OUTPUT-HEAD FACTORIZATION IMPROVES LANGUAGE MODELS AT NO INFERENCE COST

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing interventions on a language model's output head, including weight tying, learned output maps, mixture models, and low-rank variants, often combine changes to parameter sharing, rank, initialization, or function class. We separate parameter sharing from training-time factorization with matched interventions that change one named axis at a time. Multiplicative factorization of the output head substantially lowers perplexity across the evaluated corpora and architectures, yet folds to a single output projection before inference with no factorization-specific computation. Across dense tied and untied references, perplexity differs by at most 1.8%, whereas the best factorized endpoint is 14.9% lower than its initialization-matched direct reference. The main causal result holds the initial output matrix fixed and changes only its factor split, lowering perplexity by 7.5%. Direct addition provides strong alternatives but does not reach the best factorized endpoint. A local update diagnostic characterizes how factorization changes the output-head update; the endpoint effect is established by intervention. Together, these results show that training-time parameterization of the output head is a consequential deployment-neutral design choice.

1 INTRODUCTION

Language models choose each next token with an output head, a linear map from a hidden state to vocabulary scores. Prior work has modified this interface through tying or decoupling input and output embeddings, learned output maps, and lower-rank or alternative output parameterizations (Press & Wolf, 2017; Inan et al., 2017; Chung et al., 2021; Gulordava et al., 2018; Sainath et al., 2013; Yang et al., 2018). These methods are usually evaluated as whole architectures, so a comparison can simultaneously change parameter sharing, initialization, output rank, factor placement, and the gradient path from the classifier to the input embedding. Weight tying and untying are the usual ways of varying parameter sharing, and they move these other axes as well. We therefore treat tying as one axis in a matched design that separately varies sharing, initialization, factorization, factor placement, and classifier-to-embedding coupling.

We separate parameter sharing from training-time factorization and ask how each choice shapes the model produced. For a deployed output matrix W, a factorization $W = AR$ writes it as a product of trainable factor matrices A and R. The product can be multiplied out before inference, leaving the same linear map while training acts on A and R rather than directly on W. Deep-linear analyses show that this changes optimization dynamics (Saxe et al., 2014; Arora et al., 2018; Du et al., 2018), and structural-reparameterization methods likewise train expanded linear graphs that fold into compact inference graphs (Guo et al., 2020; Ding et al., 2023).

Every next-token loss directly updates the output head, and its rows correspond to identifiable vocabulary items. Its linear factors can be folded exactly into the ordinary classifier before deployment, so a benefit would change the model produced by training without adding factor-induced deployed projections, parameters, FLOPs, or latency relative to a depth-1 head with the same tie convention. Other isolated linear components can be treated similarly; the output head makes the optimization signal and practical consequence directly observable.

Our primary comparisons hold the folded initial classifier, its initial logits, data order, optimizer, and full output rank fixed wherever the contrast permits, then vary one named property of the training-time head. We sweep factorization *depth*, *tie*, *initialization gauge*, *factor placement*, and *side* (input, output,

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.

or both). Every full-width factorial arm retains the rank capacity of a conventional output head, so the experiments concern the parameterization used during optimization rather than an imposed internal rank constraint. We also state the remaining bundle in each contrast rather than treating all comparisons as single-variable interventions.

The empirical sections have distinct roles. The broad endpoint and factorial studies establish where factorization improves the deployed model and which design choices accompany that improvement. The initialization gauge is the main causal test. The local diagnostic characterizes the resulting update map, while early-loss screening evaluates a practical selection signal for the studied arms.

Together, the experiments show that a foldable output-head factorization can improve the trained language model without adding factor-induced structure to its deployed classifier. The matched contrasts separate this conclusion from output rank, input-side factorization, direct addition, literal sharing, and the classifier-gradient edge.

2 Experimental setup and parameter accounting

Experimental design and protocol. We use two complementary designs. First, a broad endpoint study compares a full-width, embedding-referenced factorized head with dense tied and untied references across C4 and WikiText-103 with Qwen2-0.5B, and TinyStories with OPT-125M. Second, a controlled Qwen2-0.5B C4 factorial separates parameter sharing, output-side factorization, initialization, factor placement, and classifier-to-embedding gradient coupling. The initialization gauge is then repeated in a larger Qwen model and in Pythia-410m.

The controlled factorial trains Qwen2-0.5B from scratch on sequence-packed C4 and evaluates next-token prediction as token-level PPL $= \exp(\text{mean token NLL})$ over $N = 150{,}000$ held-out C4 validation positions. A calibration sweep selected $T^\star = 1.0$ for the reported arms. The vocabulary has $V = 114{,}179$ tokens and hidden width $d = 896$. Training uses $2{,}546{,}505$ blocks of 256 tokens, about 652M training tokens or 1.32 tokens per model parameter, for $39{,}790$ optimizer steps with effective batch 64, learning rate 3×10^{-4}, bf16, and one epoch. Unless stated otherwise, arms use seed 56; earlier path-decomposition references use seed 42. All factorial arms share one global AdamW optimizer and learning-rate schedule; we do not tune head-specific schedules for the direct reference. Two-seed reruns of selected ablations show ~ 0.2–0.4% variation; the balanced depth-2 gauge arm differs by 0.54% across its two seeds, but this is not a matched gauge-pair replication. Position bootstraps quantify evaluation-position uncertainty rather than training-seed variance.

Head parameterizations. Let $E \in \mathbb{R}^{V\times d}$ be the input embedding, $h \in \mathbb{R}^d$ a hidden state, and sg the stop-gradient operator. The output head produces logits $hW^\top$ through a folded classifier $W \in \mathbb{R}^{V\times d}$. A subscript 0 denotes an initial value, so E_0 is the initial input embedding and W_0 the initial folded output classifier. In a full-width factorization, $A \in \mathbb{R}^{V\times d}$ is the vocabulary-sized factor and $R, S \in \mathbb{R}^{d\times d}$ are square factors. Every factorized output head we study can be collapsed to one such matrix before inference, so the deployed output operation is a single $V \times d$ projection. We use *depth* for the number of trainable multiplicative factors composing W (depth-1: one matrix; depth-2: $W = AR$; depth-3: $W = ARS$). The matched full-width comparisons hold this projection's rank capacity fixed; other controls deliberately vary tie, side, source table, or classifier-gradient coupling to locate the boundary of the effect.

Each arm is defined where it is first used; Appendix B tabulates all parameterizations, their folded W, depth/side, and initial weight for reference. For the low-rank accounting below, k is the factor rank, $P, Q \in \mathbb{R}^{d\times k}$, $O \in \mathbb{R}^{V\times d}$, $C \in \mathbb{R}^{V\times k}$, and $D \in \mathbb{R}^{d\times k}$; P_0, Q_0 denote initial values of P, Q.

Parameter accounting. Folding removes factor matrices from deployment but does not change whether the input and output tables are tied: a tied deployment stores one $V \times d$ table, whereas an untied deployment stores separate E and W tables ($2Vd$ values). Thus "no added inference parameters" refers only to changing factor depth at fixed tie convention; tied and untied endpoints are compared to their respective direct references. At C4 ($V = 114{,}179$, $d = k = 896$), the grounded head $W = EPQ^\top$ adds $2dk = 1{,}605{,}632$ parameters beyond the input table ($Vd + 2dk = 103{,}910{,}016$ total token-interface parameters). The independent-OPQ head has a separate output source table and therefore has $2Vd + 2dk = 206{,}214{,}400$ total token-interface parameters. A direct CD head has $Vk + dk$ parameters beyond E and $Vd + Vk + dk = 205{,}411{,}584$ total. These

Table 1: Full-width factorized endpoint across corpora and architectures (seed 42; canonical $N = 150{,}000$ evaluation positions). The endpoint is a full-width embedding-referenced factorization; tied and untied are dense references.

Setting	factorized	tied	untied
C4, Qwen2-0.5B	43.96	46.79	47.10
WikiText-103, Qwen2-0.5B	25.19	26.57	26.12
TinyStories, OPT-125M	3.81	3.97	4.00

Table 2: Fixed-rank ($k=d$) path decomposition, seed 42. Increments telescope along the listed order.

Step	Intervention	Folded W	PPL	Δ
0	direct low-rank, random init	$CD^\top$	55.61	—
1	+ aligned initialization	$C_0{=}E_0P_0$, $D_0{=}Q_0$	52.53	-3.08
2	+ $V{\times}d$ source-table factorization	$OPQ^\top$, $O_0{=}E_0$	44.43	-8.10
3	+ hard sharing	$EPQ^\top$	43.96	-0.47
	total			$\mathbf{-11.65}$

denominators differ substantially, so we state which one each overhead figure uses. The depth/tie/init factorial is capacity-matched at $k = d$: every arm folds to a $V \times d$ classifier with the same algebraic rank capacity, while the trainable factorization is the intervention. At $d = 896$, each additional $d \times d$ factor contributes $802{,}816$ trainable values, their optimizer state, and one factor multiplication during training; folding replaces the product with the single deployed classifier.

3 Experimental results

3.1 A foldable endpoint across corpora and architectures

We first establish the endpoint pattern across three corpora and two architectures. Table 1 compares a full-width, embedding-referenced factorized head with dense tied and untied references. The factorized endpoint improves on both dense references in every setting. Across settings, the dense tied and untied references differ by 0.7–1.7% relative to the untied reference. This modest dense-reference difference motivates the controlled decompositions of source table, initialization, factor placement, gradient routing, and sharing that follow.

3.2 A detailed decomposition of the C4 endpoint

The broad endpoint pattern in §3.1 motivates resolving the output-head choices that produce it. Table 2 provides an ordered C4 decomposition of the familiar tied-versus-untied comparison. Weight tying identifies the input and output tables, while common tied-versus-untied comparisons also move table source, initialization, factor placement, and gradient routing. In the direct full-width comparison, tied reaches 46.79 PPL and untied 47.10, a 0.67% difference in this cell. The controlled analyses below trace the larger endpoint differences that emerge as these additional design choices vary.

We use an ordered fixed-rank path to expose this bundle. Starting from a random direct factorization, we introduce aligned initialization, a vocabulary-sized source table, and finally hard sharing (Table 2). Each row becomes the next row's baseline, so the increments sum to the endpoint difference by construction. The table is an ordered decomposition of this specified seed-42 path, with the factorial in §3.3 providing matched contrasts for the individual axes.

The largest increment on this path comes from routing the classifier through a separately optimized, vocabulary-sized source table under a fixed factorization (-8.10 PPL); aligned initialization contributes -3.08, and literal hard sharing contributes -0.47. This ordering motivates the factorial in §3.3, where the relevant axes receive matched contrasts rather than being read as intrinsic shares of the path total.

Table 3: Factorial arms, Qwen2-0.5B C4, $k = d$. The main factorial uses $W_0 = E_0$, seed 56; the bottom block changes the folded initialization and is an ablation, not a depth baseline. All fold to a $V \times d$ classifier of identical rank.

Arm	depth	tie	init W_0	side	PPL
untied depth-1	1	no	E_0	out	47.02
untied depth-2	2	no	E_0	out	44.01
untied depth-3	3	no	E_0	out	42.92
tied depth-2 (shared factors)	2	yes	E_0	both	42.69
tied depth-3 (shared factors)	3	yes	E_0	both	41.86
tied depth-3 (hidden-side factor)	3	yes	E_0	both+out	43.70
input-factorized ($E{=}AR$)	2	no	E_0	in	46.86
parallel-in-factor (E_0)	parallel	no	E_0	out	43.13
Folded-initialization ablations: $W_0 = E_0P_0Q_0$					
aligned CD	2	no	$E_0P_0Q_0$	out	52.66
identity split	3	no	$E_0P_0Q_0$	out	43.84
independent OPQ	3	no	$E_0P_0Q_0$	out	44.64
parallel-in-factor ($E_0P_0Q_0$)	parallel	no	$E_0P_0Q_0$	out	57.68

Hard parameter identity is not the only effect bundled into the final step. Sharing the table also opens a gradient path from the classifier back into E, so we isolate that edge with a matched pair that gives both arms an unrestricted residual $\Delta \in \mathbb{R}^{V\times d}$ and varies only the stop-gradient on E. Coupling the edge improves PPL by -1.81 at seed 42 (42.23 against 44.05) and by -1.89 at seed 56 (42.30 against 44.19). Detaching without residual capacity instead collapses the model ($\mathrm{sg}(E)PQ^\top$ reaches 105.25 at $k = d$), because it removes E as classifier-adaptable capacity rather than merely cutting the edge. Thus the small hard-sharing increment in Table 2 should not be read as evidence that every consequence of tying is inert; it says that literal parameter identity is small along this particular path once the preceding choices are in place.

3.3 A FACTORIAL DISSOCIATION OF DEPTH, TIE, INITIALIZATION, PLACEMENT, AND SIDE

The decomposition of §3.2 entangles several axes inside its dominant step, so we now vary them one at a time, all at $k = d$ and capacity-matched to the same folded rank, holding the effective weight and initialization fixed wherever the contrast allows. Table 3 shows the main $W_0 = E_0$ factorial and the folded-initialization ablations that change the initial function. The upper block varies factorization, tying, placement, and side at the shared initial function; the lower block changes that initial function. "Parallel-in-factor" denotes a sum inside a shared factorization, $A(R_1 + R_2)$, rather than a direct sum of output tables.

Depth and tying under matched initialization. Table 3 organizes the main factorial. Within the seed-56 untied $W_0 = E_0$ ladder, each added factor helps: $47.02 \to 44.01 \to 42.92$. The aligned-CD point (52.66 at depth 2) starts instead from $E_0P_0Q_0$; it is an initialization ablation, not a counterexample to that ladder. The hidden-side factor control provides a separate placement test: its third factor acts on h, so its 43.70 does not test another table factor. Moving that factor onto the table gives tied depth-3 with shared factors at 41.86.

Tying depends on the surrounding factorization. A seed-56 full-width rerun gives tied 46.65 versus untied 47.18. At depth 2 with $W_0 = E_0$, the tied shared-factor table (42.69) beats untied (44.01) by 3.0%. Literal table sharing is therefore neither a sufficient account of the result nor interchangeable with the full tie bundle, which also changes the classifier-to-embedding gradient edge (§3.2).

The effect is output-side and initialization-dependent. Input-only factorization ($E{=}AR$) reaches 46.86, close to the E_0-initialized direct output reference (47.02), but it does not match the E_0-initialized output-side depth-2 arm (44.01). Reported input-side gains change the input function class through a learned continuous token map (Batley & Saha, 2026), not a foldable linear factorization of the table; our control isolates the latter. The parallel-in-factor controls vary whether an extra right-factor coordinate is chained or summed inside a shared factorization. Their outcomes depend

strongly on folded initialization: at $E_0P_0Q_0$, the parallel head reaches 57.68 versus 43.84 for the identity split; at E_0, it reaches 43.13 versus 42.92 for depth 3. These parameterizations are distinct from a direct sum of two full output tables, defined below.

Direct-additive controls. Let W_1 and W_2 denote full vocabulary-sized output tables. The capacity-matched residual uses $W = W_1 + UB$, with $U \in \mathbb{R}^{V\times 14}$ and $B \in \mathbb{R}^{14\times d}$; the pure sum uses $W = W_1 + W_2$. The residual reaches 44.70 PPL/0.3341 top-1 (-4.94% versus depth-1; $+11.76\%$ versus balanced depth-3 at 40.00). The larger pure sum reaches 41.36/0.3388 (-12.04% versus depth-1), but remains $+3.40\%$ behind balanced depth-3 at 40.00. Thus direct addition is a strong training-time baseline: even doubling the output table improves the endpoint substantially. All paired-position CIs exclude zero (App. C); the best factorized endpoint remains lower than either additive alternative.

A better deployed head. Tied depth-3 with shared factors reaches 41.86 versus its tied seed-56 depth-1 reference at 46.65, with lower perplexity (-10.3%). The untied balanced depth-3 gauge arm reaches the best observed point estimate, 40.00, versus the $W_0 = E_0$-matched untied seed-56 depth-1 reference at 47.02, with lower perplexity (-14.9%). The randomly initialized untied seed-56 reference is 47.18 (§2). Each factorized head folds offline to one $V \times d$ classifier, so it uses the same inference FLOPs, parameter count, and latency as its depth-1 head with the same tie convention; tied and untied families have different deployed table counts (§2). During training, depth-2 and depth-3 retain one and two additional $d \times d$ factors, respectively, with their optimizer state (§2); all arms use the same 39,790-step AdamW protocol.

3.4 Why a factorization can train differently: local update geometry

A product of factors and its folded matrix implement the same linear operation, but a factor update is not a direct update of that matrix. For simultaneous Euclidean gradient descent with step size η on $W = AR$, let $G = \partial\mathcal{L}/\partial W$. One step gives

$$\Delta W \;=\; -\eta\big[(AA^\top)G + G(R^\top R)\big] + O(\eta^2). \tag{1}$$

Thus, to first order, the folded weight follows the plain gradient G preconditioned by the factor Gram matrices $AA^\top$ and $R^\top R$, the familiar deep-linear distinction between optimizing a product and optimizing its factors (Arora et al., 2018; Saxe et al., 2014). At depth-1, $\Delta W = -\eta G$. Our training runs use AdamW, so Eq. 1 is not their literal update; it is a local Euclidean-GD diagnostic of what the factorization changes. We track its *misalignment* $\cos(\Delta W, -G)$ and per-frequency-bucket *tail amplification*, $\|\Delta W_{\text{bucket}}\|/\|G_{\text{bucket}}\|$.

Measuring the diagnostic at a matched checkpoint reference. Reading each arm at its own trained checkpoint would combine the factorization-induced update with where that arm converged and with the arbitrary factor *gauge*, since $W = (AM)(M^{-1}R)$ leaves W unchanged while changing $AA^\top$ and $R^\top R$. At each scale, we instead take one shared reference matrix $W^\star$, the trained input embedding from the final imbalanced checkpoint, and evaluate every hypothetical depth on the same $150{,}000$ held-out C4 positions. This reference is used only to make the one-step diagnostic comparable across hypothetical factorizations; it is not an additional trained arm. The loss, G, and backbone gradient $\partial\mathcal{L}/\partial h$ are then identical across depths within that scale; only the local ΔW diagnostic varies. For a depth-L product in the init gauge, where the vocabulary-sized table is $W^\star$ and the $L-1$ interior factors are I,

$$\Delta W = -\eta\big[G + (L{-}1)(W^\star W^{\star\top})G\big] + O(\eta^2). \tag{2}$$

The table factor contributes $-\eta G$; each interior factor contributes a token-Gram term.

Table 4 reports the resulting full-evaluation diagnostic. Frequency buckets are formed from the same $150{,}000$ labels at cumulative masses $50/90/99\%$. At both scales, this hypothetical init-gauge diagnostic turns the update away from steepest descent and produces large very-rare relative-amplification ratios. These ratios are not shares of total update mass, so they do not by themselves establish how much absolute update is allocated to rare rows. At 0.5B, depth 3 yields $1376.99\times$, a ratio of 1.9995 to depth 2; at 1.5B, the corresponding ratio is $767.38/383.70 = 1.9999$. These near-exact doublings are consistent with Eq. 2 and validate its first-order implementation: each additional identity-initialized interior factor contributes one more token-Gram term.

Table 4: Matched-$W^\star$ first-order Euclidean-GD diagnostic on 150,000 held-out C4 positions. At each scale, $W^\star$ is the trained input embedding from the final imbalanced checkpoint; all hypothetical heads at that scale use this same reference. Frequency buckets are defined from these labels.

	0.5B (d=896)		1.5B (d=1536)	
depth	overall cos	very-rare relative amplification	overall cos	very-rare relative amplification
1	1.000	1.00×	1.000	1.00×
2	0.107	688.67×	0.110	383.70×
3	0.087	1376.99×	0.083	767.38×

Table 5: Gauge intervention. The seed-56 init/balanced rows are matched pairs with identical $W_0 = E_0$, depth, tie, side, optimizer, and data; the final row repeats the balanced depth-2 configuration at seed 57. $I_{A,R}$ is the normalized factor-Gram mismatch defined above, measured at the earliest and final retained checkpoints.

arm	init gauge	depth/seed	final PPL	$I_{A,R}$ (earliest→final)
untied	init (imbalanced)	2/56	44.01	$0.98 \to 0.995$ (stays imbalanced)
untied	SVD-balanced	2/56	40.70	$0.91 \to 0.976$ (drifts up)
untied	SVD-balanced	3/56	**40.00**	$0.95 \to 0.984$ (drifts up)
untied	SVD-balanced replicate	2/57	40.92	$0.907 \to 0.976$ (drifts up)

Matched interventions identify endpoint effects. The relative-amplification diagnostic is evaluated in the concentrated initialization gauge. For $E_0 = U\Sigma V^\top$, its table-factor Gram term contains $U\Sigma^2 U^\top$, whereas the balanced initialization uses $U\Sigma U^\top$. The balanced intervention reaches the better endpoint (§3.5), and the concentrated-gauge diagnostic characterizes a constrained local update. The induced update is hypothetical and instantaneous, whereas final perplexity is a property of the full trajectory. The six-arm checkpoint-local rank analysis in App. E provides descriptive correlations ($|\rho| \leq 0.4$). Matched interventions identify the endpoint effect (§3.5), while the local diagnostic supplies its update-geometry characterization.

3.5 A CAUSAL DEGREE OF FREEDOM: THE INITIALIZATION GAUGE

The geometry of §3.4 depends on the gauge, and the gauge correlates only weakly with perplexity, so correlation has taken us as far as it can. What remains is to intervene on it. We hold the folded initial weight $W_0 = E_0$, the depth (2), the tie (untied), the side (output), the seed, the optimizer, the learning rate, and the data all fixed, and change only how that weight is split into factors at initialization. The init gauge puts every bit of mass in the table ($A_0 = E_0$, $R_0 = I$), which is maximally imbalanced; the SVD-balanced gauge splits $E_0 = U\Sigma V^\top$ evenly ($A_0 = U\Sigma^{1/2}$, $R_0 = \Sigma^{1/2}V^\top$). For depth three, the same SVD is split across three factors as $A_0 = U\Sigma^{1/3}$, $R_0 = \Sigma^{1/3}$, and $S_0 = \Sigma^{1/3}V^\top$. Both constructions fold to the identical $W_0 = E_0$ and produce identical initial logits, so this is a single-variable intervention on a quantity the model's initial function cannot see.

Factor imbalance. For the table factor A and first interior factor R, we report

$$I_{A,R} = \frac{\|A^\top A - RR^\top\|_F}{\|A^\top A\|_F + \|RR^\top\|_F} \in [0, 1].$$

Zero denotes matched factor Gram matrices, whereas values near one denote a split concentrated in one of the compared factors. The SVD-balanced construction has $I_{A,R}(0) = 0$ exactly; the reported arrows begin at the earliest retained checkpoint after training has started. For depth-three heads, the table reports this same first-pair statistic $I_{A,R}$, not a single aggregate over all three factors.

Table 5 reports the matched 0.5B interventions and their checkpoint trajectory summaries. In the depth-2 pair, changing only the initialization gauge yields lower final perplexity (-7.5%; paired position bootstrap over 150,000 eval positions, 95% CI $[-7.9, -7.2]$). At depth-3, the matched pair reaches 40.00 versus 42.92 (lower perplexity -6.8%, CI $[-7.1, -6.5]$). The balanced depth-2 configuration repeats at seed 57 with final PPL 40.92, 0.22 above its seed-56 value of 40.70. The

Table 6: Matched depth-2 gauge replication. In both panels, values are balanced minus concentrated initialization; endpoint CIs are paired-position 95% intervals. Endpoint PPLs are rounded to two decimals, while deltas use unrounded per-position NLL estimates.

	Endpoint perplexity			
model	concentrated	balanced	ΔPPL	95% CI
Qwen2-0.5B	44.01	40.70	-7.5%	$[-7.9, -7.2]$
Qwen2-1.5B	36.16	34.99	-3.26%	$[-3.6, -2.9]$
Pythia-410m	41.90	39.80	-5.02%	$[-5.4, -4.7]$
Per-frequency ΔPPL (frequent / mid / rare / very-rare)				
Qwen2-0.5B	-0.9% / -12.0% / -20.1% / -20.3%			
Qwen2-1.5B	-0.2% / -4.7% / -12.5% / -9.8%			
Pythia-410m	-0.7% / -7.6% / -15.2% / -16.6%			

Table 7: Third-factor comparison from the balanced depth-2 configuration. CIs are paired-position 95% intervals; PPLs are rounded to two decimals, while deltas use unrounded per-position NLL estimates.

model	balanced depth-2	balanced depth-3	ΔPPL (95% CI)
Qwen2-0.5B	40.70	40.00	-1.72% $[-2.06, -1.37]$
Qwen2-1.5B	34.99	35.14	$+0.46\%$ $[+0.09, +0.80]$

gains are tail-concentrated: at depth-2 the per-bucket ΔPPL runs -0.9% on frequent tokens, -12.0% on mid, -20.1% on rare, and -20.3% on very-rare; at depth-3 the frequent bucket is unmoved (-0.2%, n.s.) while mid, rare, and very-rare move -11.4, -18.9, and -20.6%. Each matched pair fixes the initial function and training protocol, so the initialization gauge *causally* affects the endpoint.

The gauge effect replicates at $3\times$ scale and on a second architecture. Table 6 shows that balanced initialization wins in the larger Qwen model and in Pythia-410m as well as in the main 0.5B factorial. It also organizes the reproduced frequency pattern: the effect is small on frequent tokens and larger in the less frequent buckets. The matched gauge intervention therefore has the same sign across a distinct architecture and tokenizer, placing the effect in factorized softmax heads under AdamW rather than in one model family. Its smaller 1.5B magnitude remains budget-confounded, because the larger model receives fewer tokens per parameter.

The depth-2 gauge replicates at 1.5B; third-factor behavior varies with the shared budget. Table 6 gives the replicated depth-2 gauge pair, and Table 7 gives the separate third-factor comparison. At 0.5B, adding the factor improves the balanced endpoint and is tail-concentrated; at 1.5B it moves slightly upward. The shared token budget couples model size with training exposure, so these results provide a replicated depth-2 gauge effect while motivating additional factor-count comparisons at matched training exposure.

The initial gauge steers the training trajectory. Table 5 reports how the balanced starts move from the SVD split toward higher first-pair mismatch: the depth-2 runs move from 0.91 to 0.976, and the depth-3 arm moves from 0.95 to 0.984, approaching the concentrated arm's 0.995. The arms nevertheless reach distinct endpoints; the 40.00 arm has lower final $I_{A,R}$ (0.984) than the 44.01 arm (0.995). The initial balanced split therefore steers the early trajectory and the endpoint it reaches, while final first-pair mismatch summarizes where the factors finish.

The gauge changes the update map, and the effect is robust. Since the two gauges share an identical folded weight and loss, there is no loss force along the gauge orbit. The factor gradients, however, are gauge-dependent: $g_A = g_W R$ and $g_R = A^\top g_W$. Consequently even the Euclidean-GD update map in Eq. 1 depends on the factor split; AdamW adds state-dependent, coordinate-wise dynamics to this dependence. Decay-free reruns retain balanced–imbalanced gaps of -7.4% untied and -8.2% tied. Plain SGD and SGD with momentum reproduce the gauge advantage's sign (App. H).

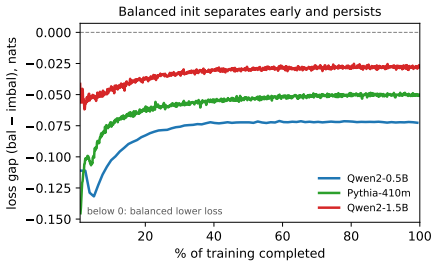

Figure 1: Training-loss gap between balanced and imbalanced gauge arms (balanced − imbalanced, nats). Beyond 1% of training, the gap is negative on every model. On the complete 0.5B run, it settles near -0.072 nats from halfway through training to the final checkpoint.

Deep-linear and implicit-bias theory motivate treating initialization shape as a meaningful intervention: in underdetermined factorized problems, initialization and update scale can affect the selected solution (Gunasekar et al., 2017; Woodworth et al., 2020; Azulay et al., 2021). That theory does not predict the sign, magnitude, or frequency structure of the perplexity effect here, nor the behavior of a full language model trained with AdamW. We establish those properties by intervention.

3.6 Early-loss screening in an eight-arm set

The gauge intervention establishes an endpoint effect, and local geometry supplies a first-order view of the associated update map at a fixed output matrix. We complement these analyses by testing whether early training loss can screen seven factorized candidate arms and one parallel-in-factor control from the shared Qwen2-0.5B C4 factorial. At each logged step, we compare loss with final perplexity using Spearman correlation and ask whether the lowest-loss arm is the eventual winner.

The eventual winner is identified by step 2000, 5% of training, and never lost thereafter. Spearman $\rho(\text{loss}_t, \text{final PPL})$ reaches $+0.93$ by 7.5% and $+0.98$ by 8.8% of the run; every early-loss pick from 1.3% of training beats the tied references. The matched gauge pairs also separate early on Qwen2-0.5B, Pythia-410m, and Qwen2-1.5B (Fig. 1). Beyond a sub-1% warmup transient on 1.5B, the balanced arm has lower loss on every model, with a magnitude that tracks the endpoint gap.

Early loss complements the causal intervention and endpoint evaluation by providing a low-cost screening signal for the evaluated head choices. The complete eight-arm curve is in Table 8.

4 Background and related work

Training-time parameterization. Deep-linear analyses show that optimizing a product of factors differs from optimizing its product directly (Saxe et al., 2014; Arora et al., 2018); Du et al. (2018) characterizes balancedness under gradient flow. More broadly, implicit-bias work shows that initialization scale and relative factor scales can affect the learned solution in underdetermined factorized models (Gunasekar et al., 2017; Woodworth et al., 2020; Azulay et al., 2021). Structural reparameterization trains expanded linear components and folds them for deployment (Guo et al., 2020; Ding et al., 2023). PiSSA also uses an SVD-derived factor initialization, but for low-rank adaptation of a pretrained frozen matrix (Meng et al., 2024); our setting trains full output heads from scratch and tests the output-head factor split by matched intervention. Weight normalization, μP, and Shampoo are further ways to alter parameterization, scaling, or update geometry (Salimans & Kingma, 2016; Yang et al., 2021; Gupta et al., 2018).

Language-model output interfaces. Weight tying shares input and output embeddings (Press & Wolf, 2017; Inan et al., 2017); later work studies decoupling (Chung et al., 2021) and the dense output-gradient effect on tied embeddings (Lopardo et al., 2026). We test that gradient edge separately. Gulordava et al. (2018) modify a tied architecture with a learned output transformation that decouples the hidden state from tied word-embedding prediction. This is close prior work on a tied output map; our contribution is the controlled dissection of sharing, factor split, factor placement, and classifier-to-embedding coupling. ALBERT factorizes vocabulary embeddings for parameter efficiency (Lan

et al., 2020); and Batley & Saha (2026) changes the input interface through a continuous token map while leaving the output head untied. Unlike our foldable linear control, Leviathan changes the input function class and deployed input computation (Batley & Saha, 2026). Our full-width transformer comparisons retain the folded output classifier's rank capacity and isolate its training-time factorization.

Output rank and the present scope. Direct low-rank output factorization predates transformers (Sainath et al., 2013), while mixture-of-softmax concerns the rank of the context-by-vocabulary log-probability matrix (Yang et al., 2018). In our factorial $k = d$, so the folded classifier retains the algebraic rank capacity of a $V \times d$ projection. The parallel-in-factor control tests addition within a factorized parameterization; Section 3.3 separately tests both a direct-additive residual and a pure $W_1 + W_2$ classifier.

5 LIMITATIONS

Fine orderings within the factorial are single-seed: arms separated by less than 1% are point estimates, and paired position bootstraps quantify evaluation-position rather than training-seed uncertainty. The five-axis factorial is one Qwen2-0.5B C4 cell. The factorial, gauge interventions, and early-loss screen all use a fixed one-epoch budget (1.32 tokens per parameter), so their endpoint ordering and screening behavior need not persist under longer training. All factorial arms share one global AdamW optimizer and learning-rate schedule; we do not tune head-specific learning rates or schedules for the direct reference, so these comparisons do not establish whether such tuning would change the observed gaps. The gauge direction replicates on Qwen2-1.5B and Pythia-410m, and the fixed-rank endpoint spans two architectures and three corpora, but neither the complete factorial nor a budget-deconfounded scale comparison is repeated elsewhere. The gauge intervention identifies a causal endpoint knob, not a complete mechanism: factorization changes the update map, gradient noise, factor conditioning, and step scales together. The capacity-matched additive residual still contains a low-rank product, while the pure direct sum has an extra vocabulary-sized training table; both direct-addition controls are one-seed measurements. Finally, early-loss screening covers one eight-arm Qwen2-0.5B C4 set, with matched gauge trajectories on the additional models, and does not establish a general selection rule for new corpora or architectures.

6 CONCLUSION

A deployed output classifier can be learned through different parameterizations. Output-head factorizations that fold to the same standard $V \times d$ projection reach lower perplexity in the evaluated settings with no added deployed parameters, FLOPs, or latency relative to a depth-1 head at the same tie convention.

The central evidence is a matched initialization-gauge intervention. Holding the folded initial classifier, initial logits, model, data, optimizer, and seed fixed, changing only the factor split lowers perplexity by 7.5%. This function-invisible degree of freedom yields tail-concentrated gains, and its sign holds in the tested Qwen and Pythia configurations and under all three optimizers. The factor split therefore causally steers the training trajectory while preserving the deployed matrix. Controls map the distinct roles of hard sharing and the classifier-to-embedding gradient edge; input-side factorization provides a separate function-class contrast, and direct addition is a strong baseline whose best endpoint remains above the best factorized endpoint.

Deployment-neutral training parameterizations are therefore model-design choices. The local diagnostic characterizes how a factorized head updates differently, while matched interventions identify the endpoint effect. Open questions are the mechanism and whether foldable reparameterizations help attention and MLP projections or survive budgets past one epoch.

AI USE STATEMENT

We used a generative AI coding assistant for software engineering tasks: implementing head parameterizations and diagnostics, orchestrating training/evaluation jobs, and drafting and editing this manuscript. We also used AI for retrieval and discovery, including finding related work. All

experimental design decisions, the choice of controls, and the interpretation of results were made by the authors; all AI-assisted code was reviewed and its outputs verified against logged run artifacts, and all quantitative claims trace to those artifacts. We take responsibility for the final content of this work.

Ethics statement

This work trains language models from scratch on public research corpora and evaluates them on held-out text from the same sources; it involves no new data collection, no human subjects, and no personally identifying information. The trained models are small research artifacts used for controlled comparisons of output-head parameterization, and are not released or deployed as general-purpose systems. Our intervention changes only how the output classifier is parameterized during training and folds to a standard classifier at inference, so it does not alter the model's function class or inference-time computation relative to the softmax baseline. We inherit the known limitations of the underlying corpora, including the possibility of biased or harmful text, and do not address them here.

Reproducibility statement

Section 2 and Appendix I give the full training and evaluation protocol (model, data, packing, steps, batch, LR, precision, seed) and the head-reconstruction and diagnostic procedures. The gauge construction and matched-weight preconditioner computation are specified in §3.4–§3.5 and Appendix I. Implementation and checkpoint artifacts are not publicly available at submission while corporate release approval is pending; Appendix I states the resulting reproducibility boundary.

References

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. arXiv:1802.06509.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 468–477, 2021.

Reza Batley and Sourav Saha. Leviathan: Decoupling input and output representations in language models. *arXiv preprint arXiv:2601.22040v2*, 2026.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.12821.

Xiaohan Ding, Honghao Chen, Xiangyu Zhang, Kaiqi Huang, Jungong Han, and Guiguang Ding. Re-parameterizing your optimizers rather than architectures. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2205.15242.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. arXiv:1806.00900.

Kristina Gulordava, Laura Aina, and Gemma Boleda. How to represent a word and predict it, too: Improving tied architectures for language modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2936–2941, 2018. ACL Anthology D18-1323.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shuxuan Guo, Jose M. Alvarez, and Mathieu Salzmann. Expandnets: Linear over-parameterization to train compact convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. arXiv:1811.10495.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1611.01462.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

Antonio Lopardo, Avyukth Harish, Catherine Arnett, and Akshat Gupta. Weight tying biases token embeddings towards the output space. *arXiv preprint arXiv:2603.26663*, 2026.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 157–163, 2017. arXiv:1608.05859.

Tara N. Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6655–6659, 2013.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6120.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of the 33rd Conference on Learning Theory*, pp. 3635–3673, 2020.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, 2021.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank rnn language model. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1711.03953.

A EARLY-LOSS SELECTION CURVE

Table 8: Early-loss selection curve over eight arms: untied depth-2 init, untied depth-2 balanced, untied depth-3 init, untied depth-3 balanced, tied depth-2 (shared), tied depth-3 (shared factors), identity split, and the parallel-in-factor control. The tied softmax reference is 46.79; total training is 39,790 steps. *winner-hit* is best-arm recovery, not token accuracy.

step	% train	Spearman(loss$_t$, final PPL)	winner-hit	pick beats softmax
500	1.3%	+0.21	no	yes
1000	2.5%	+0.67	no (picks 2nd)	yes
2000	5%	+0.76	**yes**	yes
3000	7.5%	+0.93	yes	yes
3500	8.8%	+0.98	yes	yes
$\geq$ 4000	$\geq$ 10%	$\sim$ 0.98 (plateau)	yes	yes

B HEAD PARAMETERIZATIONS

Table 9: Head parameterizations. All fold to one $V \times d$ classifier at inference. $P, Q, R, S \in \mathbb{R}^{d\times d}$ at full width; "Init W_0" is the initial folded weight.

Arm	Folded W	Depth / side	Init W_0
tied	E	1 / shared	E_0
untied	free W	1 / output	random
untied-E_0 depth-1	W	1 / output	E_0
untied-E_0 additive residual	$W_1 + UB$ ($\operatorname{rank} UB \leq 14$)	additive / output	E_0
untied-E_0 full direct addition	$W_1 + W_2$	additive / output	E_0
untied-E_0 depth-2	AR	2 / output	E_0 ($A_0 = E_0, R_0 = I$)
untied-E_0 depth-3	ARS	3 / output	E_0
parallel-in-factor (E_0)	$A(R_1 + R_2)$	parallel / output	E_0
Fixed-P0	$OP_0Q^\top$, P_0 frozen	3 / output	$E_0P_0Q_0^\top$
grounded	$EPQ^\top$	3 / output (shares E)	$E_0P_0Q_0^\top$
independent OPQ	$OPQ^\top$, $O_0{=}E_0$	3 / output	$E_0P_0Q_0^\top$
direct CD	$CD^\top$, random	2 / output	random
aligned CD	$CD^\top$, $C_0{=}E_0P_0$, $D_0{=}Q_0$	2 / output	$E_0P_0Q_0^\top$
identity split	$ARQ^\top$, $A_0{=}E_0P_0$, $R_0{=}I$	3 / output	$E_0P_0Q_0^\top$
parallel-in-factor ($E_0P_0Q_0$)	$C(D_1{+}D_2)^\top$, $D_{2,0}{=}0$	parallel / output	$E_0P_0Q_0^\top$
tied depth-2 (shared)	AR shared in/out	2 / both, tied	E_0
tied depth-3 (hidden-side)	$(AR)Q^\top$, $Q_0{=}I$	3 / both+output-Q	E_0
tied depth-3 (shared)	ARS shared in/out	3 / both, tied	E_0
input-factorized	$E{=}AR$ in, free W out	2 / input	E_0 (in)
residual twins	$EPQ^\top{+}\Delta$ vs. $\mathrm{sg}(E)PQ^\top{+}\Delta$	edge control	$E_0P_0Q_0^\top$

C DIRECT-ADDITIVE CONTROLS

Table 10: Direct-additive controls on the shared 150,000-position C4 evaluation (seed 56, $T^\star = 1.0$). The residual is near-matched to untied depth-3 training capacity; the pure sum has an additional vocabulary-sized training table. CIs are paired-position 95% intervals.

arm	PPL	top-1	vs. depth-1	vs. balanced depth-3
$W_1 + UB$, rank 14	44.70	.3341	-4.94% [$-5.24, -4.65$]	$+11.76\%$ [$+11.30, +12.21$]
$W_1 + W_2$	41.36	.3388	-12.04% [$-12.37, -11.71$]	$+3.40\%$ [$+3.03, +3.79$]

D FOLDED-WEIGHT INITIALIZATION: E_0 VERSUS $E_0P_0Q_0$

A factorized head requires a choice of initial folded weight W_0. The natural options are $W_0 = E_0$, initializing the output classifier *as* the input embedding (the tied operating point), and $W_0 = E_0P_0Q_0$, initializing it as the embedding passed through the aligned low-rank transform P_0Q_0 (the natural init of the grounded/low-rank construction). These fold to *different* initial functions, so this is a coarser choice than the initialization gauge of §3.5, which holds W_0 fixed and varies only the factor split.

Empirically E_0 is the better choice, and by a large margin: at depth-2, E_0 init reaches 44.01 PPL versus 52.66 for $E_0P_0Q_0$ (a ~ 8-PPL, $\sim 16\%$ gap); the gap narrows by depth-3 (42.92 vs. 43.84). The model is thus highly sensitive to the folded-weight initialization, but the beneficial direction is not a subtle one: starting the head at the tied solution E_0, a known-good operating point, outperforms starting it at a transformed embedding. We report this as an ablation and do not claim a mechanism for the gap; we use $W_0 = E_0$ throughout the main text, and note that this folded-weight choice is what makes an uncontrolled depth ladder appear non-monotone (§3.3), since the $E_0P_0Q_0$-initialized arms sit higher for reasons of initialization rather than depth.

E CONVERGED-CHECKPOINT GEOMETRY IS UNDERPOWERED FOR ENDPOINT RANKING

Reading each arm's induced-update geometry at its own converged checkpoint and correlating with final perplexity across six seed-56 arms yields descriptive rank correlations ($\cos +0.37$, very-rare $\cos +0.31$, very-rare mass -0.37, imbalance $+0.20$; all $|\rho| \leq 0.4$). With only six arms (and five for imbalance), this diagnostic is underpowered for a rank test. It therefore neither establishes nor rules out an endpoint association. Because every arm contributes its own checkpoint reference and factor gauge, its absolute mass values are not comparable to the matched-$W^\star$, 150,000-position values in Table 4. The table is descriptive of where each arm landed, not a test of which endpoint is better.

Table 11: Converged-checkpoint local geometry and endpoint perplexity (seed 56). This is an underpowered checkpoint-local association diagnostic; its absolute mass values use arm-specific references and are not comparable to the matched-$W^\star$ values in Table 4.

arm	PPL	overall cos	very-rare mass	imbalance
tied depth-3 (shared)	41.86	0.20	183×	0.996
tied depth-2 (shared)	42.69	0.49	40×	0.994
tied depth-3 (hidden-side)	43.70	0.06	1597×	0.997
identity split	43.84	0.08	948×	0.997
untied depth-2	44.01	0.48	44×	0.995
untied depth-1	47.02	1.00	1×	—

F WHY THE UPDATE MAP IS NOT GAUGE-INVARIANT

The two gauges are related by $(A, R) \mapsto (AM, M^{-1}R)$ with M the diagonal $\Sigma^{1/2}$ in the singular basis, so they share the identical folded weight $W_0 = E_0$ and identical loss: at fixed W the loss is exactly constant along the gauge orbit, and there is no loss force along it. The gauge-dependence enters through the update map before adaptivity. By the product rule the same output-space gradient g_W induces factor gradients $g_A = g_W R$ and $g_R = A^\top g_W$ that depend on the gauge, so even a recombined gradient-descent update differs between parameterizations. AdamW then maps these gauge-dependent factor gradients into additional state-dependent future motion. Thus identical W_0 and g_W need not produce identical trajectories.

This is also why the effect is not a per-factor learning-rate artifact: a scalar (or per-factor scalar) step size cannot reproduce a $\Sigma^{1/2}$-shaped, direction-dependent redistribution, and AdamW's per-coordinate normalization cancels a *fixed* coordinate-wise rescale exactly ($cm/\sqrt{c^2 v} = m/\sqrt{v}$), but the gauge's induced rescale co-evolves with A, R, so it is not a fixed rescale and is not cancelled.

Weight decay has a balancing tendency but is not the driver. Decoupled weight decay acts on the raw parameters, and for a fixed product the factor norm $\|A\|_F^2 + \|R\|_F^2$ is minimized by the balanced factorization, so weight decay in isolation prefers balanced. But this is only one term in the AdamW dynamics: empirically the net motion drifts toward imbalance, so the balancing tendency of decay does not dominate. A decay-free ablation confirms that balanced–imbalanced gaps persist without weight decay: -7.4% (95% CI $[-7.8, -7.1]$) untied and -8.2% ($[-8.6, -7.8]$) tied.

G QWEN2-1.5B LADDER TRAJECTORY

A protocol-matched single-seed Qwen2-1.5B replication of the six full-width ladder/reference arms (same $39{,}790$ steps and total tokens as 0.5B) preserves the grounded endpoint advantage while the intermediate ladder increments shrink, including near matched tokens/parameter (0.423 at 1.5B vs. 1.32 at the 0.5B endpoint). The matched gauge endpoint at 1.5B is also smaller than at 0.5B under the shared token budget (§3.5). We treat both observations as a scale-by-training-dynamics interaction and portability boundary, not an asymptotic scale law: several effects are still changing with checkpoint, the scales occupy disjoint absolute-PPL regimes, and 1.5B lacks seed replication. The matched-$W^\star$ geometry diagnostic now uses the same $150{,}000$-position protocol at both scales (§3.4); it reproduces depth-2 misalignment and tail amplification, but the absolute tail multiplier is smaller at 1.5B.

H THE INITIALIZATION BENEFIT HOLDS UNDER SGD AND MOMENTUM

We rerun the matched depth-2 gauge intervention (§3.5) under plain SGD and under SGD with momentum, each at a learning rate tuned for that optimizer (plain SGD 0.3; momentum 0.9 at 0.2), holding the folded initialization $W_0 = E_0$, seed, and data fixed and varying only the factor split. Under both optimizers the balanced initialization reaches lower perplexity and higher top-1 than its imbalanced twin (Table 12), matching the sign of the AdamW result. The initialization gauge is therefore a causal degree of freedom across every optimizer we tested, not a property of adaptive optimization alone.

The same table also corrects a natural but false expectation. Under gradient *flow* the balancedness $A^\top A - RR^\top$ is exactly conserved, which might suggest that (near-flow) plain SGD should preserve a balanced initialization. It does not: at a learning rate high enough to actually train, every optimizer, plain SGD, momentum, and AdamW, drifts a balanced start toward a strongly imbalanced gauge. Finite-step optimization does not inherit the flow's conservation, so "balanced" is not a resting state under any of these optimizers. (The near-balanced trajectories we observed at very small learning rates were an artifact of undertraining, not conservation.) Every comparison here is within a balanced/imbalanced pair, with the drift read from the gauge trajectory; absolute perplexities are not comparable across optimizers, since plain SGD and momentum train transformers far worse than AdamW at this budget.

Table 12: The balanced-initialization advantage replicates beyond AdamW. Each row pair is a single-variable init intervention (identical $W_0 = E_0$, seed, data; only the factor split differs). PPL and top-1 are within-pair; absolute values are *not* comparable across optimizers (SGD and momentum are far more undertrained than AdamW at this budget). Imbalance drifts up under every optimizer, no optimizer preserves a balanced start.

optimizer	init gauge	PPL	top-1	imbalance (start→final)
plain SGD	imbalanced	308	0.184	$0.96 \to 0.96$
plain SGD	balanced	**226**	**0.191**	$0.24 \to 0.83$ (drifts up)
SGD+momentum	imbalanced	454	0.165	$0.96 \to 0.96$
SGD+momentum	balanced	**330**	**0.168**	$0.11 \to 0.73$ (drifts up)
AdamW (§3.5)	imbalanced	44.01	—	$0.98 \to 0.995$
AdamW (§3.5)	balanced	**40.70**	—	$0.91 \to 0.976$ (drifts up)

I REPRODUCIBILITY

All arms: Qwen2-0.5B from scratch, C4, V=114,179, d=k=896, sequence-packed (2,546,505 blocks, 39,790 steps, effective batch 64, LR 3×10^{-4}, bf16, one epoch). Every factorized head folds to a single $V \times d$ classifier at inference; heads are reconstructed from checkpoints by their factor keys. Evaluation: token-level PPL $= \exp(\text{mean NLL})$ over N=150,000 C4 validation positions, self-calibrated $T^\star$=1.0. The matched-$W^\star$ diagnostic uses 150,000 held-out C4 positions at both scales (no backbone backprop after hidden-state collection); gauge imbalance is read from the saved training trajectories. Gauge intervention: SVD-balanced initialization A_0=$U\Sigma^{1/2}$, R_0=$\Sigma^{1/2}V^\top$ from E_0=$U\Sigma V^\top$, so A_0R_0=E_0 exactly.

All numbers in this paper were generated programmatically from retained training checkpoints, evaluation artifacts, and experiment configurations. The paper specifies the full protocol and the head-reconstruction and diagnostic procedures, but the implementation and checkpoint artifacts are not publicly available at submission while corporate release approval is pending. Accordingly, these details support procedural audit; independent end-to-end reproduction awaits approved artifact availability. As an additional internal check, the full C4 pipeline (training, evaluation, and table generation) was re-run independently and reproduced the reported values.