behind the summaries in the main text and in Appendices B–M. Each paragraph names its campaign, its verdict label (small capitals; see the Appendix guide and Appendix P), and the measured values; recurring terms are defined in Appendix C.

MB12: the trained-depth twin classes. Table 4 gives the per-pair results under the twin classes fixed in the MB12 specification before measurement: TRACKS-RESIDUAL (TR) if $|\rho_N - \rho_a| \leq 0.5$, TRACKS-NAIVE (TN) if $|\rho_N - \rho_n| \leq 0.5$, TRACKS-BOTH (TB) if both bands hold (corridor), DISTINCT-BETTER (DB) if below both by > 0.5, else DISTINCT (DT); a pair takes a class when both noise seeds agree, else AMBIGUOUS. ρ_N is evaluated at the midpoint $t{=}0.5$; the discriminating set $D{=}14$ excludes dnlnrms (non-discriminating; descriptive).

Table 4: Trained-depth norm-matched twin verdict per pair (MB12): naive/aligned barriers ρ_n/ρ_a (dex), twin residual $\rho_N(0.5)$ per noise seed, seed classes, and pair call.

pair	ρ_n	ρ_a	ρ_N N1	ρ_N N2	classes N1/N2	pair call
anchor01	2.6566	1.5952	1.9130	1.9306	TR/TR	TR
anchor02	2.0277	1.4087	1.9001	1.9185	TB/TN	AMBIGUOUS
anchor03	2.4014	1.3222	1.9186	1.9381	TN/TN	TN
anchor12	2.5044	1.3492	1.8567	1.7969	DT/TR	AMBIGUOUS
anchor13	2.0863	1.4119	1.8740	1.8138	TB/TB	TB
anchor23	1.9193	1.2802	1.8370	1.8269	TN/TN	TN
u24	1.9656	1.3625	1.8246	1.8235	TB/TB	TB
u32	3.3501	2.1391	1.8214	1.8731	TR/TR	TR
u40	5.2081	1.6678	1.7031	1.7061	TR/TR	TR
u48	4.8088	1.5610	1.9170	1.6835	TR/TR	TR
u56	4.9438	1.3182	1.7196	1.6452	TR/TR	TR
u64	4.6488	1.4913	1.7214	1.6653	TR/TR	TR
mb2	2.1328	1.3358	1.6372	1.7391	TB/TB	TB
k32f	2.0490	1.3172	1.9732	2.0013	TN/TN	TN
dnlnrms (desc.)	1.8930	1.5832	1.5933	1.6407	TB/TB	not counted

Population verdict ($k{=}14$): MIXED-VERBATIM (no majority class; the full class histogram is reported) – 6 TR / 3 TN / 3 TB / 2 AMBIGUOUS; neither ≥ 13 majority is reached.

Window-draw re-measure. A second independently drawn window set (same recipe, new seed) re-measures the principal statistics: every census-level claim re-closes; per-pair aligned barrier magnitudes shift 0.1–0.6 dex across draws (Section 3; Appendix H).

MB11 evidence: the depth-ramp falsification. Verdict RAMP-DEAD (both u56 and u40 BREACHED: no ramp schedule keeps all its cells inside the usable band; control note CONTROL-INCONSISTENT at anchor01), scored once from the second MB11 run (80 rows, validity gate $80/80$, write-once score log). Table 5 lists the mid-window perplexities. Mid-window window-mean perplexity at $t{=}0.25$ by runtime depth (fixed-$R16$ reference, then ramp depths, then fixed-$R192$ reference; $t{=}0.75$ mirrors):

Ramp endpoints at Rmax (SAFE, inside the band): u56 $61.9/61.8$ ($R64$), $79.7/73.7$ ($R192$); u40 $61.9/60.7$ ($R64$), $98.2/80.1$ ($R192$); window-max ≤ 146.5 against T_{usable} $245/241$ and the $\times 100$ guards. The midpoint breach is reproduced in-campaign ($5.29{\times}10^6$ u56, $9.51{\times}10^6$ u40 at $t{=}0.5$, $R{=}16$). Control: anchor01 $t{=}0/R64$ cell 254.4 leaves the pre-specified band 229 while $t{=}1/R64 = 83.3$ stays SAFE (CONTROL-INCONSISTENT, descriptive: on the anchor family, ramp validity is not uniform in t). Endpoint-identity checks pass throughout (blowup endpoints log-dist

Table 5: Depth-ramped averaging: mid-window means worsen monotonically with runtime depth (MB11), and ramp endpoints at Rmax stay inside the pre-specified usable band.

Mid-window PPL at runtime R	16	28	40	60	104	192
u56 (t=0.25)	353.0	367.6	389.4	437.2	563.7	814.1
u40 (t=0.25)	462.0	495.8	550.0	665.7	951.0	1679.2

3.74×10^{-3}/6.88×10^{-4} under the pre-specified blowup rule; $R16$ at $\sim10^{-6}$ rel; u56/u40 endpoints $\leq 1.9\times10^{-4}$ rel).

MB13 evidence: the dense-18 aligned residual. Verdict DENSE-RESIDUAL-LOOP-LIKE, scored once from the first MB13 run (10 rows, validity gate $10/10$, write-once score log; all endpoint checks at relative error $\leq 9.7\times10^{-9}$; scoring uses the window-mean basis throughout, as for the census). Naive arm reproduces the MB7 row means deterministically ($41.96/41.86$ endpoints, 2805.7 midpoint); aligned arm $41.96/41.86$ endpoints, interior $299.6/830.7/281.9$ (t=.25/.5/.75). $\rho_{n,13} = 1.8263$ (mean basis; MB7's median-basis $B_{18} = 1.7762$ is reported alongside), $\rho_{a,13} = 1.2977$ – inside the loop band $[1.2802, 2.1391]$ (0.0175 above floor). $\mathrm{shell}_{13} = 0.5286$ dex; collapse share 0.2894. Match diagnostics: $G1 = 0.448$, $G2 = 0.0381$ (weak head class, the same signature as the loop population), $G3 = 0.1983$. Permutation provenance: the tower-entry shared P_d plus per-block P_h/P_m on the 18 unshared blocks (the dense-model reading of the pre-specified three-class scheme; no class added or dropped).

MB15/MB19 evidence: the four-matcher sensitivity of the split. Set verdict AMBIGUOUS, scored once in the MB19 completion run (the two MB15 runs did not reach a verdict because the weight matcher did not read the loop RMSNorm gain of the normalization-ablation (dnlnrms) models; the pre-specified MB19 completion run then scored the retained $496/500$-equivalent outputs, validity gate $500/500$, write-once score log). Arm-level stability (pre-specified component tests S1, every pair keeps residual $\geq10\times$, and S2, collapse share and class gap; both required for SHELL-STABLE): *pass on all four arms* – $15/15$ pairs keep post-alignment residual $\geq10\times$ AND collapse share ≥0.50 with $|\mathrm{gap}_a| \leq 0.5$ dex (M1: $+0.459/0.739$; M2: $-0.043/1.025$; M3A: $+0.037/0.979$; M3B: $+0.035/0.980$). Function-informed matchers reproduce the M0 magnitudes (per-tag ρ_a M2 1.10–1.38, M3A 1.31–1.95, M3B 1.29–1.90; shell positive throughout); the activation-free M1 arm is the outlier (residual up to $1.77\times10^{11}\times$ on u56, shell negative on $11/15$ pairs, ρ_a extremes 8.81–11.25 on u32/u40/u56). Match diagnostics mean-over-tags: M1 G1/G2/G3 $0.101/0.022/0.084$; M2 $0.297/0.039/0.357$; M3A/M3B $\sim0.314/0.022/0.359$ (G2-weak signature reproduced on every arm). Set-level drivers of AMBIGUOUS (stored-precision flags defined in the MB19 specification): flag f6 (one M2 Sinkhorn plan row/column error above 10^{-6} at the 200-iteration convergence floor) and flag f7 (for the dense control's M3A and M3B arms, the M3 index arrays are not bit-identical to the archived control permutations); flag f5 did not fire. G2-head-attribution ladder (question Q3 of the specification): per-pair $|\mathrm{M3A}-\mathrm{M0}| \leq 0.188$ dex, $|\mathrm{M3B}-\mathrm{M0}| \leq 0.417$ dex, sign-mixed across all fifteen pairs – the weakest-matched class does not carry the residual. Dense-18 control re-test: arm-stable, all four arms DENSE-RESIDUAL-LOOP-LIKE (ρ_a1.29–1.86). dnlnrms, reported separately (not pooled): ρ_a1.486–1.848 across arms.

MB17 evidence: the usable deep axis. Verdict USABLE-OPPOSITION-NULL (no band breaches; both arms) with SMALL-STEP-FLAT (question Q4: the $R16\to R32$ change of the naïve barrier is not an increase of ≥0.2 dex on both hosts), scored once from the first MB17 run (270 rows, validity gate $270/270$, write-once score log; 18 endpoint-identity checks pass against the MB1/MB11 records). Ten-factor depth grid $R \in \{16, 20, 24, 28, 32, 48, 64, 96, 128, 192\}$ on the nine-point mixture grid of the depth-tame pair families (u56, u40). T_{usable} (re-derived as $4\times$ the R=16 endpoint maxima) $= 245/241$ holds at every grid endpoint out to R=192 (most-tame endpoint range 60.2–80.1). Naïve barrier vs usable depth: $\rho_{192} - \rho_{16} = -0.395$ (u56), -0.919 (u40); the eight additive small-step deltas $R16\to R32$ are all negative and monotone per host ($-0.133\ldots-0.316$ u56; $-0.139\ldots-0.403$ u40). Aligned-u56 residual: $+1.318$ ($R16$) rising slowly to $+1.493$ dex ($R192$; $\approx 31\times$ midpoint excess). Breach geometry: onset medians 2–3.5 (naïve) receding slowly to 6–8 (aligned); aligned damaged-position share 0.36 vs naïve 0.60–0.72; window-cluster bootstrap CIs (B=10000, seed 0) accompany every aggregate in the score log. A decay-form reading unifies the

two depth phenomena on the scored rows (an exploratory analysis, not pre-registered): the naïve barrier saturated-exponential fits $\rho_n(R) = \rho_\infty + (\rho_{16} - \rho_\infty)(R/16)^{-k}$ with host-specific (ρ_∞, k) – u56 $(4.52, 1.60)$ rmse 0.024, u40 $(4.04, 0.60)$ rmse 0.009 – while the aligned residual grows log R-linearly (fit $\rho_a(R) = \rho_{16} + s\ln(R/16)$ over the ten-factor grid above: $s = +0.078$ dex per e-fold, equivalently $+0.1806$ dex per decade by the $\ln 10$ conversion — one law, two unit conventions; the decade form retests the grid at R^2=0.967, max-residual 0.027): the barrier is a saturated shell plus a slowly growing residual, and convergence to a common basin holds family-internally, not to any universal floor (hosts differ). Two population extensions (pre-registered before measurement) extend the coverage of this law: on u40 the aligned residual again grows log R-linearly and positively ($+0.1025$ dex per decade, R^2=0.925; equivalently $+0.0445$ per e-fold) with the same late-grid turnover shape — law class and direction carry across both depth-tame hosts, while the slope is host-dependent ($|\Delta|$=0.078, outside the pre-specified 0.05 invariance band); on u24, a loop-recipe pair whose endpoints collapse at depth, the aligned channel does the opposite: while its endpoints remain usable the aligned residual grows *faster* than either tame host ($+1.36$ at $R16$ to $+2.66$ at $R96$, against $+1.48$ and $+1.76$ on u56 and u40), and then the endpoint catastrophe enters the residual's own denominator — at R=192 the two endpoints sit at 2.8×10^{13} and 3.7×10^{18} ppl while the aligned midpoint sits at 4.7×10^4 — and the barrier statistic collapses to -8.77 (value from campaign RE5; on the arithmetic-window basis the MB3 census value for the same alignment is -8.85 dex, a 0.08 dex difference between the two evaluation runs, while Table 1 reports -1.80 dex on the corpus basis). The second singularity is thereby measured, but it is an *endpoint* singularity that the aligned interpolation rescues by nine orders of magnitude, not an aligned-residual law: on the depth-tame family the log-growth law holds and the residual is the barrier; on the anchor family the endpoints die first at depth and ρ stops measuring a residual at all (the pre-specified classes return nonpositive, weak, or saturating labels; the full cell table is in the supplementary archive).

MB18 evidence: population-level reads of the 97% claim and of the twins. A re-analysis of existing cells without new computation (a new pre-registered specification; scored once, write-once score log; all 19 rehearsal checks passed before the specification was fixed). (i) Class-stratified nested bootstrap (B=20000, seed 0): collapse share 0.9727, interval $[0.8355, 1.0674]$, so the negligible-share region, which the earlier interval $[-0.97, 0.998]$ included, is excluded without re-measurement. (ii) Population twin share over the fourteen discriminating pairs: median 1.2884, 98.7% order-statistic interval $[1.1336, 1.4310]$ – norm-matched random-averaging twins overshoot the aligned residual at the median. Significance without distribution assumptions: $13/14$ per-pair shares exceed 1 (exact sign-test $p = 9.2 \times 10^{-4}$; the sole sub-1 share is u32's 0.863), and the two-stage pair→window bootstrap (B=20000, seed 0; windows resampled paired across the twin and aligned arms per draw) returns median 1.2854 with a 95% interval $[1.1475, 1.4021]$ – share-1 excludes zero at both stages (`logs/q2_twinshare_2stage.txt`). A second window draw gives the same population picture (MB36-zero, a zero-compute re-analysis of the MB35 twin cells): on the second window construction (seed 2026093501) the per-pair shares read median 1.4199 with **all fourteen shares** > 1 $(14/14)$ and order-statistic interval $[1.1628, 1.4863]$, overlapping the draw-1 interval $[1.1336, 1.4310]$ on $(1.1628, 1.4310)$; the same one-seed measurement on draw 1 reads median 1.3101, so the draw-to-draw population shift is $+0.11$. Basis disclosure: the draw-2 share uses the single available N1 twin seed with the draw-2-re-solved M2/OT-Sinkhorn aligned denominator, while the draw-1 value uses two seeds against the Hungarian census arm – the arm covariate reads class-median 1.3891 (Hungarian, draw 1) vs 1.2270 (OT, draw 1) vs 1.2086 (OT re-solved, draw 2), so the cross-draw read is population-level; the aligned-residual range on draw 2 is $[1.1306, 1.5207]$ against the draw-1 Hungarian band $[1.2802, 2.1391]$. Dependence-corrected layers hold too (record: `logs/q2_twinshare_dependence.txt`). The pair-level MB12 classification (Table 4) remains the unpooled reference. (iii) MB37 structured-direction probe (pre-registered specification; one GPU evaluation; write-once verdict `logs/mb37_verdict.txt`): at the identical per-tensor dose (audit rel $\leq 10^{-6}$), the isotropic (ISO) twin reproduces the N1 census within 3×10^{-3} (paired stream, original twin construction), and the displacement-subspace class separates cleanly: median $\Delta(\mathrm{SUB}-\mathrm{COMP}) = +1.1408$ dex (SUB: dose in the displacement subspace; COMP: in its complement), order-stat $(+0.566, +8.482)$, $14/14$ positive, two-stage paired bootstrap 95% $[+0.658, +2.158]$. Per-pair Δ spans $+0.32$ (k32f) to $+23.01$ (u64); sampled-U hosts amplify most (u56 $+12.06$, u40 $+1.43$, u24 $+0.57$), anchors sit mid-band ($+0.53$–$+2.02$); the complement is isotropic-like everywhere (Δ_{comp} median $+0.0064$), and SUB exceeds ISO on $14/14$ (Δ_{iso} median $+1.1333$). The paired random-frame control (MB37B, pre-registered addendum) shows that the effect is specific to the displacement's own subspace:

Haar-random 8-frames under the identical concentration and dose land isotropic-like on all fourteen pairs (ρ_{RAND} median 1.7287 against ρ_{ISO} 1.8469), reading $\Delta_{rs} = \rho_{\mathrm{SUB}} - \rho_{\mathrm{RAND}}$ median $+1.2665$ dex, order-stat $(+0.604, +8.774)$, $14/14$ positive, two-stage paired bootstrap 95% $[+0.711, +2.186]$. Reading: what explodes at fixed per-tensor dose is dose written along the displacement's *own* low-rank output subspace (8 of the $\geq$512 output-space directions per weight tensor; 1-D tensors carry the isotropic control) – not concentration geometry in general – and the trained displacement itself stays anomalously *benign* inside its own explosive subspace. Records for this paragraph: `logs/q2_twinshare_2stage.txt`, `logs/mb36_zero_verdict.txt`, `logs/mb37_verdict.txt`, and `logs/mb37b_verdict.txt`. (iv) Question Q5, u24: trained-depth unremarkable (twin 1.335 vs class 1.207); the deep-axis deep valley (-9.268 vs class $+2.435$) stands as a second, deep-only singularity with no class-internal second family detected.

MB16 evidence: transfer at stronger trained endpoints. Verdict V1-TRANSFER-CONFIRMED, scored once from two MB16 runs (the first ran units 1–3: naïve battery in both mixture directions, 12+12 cells, alignment arm 8 cells + permutations; the depth evaluator stopped before unit 4 produced output, and a follow-up run with a corrected host configuration completed units 4–6 (3×90 cells + geometry files); of 302 raw rows, the 294 specified by the protocol were scored, with a single write-once score log). A fresh u56-recipe pair at $5\times$ budget (25,000 steps, 1.229×10^9 tokens, seeds $3/4$ with disjoint sampler and validation streams; checkpoint-reuse probe 0) lands endpoints $38.3833/38.4401$ against the $1\times$ pair's $61.2079/60.1853$. The decomposition transfers and *amplifies*: naïve ρ 4.9438 $\rightarrow$ 7.9940 dex (both mixture directions identical to seven digits), aligned ρ_a 1.3182 $\rightarrow$ 9.5061 dex – the aligned midpoint (1.23×10^{11}) sits *above* the naïve (3.80×10^9): the gauge is negligible in this regime and the separation is total. The poles are depth-tame (R=16 $\rightarrow$ 192: 38.38 $\rightarrow$ 40.01, 38.44 $\rightarrow$ 39.37; drift 1.02–1.04$\times$; deep poles inside T_{usable}=154) while the midpoint breaches from $R\geq64$ (naïve deep rod $\sim3.3 \times 10^3$; aligned $\sim3.3 \times 10^5$). Validity gates: gate 1 passes (all endpoints < 52); gate 2 does not apply (no weak endpoint). The $2\times$ middle rung (MB16B, descriptive; scored once): endpoints $47.8004/47.0041$, naïve ρ 4.2677 (both mixture directions), aligned ρ_a 1.7772, pole drift 1.19–1.30$\times$ at R=192, breach medians 3.0–3.5 naïve / 5.0–5.5 aligned, endpoint audits clean. The completed dose ladder: endpoint quality monotone ($61.2/60.2 \rightarrow 47.8/47.0 \rightarrow 38.4/38.4$) and uninformative; gauge-resolved residual amplified at every rung (1.318 $\rightarrow$ 1.777 $\rightarrow$ 9.506 dex); naïve barrier non-monotone (4.944 $\rightarrow$ 4.268 $\rightarrow$ 7.994 dex, $2\times$ shell share shrinking). An independent-seed control pair at the same $5\times$ rung (MB23 B arm, fresh seeds $7/8$, endpoints $38.7/38.3$) reproduces the principal family instead: naïve ρ 4.03 dex, aligned ρ_a 2.15 dex, alignment benefit $R= + 1.88$ dex in-band, aligned midpoint $75\times$ better than the naïve – so the 9.51 amplification and the aligned-harmful datum above are *pair-specific*, with the independent seed replica in the principal family.

MB23-A evidence: the capacity law at n=6 – the plateau holds, the rank law does not. The width ladder is extended to six pairs with identical-recipe endpoints at d=1280 and d=2560 ($20/40$ heads, 3+3 loop blocks, seeds $0/2$; endpoint quality $39.87/39.78$ and $41.57/38.95$ at R=16), evaluated on the same certified windows and the pre-specified metric ($B = \mathrm{ppl}(t{=}.5, R{=}16)/\min(\mathrm{endpoints})$, $\log B$ in dex). The six-point series – 2.4756 (d=1024), 2.2948 (1280), 2.4321 (1536), 2.4587 (2048), 2.3172 (2560), 2.0402 (3072) – keeps the sign of the relief but not its rank strength: $\mathrm{Spearman}(\log_{10} d, \log B) = -0.6000$, which misses the pre-registered band (negative sign, $|\rho| \geq 0.7$; the pre-registered sensitivity clause for n). The shape on six pairs is the same as on four – a plateau (2.29–2.48) with a single-step relief at d=3072 – with a measured within-plateau pair variance of ±0.16 dex, the same variance class that the budget arm's independent-seed split carries (MB23-B). Deep profiles break monotonicity outright: d=1280 sits mid-ladder (logB192 = 3.3438, rod = 10.54), while d=2560 is the most stable deep mixture on the entire ladder (rod = 1.245 ($D_{\mathrm{mid}} = 1.327$ vs $\max(D_0, D_1) = 1.066$), logB192 = 2.4345 below d=1024's 2.8057); the deep-side "wider is worse" reading is thus pair-variable on this architecture family. Gauge alignment on the new rungs does not pay at trained depth on either: d=1280 reads ρ_a=2.3902 (alignment benefit $R= - 0.0954$, the first measured width pair where the aligned bridge costs more than the naïve path at trained depth, and ρ_a sits 0.25 dex above the census top 2.1391). The aligned residual completes the picture at six points: ρ_a reads 2.1635, 2.3902, 2.1768, 2.6760, 2.2639 by width through d=2560 – oscillating above the census top 2.1391 on $5/6$ rungs, with the alignment benefit R at or below the two-sided $+0.3$ floor on $4/6$ (two of them negative) – before collapsing to 1.5335 at d=3072 ($R= + 0.5067$, in-band, and the only width pair whose aligned deep profile also improves, logB192 3.2187 vs 3.9938). The preregistered monotone-non-

increasing posture is rejected as a law ($1024 \to 1280$ and $1536 \to 2048$ both rise); and the notable datum is that the relief pair and the alignment-paying pair are the *same* pair: no monotone width law survives on either axis, and the d=3072 seed pair is simply the most favorable of the six on both – pair-level variance on this family dominates width trends everywhere we can measure them. Matcher-arm scoping (Appendix E): this six-point series is the census Hungarian (G-class) reading on its own windows; the same checkpoints and identical windows re-measured under the stronger OT-Sinkhorn arm give the monotone contraction $2.2638 \to 1.6275$ (Spearman -1.0000); the two shapes are matcher-specific readings, both upper bounds at their specified arm, and the census-reference separation holds same-arm ($I_{c,\mathrm{OT}}$=$(1.1916, 1.2642)$ vs width $(1.9467, 2.3854)$, overlap 0).

MB23-B evidence: the $5\times$ budget claims do not replicate on matched independent seeds. A fresh u56-recipe endpoint pair at the identical $5\times$ budget (25,000 steps, seeds $7/8$ – disjoint from seeds $3/4$ of the ladder pair and $5/6$ of MB16B; same windows, grids, and scoring rules) lands endpoints at the same quality ($38.6708/38.3044$ on the certified windows vs the ladder pair's $38.3833/38.4401$) but a categorically different landscape. Naïve interpolation: the trained-depth barrier reads 4.03 dex (peak 5.42 dex at t=0.625; the curve is strongly right-skewed) instead of the ladder pair's 7.99 dex. With alignment: the mid-curve is symmetric ($558/530$ at t=$0.25/0.75$), the trained-depth residual is ρ_a=2.1495 dex – $+0.0104$ dex above the principal population top 2.1391, i.e., not significantly above – the alignment benefit R= + 1.8766 dex lands inside the two-sided band $[+0.3, +2.0]$, and the aligned midpoint (5,404.7) is $75\times$ *better* than the naïve one (406,792), where the ladder pair had the inverse (1.23×10^{11} above 3.80×10^{9}). The deep side agrees: aligned logB192 2.6361 beats naïve 3.1529. Two pre-registered branches fire: B1, the amplification class (9.5061 dex at $5\times$, rung-monotone $1.32 \to 1.78 \to 9.51$), is classified as *pair-specific to that ladder member*, not a budget-axis property; B2, the aligned-harmful datum (question Q4), is classified as a *single-pair phenomenon*, and its mechanism (the context-specific bipolar catastrophe core, MB23-C) is private to that pair. What survives on both pairs at $5\times$ is the core split itself: naïve separation in the 4–8 dex range and an aligned residual at the top edge of the census magnitude; what varied is everything beyond it. The reading is that the trained-depth landscape at fixed recipe and matched endpoint quality carries large pair-level variance, and dose-ladder inferences from single pairs are bounded by it.

MB20 evidence: the probability teacher and student capture, with the depth dual account. On the MB16 endpoint pair, the per-position CE cache (16 windows $\times$ 1024 positions $\times$ 2 models, same certified window construction; endpoint checks against the stored cells at relative error $\leq 1.3 \times 10^{-5}$) gives the teacher curve at zero marginal compute: $q_{1/2}$ reads 35.5952 R16 arithmetic / 34.3750 geometric – 7.3% below the stronger endpoint in both aggregations, Jensen gap pointwise $\geq -10^{-12}$, the curve symmetric ($36.10/35.60/36.14$ at $\alpha \in \{0.25, 0.5, 0.75\}$), per-window argminima in $[0.40, 0.60]$ (median 0.50), window margins over A positive on all 16 windows ($+0.043$ to $+0.152$ nats), and per-token endpoint complementarity 50.1% (47.3–52.8% by window); the per-token best-of-two oracle (27.98) lies 0.206 nats below $q_{1/2}$ and is unattainable within the fixed-α family by convexity (per-token optimum sits on the boundary). At depth the teacher itself stays tame: $q_{1/2}$ at R=192 is 35.1190 whole-NLL – $1.0216\times$ its R16 geometric read, inside the endpoint tame class (1.02–1.04) – with a symmetric deep curve and all-positive deep window margins ($+0.062$ to $+0.183$). Endpoint-initialized paired student runs (λ=α=0.5, T=1; identical seed, batch stream, and depth draws; single write-once evaluation rows, teacher cost metered separately): rungs 1/5/12.5k steps land R16 arithmetic $38.5534/\mathbf{37.3299}/\mathbf{36.4983}$ against controls $39.6864/37.7650/36.3514$ – capture of the ensemble advantage $-6\% \to 38\% \to 68\%$ (geometric $-7\% \to 37\% \to 69\%$) with no plateau at 12.5k; teacher forwards are $45/47/47\%$ of student compute time, distill wall 2.33–2.34$\times$ control. The paired control exposes the depth account: at 12.5k the pure-CE arm matches the student at R16 while R=192 develops a concentrated eruption – exactly $2/16$ windows (max 660.3) against $0/16$ for the KL arm (student max 64.7) – plain continued training's failure mode here is window-specific poisoning that probability-teacher anchoring suppresses entirely. Bounds: the teacher's own deep gain is not yet inherited by the student at 12.5k (39.5 vs 35.1 at R=192; safety transmitted ahead of gain); the warm-reset transient at 1k is a start-up effect, not the steady state; the 25k full-budget rung (paired, same protocol) completes the curve (Appendix L).

MB20 evidence on WikiText-103: cross-corpus uniformity. Sixteen windows (1025 tokens) drawn from the tail of an unseen corpus (WikiText-103 token array; same window recipe and seeds; the FineWeb-Edu evaluation windows are unchanged; existing checkpoints only, write-once rows). The probability teacher beats the stronger endpoint by 11.2% (64.66 vs $72.83/73.23$ R16 arithmetic) with a symmetric curve and all-positive window margins ($+0.098$ to $+0.168$ nats) – stronger than in-corpus. Student R16 captures transfer with an out-of-distribution discount: d12 40.4%, c12 53.7%, d25 52.3%, c25 76.2% (ensemble advantage 8.57 arithmetic on this corpus), preserving the same orderings, including the full-budget advantage of continuation. Depth dual account, cross-corpus: d12 stays tame (R=192/$R16 = 1.03$); the c12 two-window eruption on FineWeb-Edu does *not* replicate (max window 126) – corpus-specific poisoning; c25's eruption replicates and deepens (max 13,289) – model-intrinsic deep instability; d25 stays $\sim$50$\times$ below c25 at the window max on this corpus. A single additional corpus is one step of a cross-data map, not the map; the within-construction uniformity and this out-of-corpus replication together bound the claim.

MB22 evidence: the same decomposition on WikiText-103 (cross-corpus replication), plus three companion probes. A fresh u56-family endpoint pair trained with the identical recipe on the second corpus (WikiText-103; same vocab 50,304; 0.2458×10^9 tokens per endpoint, seeds 101/102; endpoint quality $45.8/48.4$ PPL at R=16) replicates the core decomposition family: naïve separation holds on $16/16$ windows (median ρ_n=2.2175, window range $[1.99, 2.54]$ dex – same magnitude family as the principal population); the aligned residual census is ρ_a=1.2065 median-basis ($16.1\times$ median residual ratio, window range 9.5–$34.4\times$; 0.07 dex below the census band floor 1.2802); the per-pair shell share 0.456 lies inside the principal per-pair range $[0.30, 0.68]$ across the fifteen pairs; and the norm-matched twins reproduce the trained-depth residual at 1.6958 share – *overshoot*, as at population level (pre-registered band $[0.91, 1.72]$: HIT). The deep family behavior transfers too: endpoints stay tame (1.17–$1.22\times$ at R=192), midpoint drifts ($2.65\times$ naïve, $2.18\times$ aligned) land inside the principal 1.2–$5.2\times$ band, and no ignition event fires. Pre-registered band misses are reported as such: the collapse-share band $[0.84, 1.10]$ was specified for a population, and the single-pair shell share of 0.456 misses it as written, while lying inside the census per-pair range given above.

MB22-C evidence: a functional comparison beyond CE between the trained midpoint and its norm-matched twin (u24 pilot). On the u24 pair (R=16, same certified windows), the trained aligned midpoint is *better* than its norm-matched twin on every functional coordinate we measured: $+1.16$ nats of CE advantage (7.054 vs 8.216), symmetric-KL distance from endpoint A of 2.771 vs 4.043 ($+46\%$; this misses the pre-registered $\geq 2\times$ band), and 0.099 top-1 agreement between the two midpoints. The consistent read: beyond the norm budget there is *real but modest* retained structure – an intermediate verdict between "twins are just noise" and "the trained midpoint halves the gauge distance", consistent with the MB18 population-level twin share (1.2884 median, overshoot class) as the ceiling of the twin account at pair level.

MB22-G evidence: gauge-completeness stress test by temperature sweep of the soft-assignment class. Within the OT-Sinkhorn arm class, two new temperature levels ($\varepsilon \in \{0.01, 0.2\}$ around the default ε=0.05) move the aligned residual by at most 0.15 dex on every one of the fourteen principal pairs ($\max 0.1487$ at ε=0.01; full-level span ≤ 0.0904): near-permutation relaxation does not absorb the residual. We take this as evidence that the residual is not an artifact of assignment hardness within the soft-permutation family; rotations live outside the exact-symmetry family of this architecture (LayerNorm is not rotation-invariant), and we do not claim a geometric proof in that direction (the pre-registered bands of both questions are met; the hard-projection permutations vary by at most 0.13 dex across levels, sign-mixed). The continuous family is also measured and excluded: global orthogonal conjugation of the residual stream fails the pre-specified validity gates on all $15/15$ census pairs (the approximation and solver-consistency gates both), with median absorption -0.3701 dex — continuous conjugation is not a usable gauge at this LayerNorm/SiLU scale, so neither the tested permutation families nor continuous orthogonal conjugation remove the residual (record in the supplementary archive).

MB22-D evidence: a principled ceiling on the usable-depth question for anchors. On the anchor01 pair at dense $R \in \{1, \ldots, 192\}$: both parents erupt early (R_{ig}=62/72 at the $+1$ nat rule), the naïve midpoint's radius explodes (ρ_{192}=12.85) through a non-functional baseline (CE 10.76

nats at R=16; classified tame in CE because the $+1$ nat rule never fires), and the *aligned* midpoint erupts at R_{ig}=32 with radius crossing ρ=1.826 – inside the MB21 band $(1.65, 2.3)$ – versus its closed-form prediction R_{halt}=25.63 (Δ+6.37, band-adjacent hit). The parents' closed-form misses (Δ+43/+50) have an identified mechanism: the mean αF over the post-ignition arc inflates the constant-slope model; the refinement "read the slope from the early arc" is added to the MB21 model. The usable-regime depth question for anchored-family deep mixtures is thereby answered in the principled-impossibility direction: the aligned midpoint crosses the ignition band at R=32, long before any usable-quality region could be reached; parents do not reach it either.

MB23-C evidence: why alignment can be harmful at $5\times$ budget – a per-position mechanism. The one regime in which the aligned midpoint is *worse* than the naïve one (the $5\times$ datum, endpoint pair trained on $5\times$ tokens: aligned 1.23×10^{11} vs naïve 3.8×10^{9} PPL, both in the exploded deep-unroll regime, against a benign principal pair at $1\times$ where alignment helps throughout) was diagnosed on three pre-registered open candidates (diagnostic, no bands): endpoint-distance collapse (H1), degraded match quality (H2), and per-position harm concentration (H3). Define the aligned-minus-naïve CE lift per (window, position) cell, 16×1024 cells at R=16 on certified windows. H1 is not supported: the relative endpoint gap is if anything smaller on the $5\times$ pair (1.3101 vs 1.3788). H2 is partially supported but quantitatively insufficient: the matched neuron alignment is weaker – $G1$ mean correlation 0.2528 vs 0.3377 (-25%), $G3$ 0.3316 vs 0.3708 (-11%), $G2$ unchanged – yet stays positive and moderate, far from the low-or-negative regime that would mechanically break the aligned average. H3 is supported in a refined form: the *shape* of harm concentration is the same on both pairs (top-1% mass share 4.06% vs 4.54%; first-decile share 9.65% vs 9.08%), but the sign and magnitude flip. On the benign pair the entire lift distribution shifts negative – median -3.29 nats, p_{95}=+1.76, p_{99}=+3.38, and the per-position mean is negative even at its worst position (-0.54): no position is harmed on average. On the $5\times$ pair the median cell flips to harmed ($+0.36$; 51.6% of cells lifted) and the positive tail explodes an order of magnitude at fixed high quantiles (p_{95}=+21.82, p_{99}=+34.99, $p_{99.5}$=+38.03; single-cell maxima are large on both pairs, $+49.04$ vs benign $+54.32$) while the negative tail is comparable ($p_{0.5}$=-32.31 vs benign -36.18) – a bipolar catastrophe, not a uniform shift. Catastrophic cells (lift $> +10$ nats) occupy 15.6% of all cells on the $5\times$ pair versus 0.26% benign ($60\times$ frequency), and they carry no fixed positional anchor: only one position is hit in $\geq 8/16$ windows (231 in $\geq 4/16$), and the pairwise window correlation of full lift profiles is $+0.0000$ mean ($+0.0019$ median). The consistent mechanism read: heavy training on the $5\times$ pair leaves the two endpoints different functions that gauge alignment cannot fully reconcile (ρ_a=9.51 on this pair, pair-specific, MB23-B); their aligned average is context-specifically bipolar – in any window, a near-random $\sim 15\%$ of positions are torn by $+10$-nat-plus CE failures, and under deep unrolls these cells ignite and dominate the PPL. At tame depth (R=16) the mean lift is actually -0.34 nats, so the $5\times$ datum is an exploded-regime statement about ignition of this unstable cell population, not tame-depth degradation; the benign pair's matched structure is uniformly helped in the bulk of the tail (benign $p_{99.5}$=+4.09). We do not claim this mechanism beyond the paired diagnostic: one harmful pair against one benign reference pair, identical architecture and protocol. The paired-diagnostic caveat is borne out: an independent-seed control at the same $5\times$ budget (MB23 B arm, seeds $7/8$) reproduces the principal family (ρ_n 4.03 dex, ρ_a 2.15 dex, R= + 1.88 dex in-band, aligned midpoint $75\times$ better than the naïve) – the catastrophe core is thereby *specific to the ladder pair*, not a $5\times$-budget class property; the seed-pair variance of the trained-depth landscape at fixed recipe and matched endpoint quality is itself the main finding.

MB24 evidence: the graded support law (depth-jitter anchor clone). Verdict J-INTERMEDIATE with midpoint subclass ANCHOR-STYLE-MIDPOINT-RESCUE, scored once from the first jitter run (90 rows, validity gate $90/90$, write-once score log `logs/jitter_verdict.txt`; outcome branches pre-registered before any measurement: J-SUPPORT-TAME iff both seeds $D_{192}\leq$2.0 and $R192\leq 2\times R16$; J-MILD-INSUFFICIENT iff $D\geq 10^4$ and $R192\geq 10^4\times R16$; else J-INTERMEDIATE). The two hosts (loop count sampled from $U\{12..20\}$, seeds 0 and 1; archived as `jitter_12_20_s0/s1`) copy every training argument of the clean fixed-K16 anchor configuration except three – fixed loop count `fixed` 16→0 (sampling on), sampled upper bound `hi` 0→20, and sampled lower bound 12 (default 1) – and the seed; budget, batch stream, and protocol are unchanged (final losses $\sim$4.0, the 35 M family range; sanity check passed). Measured on the MB17 depth grid over the same 16 certified windows (naïve arm): endpoints 53.31→1222.89 (s0) and 52.95→910.92 (s1) at $R16\to R192$ – D_{192}=22.94/17.20 on the graded ladder D_{32}=1.026/1.048, D_{64}=1.325/1.744, D_{96}=2.346/4.627 – against the depth-tame band $D_{192} \in [1.22, 1.63]$ (u56/u40

endpoints, same windows) and the anchor-class explosions (median endpoint drift 4.97×10^6–3.43×10^{14}). On the $\log_{10} D_{192}$ scale (anchors $\sim$7.0; tame 0.09–0.21; jitter 1.24–1.36) mild jitter carries the hosts 82–85% of the distance toward the tame class without reaching it: the support threshold for full tameness lies strictly between $U\{12..20\}$ and the population's sampled-U supports (24–64), and the fixed-K explosion is a genuine recipe-class fact, not an artifact of mixture construction. The naïve $j0{\times}j1$ barrier on the same windows reads 2.1855 dex at $R16$ and 1.6411 dex at $R192$ (-0.544 dex at depth: the deep unroll rescues the midpoint relative to its endpoints, the anchor-pattern hallmark). A secondary measurement on the trainer's own validation stream (different basis, not used for the verdict) agrees in direction: the fixed-$K16$ anchor degrades $R16{\to}R48$ by $2.67\times$ against $1.12/1.28\times$ on the jitter hosts. The pre-registered MB3 trigger (barrier $\geq$0.5 dex) fires, so the aligned arm is scored as a third measurement (endpoint checks pass: 53.3079/52.9527 at $R16$, rel $\leq 1.9{\times}10^{-6}$ against the naive-arm cells): permutation alignment (G1 entry $|\mathrm{corr}|$=0.3349, the weak-head class signature; G2 0.0235–0.0665, G3 0.2722–0.4121) pays $R{=}{+}0.6737$ dex at trained depth, in-band, with the aligned residual ρ_a=1.5118 dex *inside* the principal loop band $[1.2802, 2.1391]$ (naïve 2.1855), weakening to 0.8045 dex at $R192$ (naïve 1.6411). The two axes separate: endpoint tameness is graded by the depth support, while the shell-resolved residual is already loop-class at mild jitter.

P CLAIM-TO-EVIDENCE MAP

The table lists each headline number in the main text, its campaign (study code; the corresponding specification in the supplementary archive states the pre-registered predictions and the OUTCOME), the measured population, the pre-specified band or threshold when the claim is inferential, and the verdict label with its plain-language meaning. Campaign codes are study identifiers, not version numbers; every scored row is listed in the artifact index (Appendix N).

Claim	Campaign	Population / inputs	Pre-specified band / threshold	Verdict (meaning)
Recipe-conditioned naive gap 1.7556 dex (sampled-U vs anchor)	MB2 recipe falsifier	14 principal pairs (8 loop incl. two loop ablations + 6 anchor; six sampled-depth pairs only: 2.4849 dex, so including the ablations is conservative)	F2 re-derived $\leq 5{\times}10^{-4}$ before output	recipe split exists on the principal population
Post-alignment gap 0.048 dex; collapse share 0.9727; truncated floor $\geq$ 0.83 (97.5% CI)	MB3 three-class alignment; MB18 nested bootstrap ($B = 20000$)	same 14; the 16 certified windows	CI $[0.8355, 1.0000]$ after domain truncation; 30.1% overshoot replicates	CI-STANDS (gauge explains the split; floor claim, not point)
Aligned residual census $\rho_a \in [1.2802, 2.1391]$; 19–138$\times$ excess	MB3 per-pair aligned rows	15 pairs (dnlnrms descriptive sibling)	band re-derived $\leq 5{\times}10^{-4}$	shell-resolved remainder present on every pair
Matcher-class sensitivity (OT-Sinkhorn and G2 ablations reproduce; weight-only M1 catastrophic)	MB15/MB19 four-matcher battery	15 pairs $\times$ 4 matcher classes	S1/S2 pre-specified component tests per arm	AMBIGUOUS as set token (arm-level stability is the content)
Dense-18 (79M, matched control) lands residual inside the loop band; $\rho_{a,13} = 1.2977$	MB13 dense-18 alignment	10 rows, scored once	loop band confirmed $\leq 5{\times}10^{-4}$	DENSE-RESIDUAL-LOOP-LIKE (not loop-specific)

continued on next page

Claim-to-evidence map, continued

Claim	Campaign	Population / inputs	Pre-specified band / threshold	Verdict (meaning)
Twin share median 1.2884 over discriminating pairs; draw-window stability and dependence layers; structured-direction probe	MB12 twins; MB18 order-stat CI on n=14; MB36-zero draw-2 re-analysis; dependence layers (leave-one-checkpoint-out, LOCO; checkpoint clusters); MB37 subspace probe	14 discriminating pairs ($\lvert\rho_a - \rho_n\rvert > 0.5$)	98.7% interval $[1.1336, 1.4310]$; draw-2 median 1.4199, order-stat $[1.1628, 1.4863]$, $14/14 > 1$; LOCO $10/11$; cluster 95% $[1.1592, 1.3876]$; subspace elevation $+1.14$ dex $14/14$ CI95 $[+0.66, +2.16]$; random-frame control $+1.27$ dex $14/14$ CI95 $[+0.71, +2.19]$	TWIN-POPULATION-EXPLAINS + D2-STANDS + DIR-SUBSPACE-ELEVATED + SUBSPACE-SPECIFIC (population claim stable across window construction; pair-level stays mixed; dose along the displacement's own low-rank output subspace is what explodes, not concentration in general)
Anchor-family deep valleys (parents explode, midpoints drift 1.2–5.2×, rod $\sim$0)	MB1 depth legs; MB12 deep twins	6 anchor pairs, depth $R192$	pre-specified twin band; AMBIGUOUS twin atlas	valley from endpoint collapse vs norm-budget account, both reported
Sampled-U persistent residual $+1.318 \rightarrow +1.493$ dex with depth; naive barrier weakens $(-0.39/-0.92)$	MB17 ten-factor usable-depth grid	2 host pairs (u56, u40), 270 rows	USABLE-OPPOSITION-NULL; T_{usable}=245/241	usable-regime opposition null
Width ladder boundary-exact $\rho = -0.8000$ (n=4)	MB2/MB4 capacity atlas	d=1024–3072 pairs	boundary-exact threshold $\rho \leq -0.8$	capacity relief direction measured at n=4; n=6 extension $\rho = -0.6000$ MISS (plateau+step shape on six pairs)
Budget dose ladder: 2× shared-seed same-trajectory datum; 5× aligned-harmful 1.23×10^{11}	MB16 stronger endpoints (5×, 25,000 steps)	fresh u56-recipe pair at 5×	gates G1 pass; G2 inapplicable	V1-TRANSFER-CONFIRMED as *pair-specific* ladder member; amplification clause not replicated on matched independent seeds (MB23 B arm)
Diagnostic, pair-private (seed control replicates the principal family): bipolar catastrophe mechanism of the 5× ladder pair (catastrophic cells 15.6% vs 0.26% of positions, window correlation 0.00)	MB23-C diagnostic on MB16 5× pair vs u56, per-position CE lift	16 windows × 1024 positions at R=16	lift = $\text{CE}_{\text{aligned}} - \text{CE}_{\text{naive}}$; H1–H3 pre-registered open	H1 rejected; H2 partial; H3 supported as sign/magnitude flip with pair-private core

continued on next page

Claim-to-evidence map, continued

Claim	Campaign	Population / inputs	Pre-specified band / threshold	Verdict (meaning)
Six-point width-aligned residual: plateau ρ_a 2.16–2.68 (5/6 above census top), alignment benefit $R<+0.3$ on $4/6$, monotone law rejected; relief pair d=3072 also pays alignment (ρ_a=1.5335, $R=+0.5067$)	MB23 alignment of the four original width pairs plus the two new widths, single scoring (write-once verdict)	six width pairs, certified windows	R two-sided band $[+0.3, +2.0]$; $4/6$ below floor	pair-variance over width law on both axes
Additive repairs measured to fail (naive and aligned soups; channel-moment repair; depth ramp RAMP-DEAD)	MB7/MB8/ MB11	twin gauges and u56/u40 hosts	pre-specified rows per study	weight-space constructive closure measured negative
Probability teacher $q_{1/2}$ beats stronger endpoint -7.3% (in-corpus), -11.2% (WikiText-103)	MB20 teacher evaluation on FineWeb-Edu and WikiText-103	16-window CE cache, endpoint-audited	Jensen pointwise $\geq -10^{-12}$; oracle 0.206 nat below	probability-space consolidation is viable
Student capture curve $-6/38/68\%$ at $1/5/12.5$k with depth-tame KL arm; full-budget continuation wins; anchor tax measured (65%–66% recovery by anneal, arith/geo); second-pair boundary	MB20 paired students ($\lambda=\alpha=0.5$, T=1); MB38 second pair (u56, 2.5k rung); MB39 vertex release (fresh 25k run sharing d25/c25's learning-rate schedule, data stream and per-step R draws, λ 0.5→0 at 12.5k)	4 rungs + control, same seed/stream; second-pair arms on the same stream; per-rung write-once rows	pre-registered falsifier band; second-pair captures $+1.38/+1.35$ vs $+1.82/+1.77$, ratios 1.19–1.30 tame; released 25k run R16 $35.3894/34.0971$, R192 $5181.3/223.8$, teacher cost 2965.4 s vs 5987.6 s	short-to-mid budget distillation domain, anchor-tax majority account; PAIR2-INTERMEDIATE (advantage endpoint-gap-conditioned, control leads from same rung); RELEASE-FAILS (vertex release halves the teacher bill, buys neither depth tameness nor endpoint quality; full-budget boundary stands)
Radius-ignition band: self-alignment α host-constant; slope law matched on u56 (0.1419 vs. αF=0.144, within 1.46%); $\rho^* \in (1.65, 2.3)\times r16$; R^* closed form	MB21 radial census + dense CE	3 host classes; 192-length trajectories	calibrated on hosts with recorded ignition	quantitative drift model; depth-switch hypothesis falsified
Depth-support law graded, split real: jitter $D_{192}=17$–23 vs tame $[1.22, 1.63]$ vs anchors $\sim 10^6$–10^{14}; barrier 2.19→1.64 dex at depth	MB24 depth-jitter anchor clone (first run, 90 rows)	2 fresh hosts (seeds $0/1$), same 16 windows, MB17 grid	outcome branches (J-SUPPORT-TAME/J-MILD-INSUFFICIENT) pre-registered	J-INTERMEDIATE + ANCHOR-STYLE-MIDPOINT-RESCUE (recipe-class fact, not construction)

Token glossary (plain language). CI-STANDS: the confidence interval excludes the negligible region – the claim stands at the pre-registered confidence. AMBIGUOUS (MB15/MB19): the pre-specified set-level label, reported as defined; it does not alter the arm-level results. TWIN-POPULATION-EXPLAINS: at population level, the norm-matched random-averaging twins reproduce the quantity at the median – the phenomenon is mostly a norm-budget effect. DENSE-RESIDUAL-LOOP-LIKE: the matched fixed-depth control sits inside the loop band – the effect is a generic small-from-scratch fact, not loop-specific. USABLE-OPPOSITION-NULL: within the usable band, the naive-vs-aligned opposition changed with depth in the measured direction and did not breach the pre-specified threshold. RAMP-DEAD: the depth-ramped averaging family was measured not to work (both hosts BREACHED: no schedule keeps all its cells SAFE, i.e., inside the usable band). CONTROL-INCONSISTENT: the anchor control cell leaves its usable band, so the control does not behave as specified. V1-TRANSFER-CONFIRMED: the pre-registered transfer checks (question Q1) at stronger endpoints all passed, with the pre-registered amplification clause reported alongside. REPAIR-DOES-NOT-CONNECT: the repaired midpoint's barrier stays above the pre-specified 1 dex cutoff. MIXED-VERBATIM: no pre-specified majority twin class is reached, so the full class histogram is reported. SHELL-STABLE: both component tests hold on all four matcher arms. SMALL-STEP-FLAT: the naïve barrier does not rise by $\geq$0.2 dex from R=16 to 32 on both hosts. D2-STANDS: the population twin-share claim holds on the second window draw. J-SUPPORT-TAME / J-MILD-INSUFFICIENT / J-INTERMEDIATE: the pre-specified three-way outcome of the depth-jitter anchor clone – mild depth jitter fully tames the 12$\times$ unroll / is far too weak to matter / lands measurably between the anchor and tame classes (what fired). ANCHOR-STYLE-MIDPOINT-RESCUE: at depth the naïve midpoint improves *relative to its own endpoints*, the anchor family's signature read. DIR-SUBSPACE-ELEVATED: at identical per-tensor dose, writing the direction into the real displacement's top-singular output subspace damages strictly more than its complement (median separation $\geq$ 0.10 dex with the paired interval excluding 0) – direction at fixed norm measurably carries damage content. SUBSPACE-SPECIFIC: the same elevation holds against a Haar-random frame at identical concentration and dose (14/14, interval excluding 0) – the operative object is the displacement's own subspace, not concentration geometry in general. PAIR2-INTERMEDIATE: the pre-specified second-pair consolidation outcome class – neither replication (the pre-registered vertex band is overshot on both arms) nor failure (no pre-specified failure branch fires); the second unit bounds the consolidation advantage to the large-gap strong-pair regime. RELEASE-FAILS: the pre-specified MB39 vertex-release outcome – the released run lands strictly between the distill and control arms at trained depth (neither Pareto nor recovery predicate fires), so the post-vertex teacher-cost saving is measured not to preserve the budget-position of either arm. The labels of the public-pair and random-permutation checks are explained where they appear (Appendices R and S).

Q UNCERTAINTY QUANTIFICATION

Pair-cluster percentile bootstrap (B=2000; window-level batch: two-stage, pairs within class then the drawn pair's 16 windows). Estimator choice: rod and D use the ratio-of-window-means form, which reproduces all stored anchor values exactly. Small-n caveats apply to every interval (the population has fifteen pairs and fewer independent units; Appendix U). Table 6 lists the estimates.

Granularity map. Covered at window level: the recipe-split collapse, residual statistics, and the capacity rod ladder (MB1/MB3/MB2 archive per-window vectors). Not covered by any bootstrap: MB4 tracer statistics (worst-1% share, first-breach ladder) and the MB7 B_{18} barrier – those results were stored as summary statistics without per-window values, so window-level uncertainty cannot be computed without re-measurement, which this manuscript does not perform.

R PUBLIC PYTHIA PAIRS AT 70M, 160M, AND 410M

MB40 evidence: the publicly trained modern independent pair at 160 M. The core measurements are re-run on the first of three modern, publicly trained, independent same-configuration pairs we could identify: Pythia-160M duped vs deduped (`EleutherAI/pythia-160m(-deduped)`; separate training runs, same architecture and hyperparameters; the deduplication recipe difference and unpublished init seeds are acknowledged confounds). The evaluation windows mirror the main window construction exactly (the last 4 M

Table 6: Headline point estimates and 95% bootstrap confidence intervals (percentile bounds). Log-barrier statistics are in dex; rod values are ratios. The observed residual range is $\rho_a \in [1.2802, 2.1391]$ dex (no interval is given for extrema).

Statistic (bootstrap scheme)	Point	2.5%	97.5%
F2 recipe gap, naïve (dex, two-stage)	1.7556	0.0723	2.7671
F2 recipe gap, aligned (dex, two-stage)	0.0480	0.0032	0.3064
Collapse share $1 - \mathrm{F2}_a/\mathrm{F2}_n$ (truncated[a])	0.9727	0.8355	1.0000
Median residual ρ_a (dex, pair-cluster)	1.3856	1.3265	1.5433
Median residual ρ_a (dex, two-stage)	1.3856	1.3142	1.5195
TR pair fraction (pair-cluster)	0.4286	0.2143	0.7143
Both seeds in TR band (pair-cluster)	0.6429	0.4286	0.8571
Mean twin $\rho_N(0.5)$ (dex, pair-cluster)	1.8207	1.7713	1.8684
rod, d=1024	1.86	1.41	2.44
rod, d=1536	5.15	3.66	6.90
rod, d=2048	3.82	2.73	5.10
rod, d=3072	89.7	58.9	123.1
rod, d=2048 vs 2× budget	923.6	421	1634

[a]Truncation audit of the same B=20000 bootstrap replicates (iteration-level recompute, zero new measurement): 30.1% of shares formally exceed 1 (the aligned family gap flips sign on those resamples). The domain-truncated interval at 1 becomes $[0.8355, 1.0000]$ (median 0.9707); the pre-truncation interval is $[0.8355, 1.0674]$.

tokens of FineWeb-Edu shard 004, `default_rng(0)`, 16 windows of 1025 tokens; re-tokenized with the GPT-NeoX BPE tokenizer; a 2024 corpus neither endpoint trained on), and all evaluation uses a fixed, dependency-free GPT-NeoX forward pass (verified against the Hugging Face implementation to $2-4 \times 10^{-5}$ relative PPL error; permutation invariance exact to machine precision under identity tests; the head-major interleaved qkv layout is documented). H denotes $\log_{10}$ perplexity in dex. Endpoint quality reads 1.365492 dex (duped) vs 1.533746 dex (deduped) in window-mean $\log_{10}$ ppl.

Barrier and named matcher families. The naïve left-to-right barrier stands at 3.1721 dex, and the best achieved barriers are reported per named family (best family F4_OTH_G123, shell share 0.0500 of the naïve barrier gauge-explainable; single-pair observation, no universality claim licensed by one pair). Families F1 and F2 use Hungarian matching on the cost $1 - |\mathrm{corr}|$ for the residual-stream class only (G1) or for all three classes (G123); F3 and F4 use OT-Sinkhorn followed by a Hungarian projection, again for G1 or G123:

family	matched $\log B$	match $\lvert\mathrm{corr}\rvert$
`F1_HUN_G1`	3.1721	0.4342
`F2_HUN_G123`	3.2479	0.4342
`F3_OTH_G1`	3.1721	0.4342
`F4_OTH_G123`	3.0136	0.4342

Norm/direction/allocation interventions. At identical per-tensor Frobenius dose (dose audit rel $\leq 10^{-6}$ per tensor), damage written into the displacement's top-8 left-singular output subspace vs its orthogonal complement separates by -6.7851 dex (window-paired bootstrap $[-7.2773, -6.3961]$), verdict DIR-SUBSPACE-SUPPRESSED (damage in the displacement's own subspace is lower than in its complement); the isotropic twin's layerwise-versus-tensorwise allocation comparison is reported descriptively.

Representation-level repair (stitching). A ridge-linear map between mid-network activations composes a staggered hybrid at every inter-block boundary; the worst seam's recovery fraction of the naïve midpoint damage reads 0.8985, verdict P-STITCH-REPAIRS (pre-specified recovery-fraction classes: ≥ 0.5 repairs, ≤ 0.2 fails, mixed otherwise).

Probability teacher. The mixture of endpoint distributions (computed with `logaddexp`; the Jensen-bound check passes) has relative ensemble advantage 0.14849 with endpoint gap 0.12322 at best $\alpha = 0.500$.

Composite verdict: MOD-BOUNDED (no named matcher removes most of the barrier). The pre-registered specification with all verdict classes and confounds is `versions/MB40_modern_pair_replication_spec.md`.

Consolidation arms under the $q_{0.500}$ teacher (MB40, experiment E6). A-init students trained 500 steps (batch 8×1024, identical LR clock and paired stream draws across arms; the budget amendment is documented in the specification): plain CE reaches a window-mean H of 2.2252 dex and CE + 0.5 KL at $\alpha^\star = 0.500$ reaches 2.1231 ($\Delta = -0.1020$ dex), against the A endpoint's 1.365492; moves from the shared init are $+0.8597$ (CE) and $+0.7576$ (KL) in the entropy basis H (lower is better — both students degrade from init), and the KL arm's degradation from init measures 0.8813 of the CE arm's (relative degradation, smaller is better), teacher-forward cost logged arm-wise (CE 0, KL 1 per step). Under the pre-specified gap the mixed teacher repays its forward cost on this short paired budget (verdict P-DISTILL-LEADS). The comparison is arm-internal at matched steps and draws; it does not include the teacher forward passes in a compute comparison, and the students sit far above the endpoints at this budget by construction.

MB40b: the 70 M scale companion (Pythia-70M duped vs deduped). The same fixed protocol is re-run at $\approx 0.43\times$ the parameter count (70 M vs 162 M) on the second modern, publicly trained independent pair we could identify (6L/d512/8H, GPT-NeoX BPE, Pythia v1 recipe; same acknowledged confounds as at 160 M — deduplication recipe difference, unpublished init seeds — plus a visibly larger endpoint-quality gap: duped 36.6312 vs deduped 108.5768 PPL (window-mean basis; corpus-level $H = 1.5417$ vs 1.8458 dex), gap 0.3041 dex against 0.168 dex at 160 M). Windows and forward path are construction-identical to the 160 M measurements above (pre-registered specification `versions/MB40b_70m_scale_point_spec.md`; checkpoint MD5 hashes recorded; corpus: tail of FineWeb-Edu shard 004). The naïve left-to-right barrier stands at 4.3999 dex; the four named matcher families read:

family	matched $\log B$
`F1_HUN_G1`	4.3999
`F2_HUN_G123`	4.2963
`F3_OTH_G1`	4.3999
`F4_OTH_G123`	4.6014

(match $|\mathrm{corr}|$ 0.4278). Two results stand out: the best family is `F2_HUN_G123` (shell share 0.0236), and one family (`F4_OTH_G123`) lands *above* the naïve barrier. The pre-specified predicates resolve as follows: the best matched barrier (4.2963) stays above half the naïve barrier, and the stability predicate passes, so, as at 160 M, the named matchers remove little of the barrier at 70 M. Twin comparisons are not run at this scale point (outside the specification's scope); the verdict file is `logs/mb40b_verdict.txt` (write-once), and the numbers above come from `results_mb40b/e1_e2.json`.

MB43: the 410 M scale extension (Pythia-410M duped vs deduped). The same protocol is re-run once more at $\approx 2.5\times$ the 160M parameter count (405 M vs 162 M estimated) on the third and largest modern, publicly trained independent pair we could identify (24L/d1024/16H, GPT-NeoX BPE, Pythia v1 recipe; same acknowledged confounds, with the endpoint-quality gap smaller, 0.0045 dex — duped 17.1157 vs deduped 17.3430 PPL (window-mean basis; corpus-level $H = 1.2120$ vs 1.2165 dex) — against 0.168 dex at 160 M and 0.3041 dex at 70 M). Windows and forward path are construction-identical (pre-registered specification `versions/MB43_modern_pair_scale_ext_spec.md`; checkpoint MD5 hashes recorded; corpus: tail of FineWeb-Edu shard 004; per-family endpoint permutation invariance additionally gated at $\leq 10^{-4}$ relative). The naïve left-to-right barrier narrows to 2.0542 dex; the four named matcher families read:

family	matched $\log B$
`F1_HUN_G1`	2.0542
`F2_HUN_G123`	2.4481
`F3_OTH_G1`	2.0542
`F4_OTH_G123`	2.4739

The best family is `F1_HUN_G1` (shell share 0.0000), i.e. the named families remove *nothing* of the barrier at this scale. As at 70 M and 160 M, the named matchers remove little or none of the barrier, so this pattern holds at all three tested scales (naïve barriers 4.3999/3.1721/2.0542 dex, shell shares 0.0236/0.0500/0.0000 at 70/160/410 M). The narrowing of the raw barrier with scale ($4.3999 \to 3.1721 \to 2.0542$ dex) is an observation on three recipe-confounded pairs, not a scaling law. The verdict file is `logs/mb43_verdict.txt` (write-once), and the numbers above come from `results_mb43/e1_e2.json`.

MB44: ridge-stitch repair at the 410 M scale point. The positive repair result on the public pairs — the layer-wise residual-stream ridge stitch of MB40 (experiment E5), recovering 0.8985 of the naïve midpoint damage at 160 M — is re-run byte-identically at 410 M (pre-registered specification `versions/MB44_stitch_410m_ext_spec.md`; the only documented deviation is a numerically identical activation-caching rewrite; single hybrid direction, worst-layer reading; a cross-check gate requires the probe to reproduce the MB43 endpoint and naïve-midpoint values within 5×10^{-4}, and it does). At the 410M scale point the naïve midpoint damage is 2.0542 dex against the stronger endpoint's corpus-level $H = 1.2120$ dex, and the per-layer hybrid barrier reads:

switch layer l	hybrid H	+dex above endpoint
l=1	1.3006	+0.0885
l=2	1.2991	+0.0871
l=3	1.3048	+0.0927
l=4	1.3029	+0.0909
l=5	1.2920	+0.0799
l=6	1.2862	+0.0742
l=7	1.2845	+0.0724
l=8	1.2814	+0.0694
l=9	1.2822	+0.0701
l=10	1.2821	+0.0700
l=11	1.2877	+0.0756
l=12	1.2865	+0.0744
l=13	1.2908	+0.0788
l=14	1.2912	+0.0791
l=15	1.2952	+0.0831
l=16	1.3016	+0.0896
l=17	1.2943	+0.0822
l=18	1.2909	+0.0788
l=19	1.2909	+0.0789
l=20	1.2914	+0.0793
l=21	1.2999	+0.0879
l=22	1.3021	+0.0901
l=23	1.3056	+0.0936

The worst-layer recovery fraction is 0.9545 (best at l=8, $H = 1.2814$; worst at l=23, $H = 1.3056$), above the pre-registered repair threshold of 0.5, so representation-level repair succeeds on both tested modern-pair scales (0.8985 at 160 M, 0.9545 at 410 M). The scope is as specified: one stitch class (layer-wise ridge, single direction), one confounded pair family; it is not a scaling claim and the reverse direction is unmeasured. The verdict file is `logs/mb44_verdict.txt` (write-once), and the numbers above come from `results_mb44/e5.json`.

MB45: direction and teacher measurements at the 410 M scale point. The displacement-direction probe (MB40, experiment E3) and the probability-teacher mixture (experiment E4) are re-run construction-identically at 410 M (pre-registered specification `versions/MB45_dirteacher_410m_spec.md`; displacements taken after MB43's best-family permutation, re-loaded from the MB43 record and re-verified endpoint-invariant to 10^{-4} relative inside the probe; the cross-check gate re-verifies the endpoint values to 5×10^{-4}, and it does). The twin readings (ρ: $\log_{10}$ of midpoint ppl over the better of the endpoint and full-displacement ppl; midpoint and t=1 corpus-level H; ISO: isotropic twin, SUB: dose in the displacement's top-8 subspace, COMP: dose in its complement):

intervention	ρ	midpoint H	t=1 H
ISO	4.6504	5.8648	7.1351
SUB	4.0678	5.2896	7.9904
COMP	4.6269	5.8397	7.1302

The isotropic twin's tensorwise-global allocation reads t=1 $H = 13.0073$, far above the layerwise allocation in the ISO row above, reported descriptively as at 160 M. Subspace-vs-complement damage separates by -0.5591 dex (window-paired bootstrap $[-0.6203, -0.4920]$ over 16 windows): the pre-specified predicates give DIR-SUBSPACE-SUPPRESSED, verdict DIR-SUPPRESSED-REPLICATES — the in-subspace shelter of the 160M main pair replicates at 410 M (-6.7851 dex at 160 M, -0.5591 dex at 410 M); the narrowing of the gap magnitude with scale is a two-point observation, not a scaling law. The probability-teacher α-ladder (computed with `logaddexp`; the Jensen check passes) reads:

α	hull H	mean ce (nats)
0.000	1.2120	2.79083
0.125	1.1985	2.75955
0.250	1.1914	2.74322
0.375	1.1876	2.73451
0.500	1.1864	2.73178
0.625	1.1876	2.73465
0.750	1.1915	2.74361
0.875	1.1989	2.76060
1.000	1.2165	2.80109

The hull's relative ensemble advantage is 0.02483 at interior $\alpha^\star = 0.500$ ($H_{\min} = 1.1864$), with endpoint gap 0.00368: the teacher advantage persists below the stronger endpoint at this scale as well, at a smaller magnitude than at 160 M (0.14849; descriptive, no threshold class). Scope: one direction construction (top-8 left-singular subspace at matched Frobenius dose), one confounded pair family, bootstrap intervals over 16 windows. The verdict file is `logs/mb45_verdict.txt` (write-once), and the numbers above come from `results_mb45/e3.json` and `results_mb45/e4.json`.

S RANDOM-PERMUTATION AND CROSS-FIT CHECKS

Mechanism check: random permutations at identical marginals (MB41). The attribution of the shell to activation matching (Section 3.2) is tested against the simplest counterfactual with identical aggregates: 20 uniform random permutations per pair in the same three gauge classes as the named matcher families (G_1 shared residual basis, G_2 whole heads, G_3 MLP hidden units), placed by the same row/column rules as the matchers but using *no* activation statistics (pre-registered specification MB41; endpoint-invariance gate: relative deviation $\leq 7.6e-06$ across draws and pairs). Per pair we record the share of the naïve log-barrier recovered by the best random draw ($1 - \log B_{\mathrm{rand,min}} / \log B_{\mathrm{naive}}$) against the matched family's ($1 - \log B_a / \log B_{\mathrm{naive}}$). The mediation is *partial*: median 0.232 (random) vs 0.374 (matched) across the 15 pairs, with `u24` at -0.099 and `u56` at $+0.551$; the shares split by recipe class — on the exploded-barrier sampled-U hosts (`u40`–`u64`) uniform permutations alone recover roughly half (up to $+0.551$); on anchor-class pairs the recovered share stays at or below 0.27. Verdict: P-PARTIAL-LMC (random permutations recover part, but not all, of the matched share); every per-pair row is listed in Table 7.

Selection-bias check: cross-fit permutations on a disjoint window draw. The matchers' permutations were selected on the first window draw; the residual could in principle reflect that same-window selection. We therefore re-apply each pair's fixed best (draw-1, OT-Sinkhorn) permutations to the midpoint activations of a second, disjoint window draw of the same 16 windows of 1024 positions (pre-registered specification MB41), and record ρ_a with draw-1 permutations on draw-2 windows ($\rho_\times$) against the same-draw reference (ρ_{ref}: draw-2 permutations on draw-2 windows), residuals on the window-mean arithmetic basis of the main text. All 14 pairs pass the endpoint-invariance audit (permuted-B vs. B's own second-draw endpoint, rel. 1.4×10^{-6}–3.4×10^{-5} per pair, under the pre-specified 3×10^{-3} line). The per-pair shift $\rho_\times - \rho_{\mathrm{ref}}$ has median -0.0058 dex

Table 7: MB41: uniform random permutations at marginals identical to the named matcher families recover much less of the naive barrier than activation matching on anchor-class pairs, but up to half on the exploded sampled-U class — the gauge reading is recipe-class-conditioned, and the matched-family attribution is partial, not exclusive.

Pair	$\text{share}_{\text{rand}}$ (best of 20)	$\text{share}_{\text{match}}$	$\log B_{\text{rand,min}}$ (dex)
anchor01	$+0.263$	0.400	1.9573
anchor02	$+0.045$	0.305	1.9360
anchor03	$+0.266$	0.449	1.7617
anchor12	$+0.232$	0.461	1.9233
anchor13	$+0.082$	0.323	1.9158
anchor23	$+0.054$	0.333	1.8164
dnlnrms	$+0.080$	0.164	1.7421
k32f	$+0.121$	0.357	1.8010
mb2	$+0.170$	0.374	1.7694
u24	-0.099	0.307	2.1595
u32	$+0.238$	0.361	2.5520
u40	$+0.487$	0.680	2.6698
u48	$+0.487$	0.675	2.4677
u56	$+0.551$	0.733	2.2220
u64	$+0.514$	0.679	2.2575

and median magnitude 0.0346 dex (class medians $+0.0043$ dex anchor, -0.0246 dex loop), so the same-window selection-bias check reads NEGLIGIBLE against the pre-registered 0.05 dex line (Table 8); the residual is a property of the checkpoint pair, not of the window set used to pick the permutations.

Table 8: Cross-fit check: fixed draw-1 matching permutations applied to a disjoint second window draw move the trained-depth residual by a median -0.0058 dex ($n = 14$; median magnitude 0.0346 dex) — NEGLIGIBLE by the pre-specified 0.05 dex predicate.

Pair	$\rho_\times$ (dex; draw-1 perm)	ρ_{ref} (dex; draw-2 perm)	$\Delta = \rho_\times - \rho_{\text{ref}}$ (dex)
anchor01	1.2555	1.3052	-0.0498
anchor02	1.2234	1.1484	$+0.0750$
anchor03	1.2366	1.2380	-0.0014
anchor12	1.2317	1.2217	$+0.0100$
anchor13	1.1939	1.2069	-0.0130
anchor23	1.2093	1.1792	$+0.0301$
u24	1.1868	1.1735	$+0.0134$
u32	1.3996	1.4386	-0.0390
u40	1.3086	1.5207	-0.2120
u48	1.2003	1.2104	-0.0101
u56	1.1691	1.3586	-0.1894
u64	1.1468	1.1306	$+0.0162$
mb2	1.2975	1.1476	$+0.1499$
k32f	1.0697	1.1726	-0.1029

T ALIGNMENT BENEFIT AND ENDPOINT DISTANCE ARE WEAKLY ASSOCIATED ACROSS PAIRS

Across all 23 pairs with a recorded matcher permutation, the alignment benefit $R = \log \mathrm{B}_n - \rho_a$ ranges from -0.22 to $+3.63$ dex, yet the permutation moves the endpoints' relative weight-space distance by at most 4.4% (mean 0.1%; Spearman $|\rho| \leq 0.24$ for R against pre-/post-permutation gap and gap-closing metrics; ≤ 0.07 on the principal-15 subset). This weak association makes relative weight distance an inadequate explanation of the observed matching gains. It is consistent with a role for functional correspondence, but does not isolate that mechanism. (*Exploratory census: no pre-registered band, no multiple-comparison correction; the unique-tensor gap convention is primary; the tied-embedding effect is reported on two explicit bases —*

(i) its share of the naïve squared gap (unique basis, mean over the 23 pairs) 40.3%*; (ii) the mean per-pair inflation of the relative naïve gap from re-counting the tied embedding* 1.77% *(*`versions/mb24b_embshare_addendum2.txt`*).)*

U EXTENDED DISCUSSION OF LIMITATIONS

This section expands the "Limitations and future work" paragraph of Section 8; Appendix M.1 gives the scope statement of the individual measurements.

Scale and statistical units. All controlled measurements use one 35 M weight-tied family trained on 0.25 B tokens (median trained-depth endpoint perplexity 55–58). Most configurations contribute one seed pair, and the six anchor pairs reuse four checkpoints, so independent units number well below fifteen. Evaluation windows are not training replicates, pairs sharing checkpoints are not independent replications, and several intervals are wide (e.g., $[0.07, 2.77]$ dex for the raw recipe contrast). That contrast is observed in a population with ablations and dependent anchor pairs, not as a controlled single-factor effect of sampling. Checkpoint-aware resampling does not remove the pairing-design confound, and its removal by alignment does not mean that every pair's barrier is mostly removable or that recipe has no functional effect.

Measurement scope. Achieved residuals are upper bounds for the tested matchers and estimators, not minima over all valid permutations or bounds over other symmetries or repairs, and individual pair residuals are window-dependent. The two perplexity estimators are not interchangeable, so the best-of-four (12.5–24.2×) and original-arm (18.8–112.2×) ranges are not directly comparable; 0.5 dex is a descriptive cutoff, not a universal definition of linear mode connectivity; and b_R is not the maximum excess along the full path. Norm-matched twins reproduce the residual's scale without showing that norm alone explains each pair (pairwise classification 6/3/3/2), and the dense control does not establish equal magnitude across architectures.

Transfer. On three public Pythia pairs, which confound training data and initialization, the best named matcher removes only 2.4%, 5.0%, and 0% of the log-barrier at 70M, 160M, and 410M, and the direction contrast reverses; the large alignment benefit is therefore a scoped result for the 35 M population. The six width pairs are not compute-matched (each receives 0.246 B tokens), show no monotone trend, establish no scaling law, and do not test the previously studied temperature calibration (Ito et al., 2026). The budget sibling is a same-trajectory comparison rather than an independent-mode replication, and its better trained-depth interpolation does not bring better deep-runtime behavior.

Repairs and consolidation. Continued-training recovery is slower on pairs whose twin does not track the aligned residual, and deep-runtime damage can persist after trained-depth recovery (Appendix F.2). REPAIR was not designed for weight-tied recurrent computation; our REPAIR comparison covers only the tested configurations and does not show that renormalization-based repair must fail in such models. Depth-ramped interpolation schedules did not keep intermediate points usable on the two loop pairs with persistent deep barriers (Appendix M). Distillation's advantage is in deep-runtime quality, not trained-depth quality on every rung: at full budget, pure cross-entropy continuation beats the distilled student at both runtime depths (annealing the distillation weight to zero recovers 65% of this deficit); on the fresh closer pair, plain continuation has larger captured fractions (1.82/1.77 versus 1.38/1.35, arithmetic/corpus); and at matched compute, plain continuation is better at trained depth at both comparison points. The 50.1% win rate motivates rather than establishes complementarity, and each consolidation setting uses one or two pairs, so the results support budget- and pair-dependent consolidation rather than a general repair advantage over additional training. We do not compare against every merging algorithm or against representation- or circuit-level repairs.