ROLE-DECOUPLED ATTENTION RESIDUALS: SEPARATE QK AND V READS IN BLOCK ATTNRES

Anonymous authors
Paper under double-blind review

ABSTRACT

Block Attention Residuals use one content-dependent depth mixture to construct queries, keys, and values. This couples the representations that determine attention scores to the value content aggregated under those scores. Role-Decoupled Attention Residuals (RD-AttnRes) retain the parent Q/K route and learn a second route for V over the same Block sources. Tying the route queries recovers Block AttnRes; untying them adds one model-width vector per attention layer and no second token-to-token attention operation. After 2B FineWeb-Edu tokens, RD-AttnRes lowers fixed-prefix validation NLL in all five paired seeds at both 120M and 343M parameters, with mean within-seed perplexity reductions of 2.97% and 2.43%. In a single-seed 1.5B pair trained for 10B tokens, RD lowers endpoint perplexity by 2.22%, leads at all 20 shared quick-validation checkpoints, and reaches the same validation NLL with 8.89% fewer training tokens on average across 17 interpolation-supported targets. The four primary zero-shot scores improve by 1.77 percentage points on average. Tested controls and fixed-checkpoint interventions show that trained models learn and use the $QK \mid V$ separation.

1 INTRODUCTION

A PreNorm Transformer exposes each sublayer to a residual state that sums all earlier updates (Xiong et al., 2020). Attention Residuals replace this fixed sum with a learned read over depth. Their blockwise form, Block Attention Residuals (Block AttnRes), routes over the embedding, completed-block outputs, and the current block's partial state (Kimi Team et al., 2026). At each attention layer, one mixture supplies all three projections: Q, K, and V.

Queries and keys determine the token-attention weights; values provide the content aggregated under those weights. Block AttnRes nevertheless forces both roles to draw from the same depth distribution. We ask whether this coupling costs performance when the source set, MLP route, write-back rule, and token-attention operation remain unchanged.

Role-specific cross-layer reads already appear in MUDDFormer and DeepCrossAttention (Xiao et al., 2025; Heddes et al., 2025). RD-AttnRes asks the narrower question inside Block AttnRes: what changes when only its shared QKV depth read is split into QK and V?

Role-Decoupled Attention Residuals (RD-AttnRes) introduce a second learned query over the parent source stack. One mixture feeds Q/K and the other feeds V; both enter the same causal self-attention. The MLP route, source construction, write-back, and final mixer remain unchanged. For L attention layers of width d_{model}, the added queries contain Ld_{model} parameters in total.

Tying the two queries recovers Block AttnRes exactly in real arithmetic. When the centered depth keys have nonzero rank, untying them realizes pairs of depth distributions unavailable to a shared query. Under the additional rank condition in Section 3.3, it also introduces local attention-output directions outside the shared-route update space.

Across five paired seeds at each of the 120M and 343M scales, RD reaches lower NLL after 2B FineWeb-Edu tokens (Penedo et al., 2024), with mean within-seed perplexity reductions of 2.97% and 2.43%. The single 1.5B pair extends the comparison to 10B tokens: RD remains ahead throughout the shared 5–9.75B trajectory, lowers 10B endpoint perplexity, reaches the same validation NLL with 8.89% fewer training tokens on average, and improves all four primary zero-shot metrics.

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.

Execution and redundant-parameter controls remain near the Block parent, the tested role partitions select QK|V as the strongest equal-cost two-route design, and route interventions raise NLL at three fixed checkpoints. At the L endpoint, QK/V divergence is nonzero at all 23 non-degenerate read sites. RD adds below 0.008% parameters at every scale; measured time per token rises by 4.96% at M and 1.51% at L. Across the tested settings, the shared QKV depth read behaves as a genuine modeling constraint rather than a harmless parameter tie.

2 Related Work

Learned cross-layer aggregation has taken several forms. DenseFormer learns input-independent weights for mixing depth states, while Hyper-Connections expands the residual state into multiple streams coupled by learned read, write, and mixing maps (Pagliardini et al., 2024; Zhu et al., 2025). Attention Residuals (AttnRes) compute a token-dependent softmax over previous residual-writing outputs; Block AttnRes groups these outputs into block-level sources and is the direct parent of RD-AttnRes (Kimi Team et al., 2026). Block AttnRes is also deployed in Kimi K3, a 2.8T-parameter sparse MoE with 104B activated parameters that reports strong results across coding, agentic, knowledge, reasoning, and vision benchmarks (Kimi Team, 2026). This system-scale deployment supports the practical viability of the parent architecture; our experiments address the separate question of whether its attention-side read should be role-decoupled.

Cross-layer mechanisms have also treated attention roles asymmetrically. Value Residual Learning mixes the first-layer value representation into later-layer values, while Depth-Attention mixes value states across depth before sequence attention (Zhou et al., 2025; Zeng et al., 2026). DeepCrossAttention constructs Q, K, and V through independent reads over a shared history, and MUDDFormer learns dynamic mixtures for Q, K, V, and the residual stream (Heddes et al., 2025; Xiao et al., 2025). RD takes the narrower QK|V partition suggested by the self-attention computation and embeds it in Block AttnRes while preserving the parent's source construction, MLP route, and write-back. Query tying recovers the parent exactly.

Delta Attention Residuals retain the running residual state and add a routed mixture of past updates; the public Delta-V variant also uses QK|V routing inside this additive operator (Luo et al., 2026a; Luo & Cai, 2026). RD shares the role partition but applies it to the Block AttnRes read, and tying its queries recovers Block AttnRes rather than Delta AttnRes. The parent operator and source/write rule are therefore different. Other AttnRes extensions vary the score representation or feature granularity: Low-Rank AttnRes uses low-dimensional depth keys, and Multi-Head AttnRes assigns separate depth distributions to feature subspaces rather than attention roles (Su, 2026; Luo et al., 2026b).

3 Role-Decoupled Attention Residuals

Standard PreNorm exposes a cumulative state h_ℓ to each sublayer. Figure 1 focuses on the two depth-routed architectures: Block AttnRes constructs one mixture $x_{\ell,c}$ for Q, K, and V, whereas RD-AttnRes supplies $x_{\ell,c}^{\mathrm{QK}}$ to Q/K and $x_{\ell,c}^{\mathrm{V}}$ to V.

3.1 Standard and Block Residual Reads

Standard decoder residuals. A standard PreNorm decoder passes a cumulative residual state through normalized attention and MLP sublayers. Earlier writes enter each later layer with coefficient one rather than as separately weighted depth sources.

Block Attention Residuals. Block AttnRes (Kimi Team et al., 2026) replaces the cumulative attention read with routing over a compact source collection. For attention layer ℓ in block g, let E denote the embedding, C_j the sum of attention and MLP outputs written within completed block j, and R_ℓ the current block's partial sum immediately before layer ℓ. The attention sources are

$$\mathcal{S}_\ell = \begin{cases} (E, C_0, \ldots, C_{g-1}), & \text{at the first attention read in block } g, \\ (E, C_0, \ldots, C_{g-1}, R_\ell), & \text{at later attention reads in block } g. \end{cases} \tag{1}$$

A completed source C_j contains the writes made inside block j; the first attention read starts without R_ℓ.

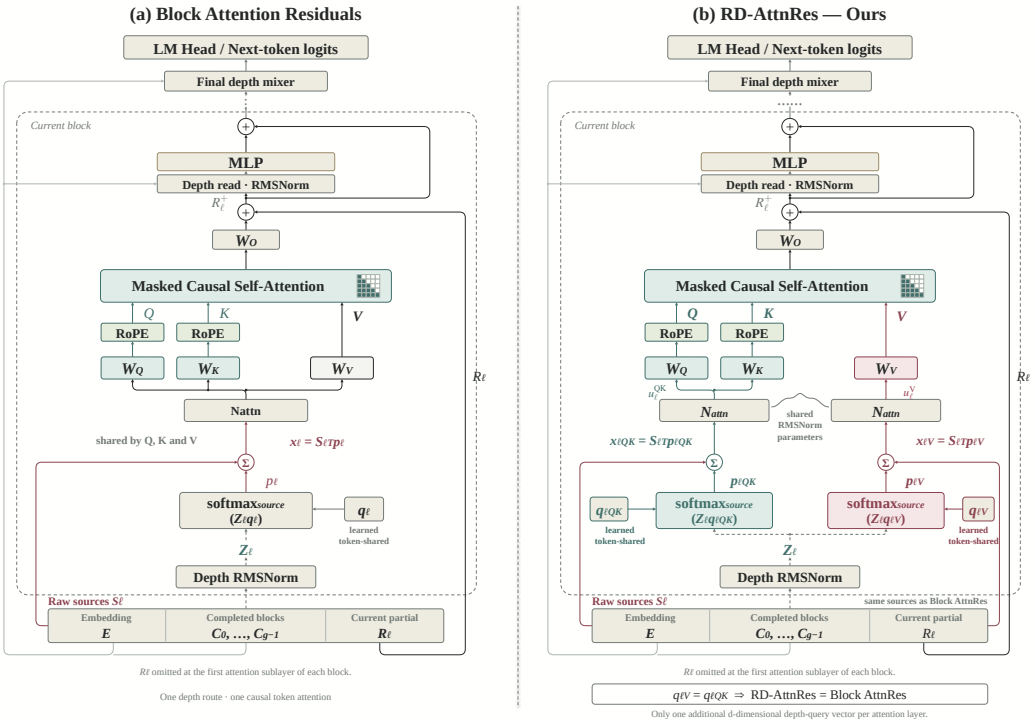

Figure 1: Block AttnRes and RD-AttnRes. (a) The parent uses one source distribution for Q, K, and V. (b) RD retains the same sources but gives Q/K and V separate routes. Both execute one masked causal self-attention; the MLP route, block write-back, and final mixer are unchanged. The schematic omits batch–token.

At batch–token position $c = (b, t)$, stack the m_ℓ sources as $S_{\ell,c} \in \mathbb{R}^{m_\ell \times d}$, where $d = d_{\text{model}}$, and apply depth RMSNorm row-wise (Zhang & Sennrich, 2019):

$$Z_{\ell,c} = N_\ell^{\text{depth}}(S_{\ell,c}). \tag{2}$$

A learned depth query $q_\ell \in \mathbb{R}^d$, shared across batch and token positions, then gives

$$\begin{aligned} p_{\ell,c} &= \text{softmax}_{\text{source}}(Z_{\ell,c} q_\ell)\,, \\ x_{\ell,c} &= S_{\ell,c}^\top p_{\ell,c}. \end{aligned} \tag{3}$$

The softmax runs across sources and uses no $1/\sqrt{d}$ factor. The normalized matrix $Z_{\ell,c}$ supplies the scores, while $S_{\ell,c}$ supplies the raw vectors being mixed. Although q_ℓ is token-shared, $p_{\ell,c}$ varies with the source content in $Z_{\ell,c}$. The equations describe the single-head depth-routing operator used by the main Block AttnRes and RD models in our experiments. The route query spans the full feature width d, and the source softmax is applied once over the source axis. A feature-partitioned multi-head variant is considered separately as the Block-MHAR baseline. After attention-input RMSNorm, the same mixed state generates all three attention representations:

$$\begin{aligned} u_{\ell,c} &= N_\ell^{\text{attn}}(x_{\ell,c}), \\ Q_{\ell,c} &= \text{RoPE}\Big(W_\ell^Q u_{\ell,c}\Big)\,, \qquad & K_{\ell,c} &= \text{RoPE}\big(W_\ell^K u_{\ell,c}\big)\,, \\ V_{\ell,c} &= W_\ell^V u_{\ell,c}. \end{aligned} \tag{4}$$

Lower-case q_ℓ is the depth query; uppercase Q, K, and V are token-attention representations, with rotary position embeddings applied to Q/K (Su et al., 2021). The source softmax acts over depth, while the masked causal softmax acts over positions $\tau \le t$. Let A_ℓ and M_ℓ denote the attention and MLP writes. After W_ℓ^O, the attention write updates the current partial state as

$$R_\ell^+ = R_\ell + A_\ell, \qquad R_{\ell+1} = R_\ell^+ + M_\ell, \tag{5}$$

where the MLP obtains its input through the parent's single-route read over the sources updated by A_ℓ. At the block boundary, the completed partial is stored as C_g and the next block starts without an empty partial source. For G blocks, the final mixer reads $E, C_0, \ldots, C_{G-1}$ before the final RMSNorm and language-model head.

3.2 Role-Decoupled Attention Read

Block AttnRes uses one depth distribution for the representations that form attention scores and for the values aggregated by those scores. RD-AttnRes assigns this formerly shared read to q_ℓ^{QK} and introduces an independent q_ℓ^{V}. Both queries score $Z_{\ell,c}$ and mix $S_{\ell,c}$. For $\rho \in \{\mathrm{QK}, \mathrm{V}\}$,

$$\begin{aligned} p_{\ell,c}^{\rho} &= \mathrm{softmax}_{\mathrm{source}}(Z_{\ell,c} q_\ell^{\rho}), \\ x_{\ell,c}^{\rho} &= S_{\ell,c}^{\top} p_{\ell,c}^{\rho}. \end{aligned} \tag{6}$$

The two mixed states are separately processed by the same parameterized attention RMSNorm:

$$\begin{aligned} u_{\ell,c}^{\rho} &= N_\ell^{\mathrm{attn}}\left(x_{\ell,c}^{\rho}\right), \qquad \rho \in \{\mathrm{QK}, \mathrm{V}\}, \\ Q_{\ell,c} &= \mathrm{RoPE}\left(W_\ell^{Q} u_{\ell,c}^{\mathrm{QK}}\right), \\ K_{\ell,c} &= \mathrm{RoPE}\left(W_\ell^{K} u_{\ell,c}^{\mathrm{QK}}\right), \\ V_{\ell,c} &= W_\ell^{V} u_{\ell,c}^{\mathrm{V}}. \end{aligned} \tag{7}$$

The two calls share the RMSNorm parameters, but their inputs and tokenwise normalization factors generally differ. The Q/K branch forms the representations used in token-matching scores; the V branch supplies the content aggregated by those scores. The resulting Q, K, and V enter the masked causal self-attention inherited from Block AttnRes. Its output projection, MLP depth read, block write-back, final mixer, final RMSNorm, and language-model head are also inherited.

Tying the route queries gives

$$q_\ell^{\mathrm{V}} = q_\ell^{\mathrm{QK}} \quad \Longrightarrow \quad p_{\ell,c}^{\mathrm{V}} = p_{\ell,c}^{\mathrm{QK}}, \qquad x_{\ell,c}^{\mathrm{V}} = x_{\ell,c}^{\mathrm{QK}}. \tag{8}$$

Under this constraint, the attention read reduces to Block AttnRes. For L attention layers of width d, the additional route queries contribute Ld parameters. Relative to Block AttnRes, RD additionally evaluates one V-route depth-score vector, one source softmax, and one raw-source mixture per layer; the resulting V-route state receives one additional invocation of N_ℓ^{attn} using the existing RMSNorm parameters. The token-attention term remains unchanged.

The L-scale Block/RD comparison uses one full-width depth-routing head per attention-side route. Block AttnRes uses one shared route for Q, K, and V, whereas RD-AttnRes uses separate single-head routes for Q/K and V over the same source set. Figure 1 depicts this role split.

3.3 Why QK|V and Local Route Freedom

Within one token-attention head, perturbing Q/K directly changes the token weights, while perturbing V directly changes the content aggregated under those weights; Appendix D.1 gives the exact differential. QK|V is the only two-route partition aligned with this split. Section 5.4 tests it against the equal-cost alternatives.

Tying the two queries embeds Block AttnRes in RD-AttnRes. Fix one read and batch–token position, let $S, Z \in \mathbb{R}^{m \times d}$ contain its raw sources and normalized depth keys, and define $p(q) = \mathrm{softmax}(Zq)$, $p_i = p(q_i)$, and $\Pi_m = I_m - \mathbf{1}\mathbf{1}^{\top}/m$. Softmax shift invariance gives

$$p_1 = p_2 \iff \Pi_m Z(q_1 - q_2) = 0. \tag{9}$$

If $\mathrm{rank}(\Pi_m Z) > 0$, the realizable shared-route pairs form a strict subset of the decoupled pairs. A realized pair produces distinct mixed states exactly when $S^{\top}(p_1 - p_2) \neq 0$.

The preceding statements are pointwise. For the complete causal map, fix the sources, collect routed inputs over the batch and sequence into $X(q)$, and set $J = \partial \,\mathrm{vec}\, X(q)/\partial q$. Let $\mathcal{F}(X^{\mathrm{QK}}, X^{\mathrm{V}})$

include RMSNorm, projections, RoPE, masked causal attention, and W^O. At (X, X), define $A = D_1\mathcal{F}(X, X)$ and $B = D_2\mathcal{F}(X, X)$. The shared and decoupled first-order output perturbations lie in

$$\mathrm{Im}((A + B)J) \quad \text{and} \quad \mathrm{Im}([AJ\ BJ]), \tag{10}$$

respectively. RD-AttnRes exposes an additional local output direction exactly when

$$\mathrm{rank}\,[AJ \quad BJ] > \mathrm{rank}((A + B)J). \tag{11}$$

This is a local statement at fixed sources and parameters; it does not assert an optimization or convergence result.

Let $a = \nabla_{q^{\mathrm{QK}}}\mathcal{L}$ and $v = \nabla_{q^{\mathrm{V}}}\mathcal{L}$ be the branchwise route gradients in the decoupled parameterization, evaluated at a tied point. With

$$\mu = \frac{q^{\mathrm{QK}} + q^{\mathrm{V}}}{2}, \qquad \delta = q^{\mathrm{V}} - q^{\mathrm{QK}}, \tag{12}$$

the corresponding derivatives are

$$\nabla_\mu\mathcal{L} = a + v, \qquad \nabla_\delta\mathcal{L} = \frac{v - a}{2}. \tag{13}$$

The tied Block-AttnRes query receives $a + v$; RD-AttnRes additionally exposes the role-difference derivative $(v - a)/2$.

4 Experimental Setup

Comparison structure. The replicated study pairs RD-AttnRes with its direct Block AttnRes parent at S and M scales. Within each pair, the attention-side route factorization is the only model-definition change. A single-seed L-scale pair tests the same comparison at approximately 1.5B parameters and 10B training tokens. A common-protocol M-scale study compares RD-AttnRes with Standard PreNorm, Block-MHAR, and Projected LR-Block. Tied-route controls and alternative role partitions isolate the architectural change, while route diagnostics and checkpoint interventions test whether trained models use the additional freedom. Separate C4 training, frozen-checkpoint cross-corpus evaluation, zero-shot tasks, and matched timing measurements assess distributional scope and computational cost.

Data and tokenizer. The main runs use the `sample/10BT` subset of FineWeb-Edu (Penedo et al., 2024), which contains 9,672,101 documents in 14 source shards. A fixed salted BLAKE2b-64 hash assigns complete documents to 99%/0.5%/0.5% training, validation, and test splits. We train a 32,768-token byte-level BPE tokenizer (Sennrich et al., 2016) on a deterministic 1/20 sample of the training documents. The packed splits contain 9,892,016,691, 50,328,166, and 49,144,462 tokens, respectively.

S- and M-scale models. All models are decoder-only causal Transformers with four layers per residual-history block and a context length of 1,024. The S configuration has 16 layers, 4 blocks, $d_{\mathrm{model}} = 576$, $d_{\mathrm{ff}} = 2880$, and 36 attention heads; Block and RD contain 119,791,296 and 119,800,512 parameters. The M configuration has 24 layers, 6 blocks, $d_{\mathrm{model}} = 832$, $d_{\mathrm{ff}} = 4160$, and 52 heads; the corresponding counts are 343,039,424 and 343,059,392. Each main run consumes 2,000,027,648 tokens in 61,036 updates, with 32,768 tokens per update. We use AdamW (Loshchilov & Hutter, 2019) with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, BF16 autocast, and FP32 cross-entropy. Weight decay is 0.1 for matrix and embedding weights and zero for bias, normalization, and routing parameters. After 1,221 linear-warmup updates, the learning rate follows cosine decay from 3×10^{-4} to 3×10^{-5} for S and from 2×10^{-4} to 2×10^{-5} for M.

S- and M-scale pairing and evaluation. The main Block–RD study uses seeds 123, 2025, 2026, 3407, and 7777. Within a pair, the models share the initialization of every common parameter, the initial QK and V routes implement the same mixture, and both runs receive the same training sequences in the same order. The optimizer, schedule, token budget, and validation prefix are also matched. Endpoint NLL is computed over a fixed 10,485,760-token validation prefix. We report all paired endpoints, mean within-seed perplexity reduction, and a two-sided 95% paired t-interval with

seed as the unit of replication. These intervals apply only to the replicated S/M experiments. The residual-routing comparison uses same-parent variants and the common M-scale seeds 123, 2025, and 2026. Capacity and role-partition controls use matched S-scale 0.5B-token runs. Route diagnostics and post-hoc interventions use three fixed RD checkpoints.

For distributional scope, we train three paired S-scale models per method on C4 (Raffel et al., 2020) for 1B tokens. We also evaluate four paired M-scale FineWeb-Edu checkpoints on WikiText-103 (Merity et al., 2017), PG19 (Rae et al., 2020), the near-deduplicated CodeParrot validation corpus (CodeParrot, 2022), and C4 without further training, followed by zero-shot HellaSwag, PIQA, ARC-Easy, and LAMBADA evaluation (Zellers et al., 2019; Bisk et al., 2020; Clark et al., 2018; Paperno et al., 2016).

L-scale configuration. The L configuration has 24 layers in six four-layer blocks, width 1,776, MLP width 8,880, and 111 token-attention heads of dimension 16. For the Block/RD comparison, depth routing is single-headed over the full model width: Block uses one shared QKV route, while RD uses separate QK and V routes over the same block-level source setAt this scale, we compare the single-head Block AttnRes parent with its role-decoupled RD counterpart. Block and RD contain 1,496,761,296 and 1,496,803,920 parameters. Both runs use seed 9013 and the same data, tokenizer, global batch, token budget, and learning-rate protocol. Block and RD both runs end at step 305,176 after consuming 10,000,007,168 tokens. The learning rate warms for 9,155 updates to 1.5×10^{-4} and then decays toward 1.5×10^{-5} over a 457,764-update horizon, which extends beyond the run. Quick evaluations occur every 7,630 steps on a fixed 1,048,576-token validation prefix. The 10B endpoint uses a 10,485,760-token validation protocol and is compared only within that protocol. Task construction follows a fixed Language Model Evaluation Harness snapshot (Gao et al., 2024); model scoring uses the project-specific evaluator recorded by hash. Downstream evaluation is zero-shot with a common evaluator, tokenizer, and task splits. Primary metrics are normalized accuracy for HellaSwag, PIQA, and ARC-Easy, and accuracy for LAMBADA. Appendix C gives the complete protocol.

5 Results

5.1 Direct Comparison with the Block Parent

Tying the route queries matches the Block parent on the FP32 recovery check: both constructions report a loss of 4.7543015, with a maximum absolute logit difference of 2.67×10^{-5} and no tokenwise argmax mismatches. At the network's first attention read, only one source is available, so both queries produce the same constant route and the measured QK/V difference is zero.

When the two routes are trained independently, every matched endpoint favors RD-AttnRes (Table 1). After 2B FineWeb-Edu tokens, the mean within-seed perplexity reduction is 2.97% at S and 2.43% at M. Individual reductions span 2.66–3.45% at S and 1.19–4.21% at M; both paired NLL intervals lie above zero under the Block-minus-RD convention.

Table 1: Paired FineWeb-Edu endpoints after 2B training tokens. We define $\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{Block}} - \mathrm{NLL}_{\mathrm{RD}}$, so positive values favor RD-AttnRes. The 95% paired t-intervals, with seed as the unit of replication, are $[0.026348, 0.033887]$ at S and $[0.009812, 0.039570]$ at M.

	S (approximately 120M)				M (approximately 343M)			
Seed	Block NLL	RD NLL	ΔNLL	PPL red.	Block NLL	RD NLL	ΔNLL	PPL red.
123	3.167001	3.137957	0.029044	2.86%	3.024623	3.006487	0.018136	1.80%
2025	3.157996	3.131043	0.026953	2.66%	3.036566	3.007358	0.029209	2.88%
2026	3.174608	3.145236	0.029372	2.89%	3.034505	2.991464	0.043042	4.21%
3407	3.178064	3.147977	0.030086	2.96%	3.030800	3.009731	0.021070	2.08%
7777	3.184249	3.149117	0.035132	3.45%	3.012970	3.000973	0.011996	1.19%
Mean	3.172384	3.142266	0.030118	2.97%	3.027893	3.003202	0.024691	2.43%

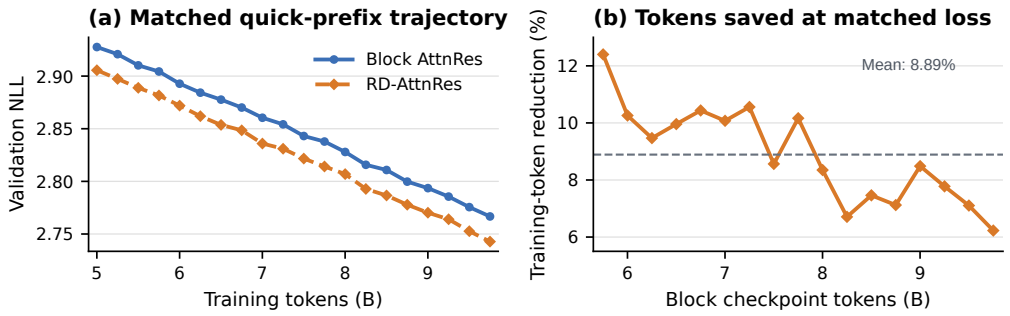

Figure 2: Single-seed L-scale comparison on the common 1,048,576-token quick-validation prefix; Block denotes the single-head Block AttnRes parent. (a) Observed 5.00–9.75B trajectories; connecting segments guide the eye. (b) Percentage reduction in training tokens needed by RD to reach each estimable Block validation NLL, obtained by piecewise-linear interpolation of the observed RD trajectory. The 10B long-prefix endpoint is excluded, and no extrapolation is used.

5.2 Large-Scale 10B-Token Comparison

At 10,000,007,168 training tokens, RD-AttnRes reaches 2.656103 NLL on the long fixed validation prefix, versus 2.678519 for Block AttnRes; the corresponding perplexities are 14.241 and 14.564, a 2.217% reduction. Figure 2(a) shows that RD is also lower at all 20 common quick-prefix checkpoints from 5.00B through 9.75B. These are serial observations from one seed pair, not independent replications. The 10B endpoint uses ten times as many validation tokens and is compared only within that protocol.

For a Block checkpoint (T_i, b_i), we linearly interpolate the RD quick trajectory to obtain the token count $\widehat{T}_{\mathrm{RD}}(b_i)$ at the same NLL. We define the equal-loss token reduction as $100\big(T_i - \widehat{T}_{\mathrm{RD}}(b_i)\big)/T_i$. Across the 17 targets supported without extrapolation, RD reaches the same validation NLL with 8.89% fewer training tokens on average. Combining this reduction with the recorded run-average throughput gives an estimated 7.51% shorter time to the same NLL for this pair.

At the 10B endpoint, RD improves the equal-weight macro average of the four primary zero-shot metrics by 1.77 percentage points. Appendix C gives the complete task results, checkpoint record, and evaluation audit.

5.3 Comparison with Other Residual-Routing Designs

At M scale, RD is also the strongest same-parent routing design tested under the common protocol (Table 2). Standard PreNorm supplies the no-routing reference; Block-MHAR partitions features across depth routes, whereas Projected LR-Block changes the representation used to score depth sources.

Table 2: Common-protocol M-scale comparison after 2B FineWeb-Edu tokens. All methods use the same seeds, data order, optimization schedule, and fixed 10,485,760-token validation prefix.

Method	Parameters	Seed 123	Seed 2025	Seed 2026	Mean NLL ↓
Standard PreNorm	342.958M	3.076376	3.079593	3.077487	3.077819
Block AttnRes	343.039M	3.024623	3.036566	3.034505	3.031898
Block-MHAR ($H = 8$)	343.039M	3.026530	3.031460	3.025904	3.027965
Projected LR-Block ($r = 32$)	346.822M	3.016373	3.018024	3.014562	3.016320
RD-AttnRes	343.059M	**3.006487**	**3.007358**	**2.991464**	**3.001769**

On the three shared seeds, Block-MHAR, Projected LR-Block, and RD-AttnRes lower the Block mean NLL by 0.003933, 0.015578, and 0.030129, respectively. RD-AttnRes is lower than both alternatives in each seed and has a mean NLL margin of 0.014551 over Projected LR-Block. Full

AttnRes, Direct Delta, Sliced LR-Block, and Projected LR+RD have different scale, token, seed, or endpoint coverage and therefore appear in the appendix rather than this ranking.

5.4 Attributing the Gain to Role-Decoupled Routing

Execution and redundant-parameter controls. At S/seed-123/0.5B, Block AttnRes reaches 3.541607 NLL. The tied-QKV, dual-route average, and frozen-V controls reach 3.541253, 3.540951, and 3.539314, all within 0.0023 of the parent. Under the same budget, RD-AttnRes reaches 3.499386, a reduction of 0.042221. Tied-QKV shares the query while retaining the two-route execution path; dual-route average constructs two routed states and averages them before the QKV projections; frozen-V leaves the V query at zero, hence its source weights uniform, while QK adapts. None reproduces the RD reduction. At the symmetric initialization, the average control has no first-order antisymmetric route direction, so it matches duplicated computation and allocated queries rather than RD's role-assigned freedom.

Role partition. The S/seed-42/0.5B role-partition study changes which attention roles share a route. Shared-2Q averages two query vectors before the source softmax and therefore still produces one QKV route; its endpoint is nearly identical to Block. This differs from the dual-route average above, which first constructs two routed states and then averages those states. The three equal-cost two-route variants fix route count while changing only the grouping. In this run, QK|V is the best of these variants and captures 96.8% of the NLL reduction obtained by splitting all three roles (Table 3(a)).

5.5 The Learned Routes Diverge and Affect the Model

At the tied point, each of the three checkpoints has a nonzero gradient along the role-difference coordinate. Their global $\cos(a, v)$ values range from -0.447 to 0.226, showing that the two branch gradients are not simple copies.

After training, QK and V differ at every non-degenerate read site. Table 3(b) reports route JS divergences and mixed-state cosines. The first, single-source site returns JS$= 0$ and cosine$= 1$. The JS values compare route probabilities, while the state cosines confirm that the different distributions produce different attention inputs. Appendix E.4 specifies the reductions.

Table 3: Mechanism evidence for role-decoupled routing. Panel (a) compares role partitions at S/seed-42/0.5B against the Block baseline 3.518998, with $\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{Block}} - \mathrm{NLL}_{\mathrm{variant}}$; positive values favor the variant. Panel (b) reports route-separation probes defined in Appendix E.4. Panel (c) reports $\Delta_{\mathrm{int}} = \mathrm{NLL}_{\mathrm{intervened}} - \mathrm{NLL}_{\mathrm{trained}}$ on fixed checkpoints; $\lambda = 0$ ties the V query to QK, while $\lambda = 1$ is the trained model.

(a) Role partition

Variant	Validation NLL	ΔNLL
Shared-2Q	3.518879	0.000119
QV|K	3.500998	0.018000
KV|Q	3.508192	0.010806
QK|V (RD)	**3.458975**	**0.060023**
Q|K|V	3.456997	0.062001

(b) Route-separation probes

Checkpoint	$\cos(a, v)$	JS (nats)	State cos.
S/2B, seed 7777	0.226	0.191	0.629
M/2B, seed 7777	−0.447	0.205	0.634
S/0.5B, seed 42	−0.031	0.224	0.595

(c) Functional interventions: Δ_{int} relative to the trained checkpoint

Checkpoint	$\lambda = 0$	0.25	0.50	0.75	1.00	1.25	V shift	QK↔V
S/2B, seed 7777	1.904	1.483	0.641	0.103	0.000	0.048	0.139	3.419
M/2B, seed 7777	1.384	0.640	0.169	0.031	0.000	0.135	0.121	3.224
S/0.5B, seed 42	1.998	1.362	0.466	0.066	0.000	0.035	0.121	3.114

Interventions on the same checkpoints connect this separation to the trained function (Table 3(c)). The query sweep sets $q_\ell^{\mathrm{V}}(\lambda) = q_\ell^{\mathrm{QK}} + \lambda(q_\ell^{\mathrm{V}} - q_\ell^{\mathrm{QK}})$ simultaneously at every layer. Moving the V query toward the QK query raises NLL at every sampled point and adds 1.384–1.998 NLL at complete tying. Extending the learned difference to $\lambda = 1.25$ adds 0.035–0.135 NLL. Shifting V-route weights across batch examples while preserving layer, sequence position, and source count

adds 0.121–0.139; swapping the QK and V routes adds 3.114–3.419. The checkpoints depend on the magnitude, token-conditioned assignment, and role orientation of the learned route split. These interventions measure the cost of disrupting a co-adapted RD computation; the paired training runs estimate the gap to a separately optimized Block model.

At the L endpoint, QK/V Jensen–Shannon divergence is nonzero at all 23 non-degenerate read sites; their equal-weight mean and maximum are 0.171838 and 0.617517 nats. The endpoint therefore does not collapse to the shared-route subspace (Appendix C.5).

5.6 Transfer across Training and Evaluation Distributions

Under C4 training, RD is lower in all three S-scale pairs after 1B tokens, with a mean PPL reduction of 3.33% and a 95% paired t-interval of $[-0.003901, 0.071806]$ for Block-minus-RD NLL. Without further training, RD also lowers PPL in all 16 M-scale seed–corpus pairs on WikiText-103, PG19, CodeParrot, and C4. Zero-shot transfer is task-dependent: HellaSwag, PIQA, and LAMBADA improve, while ARC-Easy normalized accuracy is essentially unchanged.

5.7 Computational Cost

RD-AttnRes adds 9,216 parameters at S, 19,968 at M, and 42,624 at L, never more than 0.00769% of the corresponding Block model. Across a L scale on an RTX PRO 6000, throughput changes from 9,253.743 to 9,115.990 tokens/s (1.51% more time per token), implying 300.179 and 304.715 hours over the common 10B-token budget. Despite the higher per-token cost, the interpolated token reduction and run-average throughput imply an estimated 7.51% shorter time to matched NLL. L-scale memory is reported in Appendix C.4. By contrast, the downstream scoring suite took 43.0% longer for RD (Appendix C.3).

6 Discussion and Limitations

Across the ten paired S/M runs, independently adaptive QK/V reads consistently lower language-modeling loss. The S-scale controls show that duplicated route computation, symmetric state averaging, and a uniform V route do not reproduce the reduction. QK|V is also the best equal-cost partition and captures 96.8% of the gain from splitting all three roles. This matches self-attention: Q/K determine token weights, while V supplies the content carried under them. The consequential change is role assignment, not route count or a second token-attention operation. Untying also changes the optimizer parameterization, so the evidence identifies the useful trained design without assigning the gain to one optimization mechanism.

Trained models use the added direction. Each examined checkpoint has a nonzero role-difference gradient at the tied point; after training, route distributions and mixed states differ at every non-degenerate site. Tying, shifting, or swapping the routes raises NLL. These interventions test functional dependence in co-adapted models, while the paired Block–RD runs quantify the training gain.

At L scale, RD reaches 17 comparable Block losses with 8.89% fewer tokens on average. Although each token takes 1.51% longer, throughput still implies 7.51% less time to the same NLL. Its 20/20 quick-checkpoint lead persists across 5–9.75B, although one pair cannot estimate L-scale seed variation. Language-modeling transfer favors RD in all 16 frozen-checkpoint cross-corpus pairs and all three C4-training pairs; zero-shot changes are smaller and task-dependent. Tests span 120M–1.5B parameters, English-heavy web data, and 1,024-token contexts; other contexts and distributions remain open.

7 Conclusion

RD-AttnRes gives Q/K and V separate depth reads while preserving Block's source stack and token-attention operation. It lowers loss across 120M–1.5B models. At 1.5B, matched loss requires 8.89% fewer tokens and an estimated 7.51% less time. Controls and interventions identify adaptive role separation as consequential: relaxing the shared QKV read improves the tested Block models.

AI USE STATEMENT

The authors did not use generative AI tools in the research or preparation of this manuscript.

ETHICS STATEMENT

This work studies language-model architecture on FineWeb-Edu and C4, web corpora that may retain social biases, personal information, or copyrighted text despite dataset curation. We do not release generated text or make deployment claims. Architectural efficiency and quality improvements can lower experimentation cost, but may also broaden access to capabilities that inherit the risks of web-scale training data. Any downstream deployment requires separate data-governance, safety, and bias evaluation.

REPRODUCIBILITY STATEMENT

Appendix E specifies the dataset split, tokenizer, model shapes, optimization, seeds, evaluation budget, parameters, hardware, and software environment. Appendix A reports every replicated S/M endpoint, Appendix F states the estimands and intervals, and Appendix G records the S/M compute environment and matched runtime measurements. Appendix C gives the L-scale configuration, complete shared quick-prefix record, interpolation rule, 10B endpoint, downstream audit, efficiency record, and layerwise route divergences.An anonymized reproduction package is included in the supplementary material, containing the model implementations, training and data-processing code, evaluation scripts, run configurations, tokenizer artifacts, software-environment records, and scripts for auditing the supplied summary records and exporting the reported tables and data figures. Retained checkpoints and training logs are not included in the submission and will be released separately.

REFERENCES

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

CodeParrot. CodeParrot dataset after near deduplication (validation). Hugging Face dataset repository, 2022. URL https://huggingface.co/datasets/codeparrot/codeparrot-valid-near-deduplication. Accessed 2026-09-05.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 2024. URL https://zenodo.org/records/12608602.

Mike Heddes, Adel Javanmard, Kyriakos Axiotis, Gang Fu, Mohammadhossein Bateni, and Vahab Mirrokni. DeepCrossAttention: Supercharging transformer residual connections. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22881–22903. PMLR, 2025. URL https://proceedings.mlr.press/v267/heddes25a.html.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. doi: 10.48550/arXiv.2607.24653. URL https://arxiv.org/abs/2607.24653.

Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, et al. Attention residuals. *arXiv preprint arXiv:2603.15031*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Cheng Luo and Zefan Cai. Delta attention residuals: Source representation and per-stream routing for cross-layer information flow. Open Attention Residuals repository, 2026. URL https://github.com/wdlctc/open-attention-residuals/tree/16a28d4. Fixed commit 16a28d4; accessed 2026-09-05.

Cheng Luo, Zefan Cai, and Junjie Hu. Delta attention residuals. *arXiv preprint arXiv:2605.18855*, 2026a.

Cheng Luo, Zefan Cai, and Junjie Hu. Multi-head attention residuals. *arXiv preprint arXiv:2607.27230*, 2026b.

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

Matteo Pagliardini, Amirkeivan Mohtashami, François Fleuret, and Martin Jaggi. DenseFormer: Enhancing information flow in transformers via depth weighted averaging. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136479–136508, 2024. doi: 10.52202/079017-4336.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144/.

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.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Jonathan Su. Low-rank attention residuals. *arXiv preprint arXiv:2607.09694*, 2026.

Da Xiao, Qingye Meng, Shengping Li, and Xingyuan Yuan. MUDDFormer: Breaking residual bottlenecks in transformers via multiway dynamic dense connections. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68440–68458. PMLR, 2025. URL https://proceedings.mlr.press/v267/xiao25d.html.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Boyi Zeng, Yiqin Hao, Zitong Wang, Shixiang Song, He Li, Feichen Song, Yifan Liu, Ziwei He, Xinbing Wang, and Zhouhan Lin. Depth-attention: Cross-layer value mixing for language models. *arXiv preprint arXiv:2606.05014*, 2026. doi: 10.48550/arXiv.2606.05014. URL https://arxiv.org/abs/2606.05014.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 2019.

Zhanchao Zhou, Tianyi Wu, Zhiyun Jiang, Fares Obeid, and Zhenzhong Lan. Value residual learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28341–28356. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1375. URL https://aclanthology.org/2025.acl-long.1375/.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f1e8ff97057e0c3f6bcb7018c78d4df1-Abstract-Conference.html.

A Complete Paired Results

Table 4: All replicated S/M endpoints after 2,000,027,648 training tokens. ΔNLL is Block minus RD, and relative PPL reduction is $(\mathrm{PPL}_{\mathrm{Block}} - \mathrm{PPL}_{\mathrm{RD}})/\mathrm{PPL}_{\mathrm{Block}}$; positive values favor RD.

Scale	Seed	Block NLL	RD NLL	Δ NLL	Block/RD PPL	PPL red.
S	123	3.16700	3.13796	0.02904	23.736/23.057	2.86%
S	2025	3.15800	3.13104	0.02695	23.523/22.898	2.66%
S	2026	3.17461	3.14524	0.02937	23.917/23.225	2.89%
S	3407	3.17806	3.14798	0.03009	24.000/23.289	2.96%
S	7777	3.18425	3.14912	0.03513	24.149/23.315	3.45%
M	123	3.02462	3.00649	0.01814	20.586/20.216	1.80%
M	2025	3.03657	3.00736	0.02921	20.834/20.234	2.88%
M	2026	3.03451	2.99146	0.04304	20.791/19.915	4.21%
M	3407	3.03080	3.00973	0.02107	20.714/20.282	2.08%
M	7777	3.01297	3.00097	0.01200	20.348/20.105	1.19%

C4 training pairs. The separate S-scale C4 runs use a 1B-token budget. RD is lower in all three pairs: Block/RD NLL is 3.475379/3.454070 for seed 123, 3.465589/3.414718 for seed 2025, and 3.468689/3.439013 for seed 2026. Their mean within-seed PPL reduction is 3.33%; the 95% paired t-interval for Block-minus-RD NLL is $[-0.003901, 0.071806]$.

B Residual-Routing Variants outside the Common Protocol

Table 5: Additional residual-routing results with heterogeneous scale, token, seed, or endpoint coverage. A dagger marks an archived aggregate for which individual endpoints are unavailable.

Method	Scale / tokens	Seeds	Endpoint NLL	Design question
Projected LR+RD ($r = 32$)	M / 2B	3†	mean 3.00544†	Composition
Sliced LR-Block	M / 2B	123	3.017827	Fixed low-rank key
Direct Delta Block	M / 2B	123/2025	mean 3.053003	Source/write rule
Full AttnRes	S / 1B	123	3.310178	Source granularity

These runs answer separate design questions and are not mixed into the common-protocol ranking. The archived *Direct Delta Block* run predates and is distinct from the public Delta-V implementation discussed in Related Work; it does not provide an RD–Delta-V comparison. The Projected LR+RD archive retains its three-seed mean but not its individual endpoints.

C LARGE-SCALE 10B-TOKEN EXPERIMENT

C.1 CONFIGURATION AND PROTOCOL

Both methods use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$, global-norm clipping at 1.0, BF16 autocast, and the same FineWeb-Edu data, tokenizer, seed, global batch, token budget, and learning-rate protocol. Weight decay is 0.1 for matrix parameters and zero for bias, low-dimensional, and routing parameters. The packed training stream cycles deterministically after one pass. After warmup, cosine decay is parameterized over a 457,764-update horizon; the 10B runs stop at update 305,176, before reaching the specified minimum learning rate.

Quick evaluations use a fixed 1,048,576-token validation prefix. The 10B endpoint uses a 10,485,760-token validation protocol and is compared only within that protocol. The archived Block `eval_metrics.jsonl` is unavailable before 5B. A frozen report retains its 0.25B and 0.5B endpoints, but the 0.75B–4.75B values remain missing and are neither interpolated nor estimated.

Table 6: L-scale training configuration. Shared entries apply to both methods.

Item	Block AttnRes	RD-AttnRes
Parameters	1,496,761,296	1,496,803,920
Layers / residual blocks	24 / 6, with 4 layers per block	
Model / MLP width	1,776 / 8,880	
Attention heads / head dimension	111 / 16	
Depth-routing heads per route	1	
Attention-side routes	1 shared QKV route	2 routes (QK and V)
Context length / vocabulary	1,024 / 32,768	
Microbatch / accumulation	4 / 8	
Global batch / tokens per update	32 sequences / 32,768 tokens	
Updates / training tokens	305,176 / 10,000,007,168	
Seed	9013	
Peak / minimum learning rate	1.5×10^{-4} / 1.5×10^{-5}	
Warmup / scheduler horizon	9,155 / 457,764 updates	
Quick-evaluation cadence / size	7,630 updates / 1,048,576 tokens	
Archived execution	resumed at 5B; 0–10B total	from scratch; 0–10B total

C.2 QUICK-PREFIX TRAJECTORY AND 10B ENDPOINT

Across the 20 quick-prefix checkpoints, the arithmetic mean of the checkpointwise Block-minus-RD NLL differences is 0.022643, with a range of 0.020813–0.024502. The corresponding mean relative perplexity reduction is 2.239%, with a range of 2.060–2.420%. These checkpoints are correlated observations from one training pair, not units of replication.

For the equal-loss token saving rate in Figure 2, take a Block checkpoint (T_i, b_i) and locate adjacent RD quick-prefix observations (t_j, r_j) and (t_{j+1}, r_{j+1}) satisfying $r_j \geq b_i \geq r_{j+1}$. We use

$$\widehat{T}_{\mathrm{RD}}(b_i) = t_j + (t_{j+1} - t_j)\frac{r_j - b_i}{r_j - r_{j+1}}, \qquad S_i = 100\frac{T_i - \widehat{T}_{\mathrm{RD}}(b_i)}{T_i}. \tag{14}$$

The target lies inside the observed RD range at 17 Block checkpoints from 5.75B through 9.75B. The checkpointwise mean of S_i is 8.89%: RD reaches the same validation NLL with 8.89% fewer training tokens on average. The corresponding mean absolute lead is 671,630,219 tokens. Combining each interpolated token requirement with the recorded run-average throughput gives a mean estimated time-to-NLL reduction of 7.51%. The 5.00B, 5.25B, and 5.50B Block losses are already above the first RD value at 5.00B; estimating their crossings would require backward extrapolation, so they are omitted. The calculation also excludes the 10B endpoint protocol.

As a checkpoint-only sensitivity check, we replace the interpolated crossing by the first recorded RD checkpoint whose NLL reaches or improves upon each of the same 17 Block targets. The mean token reduction is 7.18%, requiring no interpolation between observations; combined with run-average throughput, it gives an estimated 5.77% reduction in time to the recorded target. This statistic locates attainment at checkpoint resolution rather than estimating the first crossing. Exact values for both calculations are provided in `data/l_scale_token_lead.csv`.

Table 7: Quarter-billion quick-prefix record from 5.0B through 9.75B, followed by the 10B long-prefix endpoint. The final row, marked †, should not be read as a continuation of the same validation series. PPL values are recomputed as $\exp(\mathrm{NLL})$; the final column is the relative RD PPL reduction.

Training tokens	Step	Block NLL	RD NLL	Block/RD PPL	PPL red.
5,000,396,800	152,600	2.927588	2.905573	18.683/18.276	2.177%
5,250,416,640	160,230	2.920813	2.897320	18.556/18.126	2.322%
5,500,436,480	167,860	2.910189	2.888849	18.360/17.973	2.111%
5,750,456,320	175,490	2.904354	2.881594	18.253/17.843	2.250%
6,000,476,160	183,120	2.892764	2.871951	18.043/17.671	2.060%
6,250,496,000	190,750	2.884257	2.861910	17.890/17.495	2.210%
6,500,515,840	198,380	2.877634	2.853650	17.772/17.351	2.370%
6,750,535,680	206,010	2.870122	2.848490	17.639/17.262	2.140%
7,000,555,520	213,640	2.860430	2.835928	17.469/17.046	2.420%
7,250,575,360	221,270	2.854163	2.831000	17.360/16.962	2.290%
7,500,595,200	228,900	2.843073	2.821647	17.168/16.805	2.120%
7,750,615,040	236,530	2.837824	2.814147	17.079/16.679	2.340%
8,000,634,880	244,160	2.827929	2.806917	16.910/16.559	2.079%
8,250,654,720	251,790	2.815750	2.792797	16.706/16.327	2.269%
8,500,674,560	259,420	2.810801	2.786613	16.623/16.226	2.390%
8,750,694,400	267,050	2.799752	2.777814	16.441/16.084	2.170%
9,000,714,240	274,680	2.793573	2.770231	16.339/15.962	2.307%
9,250,734,080	282,310	2.785535	2.764007	16.208/15.863	2.130%
9,500,753,920	289,940	2.775535	2.752676	16.047/15.685	2.260%
9,750,773,760	297,570	2.766676	2.742794	15.906/15.530	2.360%
10,000,007,168†	305,176	2.678519	2.656103	14.564/14.241	2.217%

C.3 Zero-shot evaluation audit

The downstream evaluator uses batch size 16, maximum sequence length 1,024, zero-shot prompting, BF16 autocast, FP32 log-softmax, and no sample-level `limit`. Every task passes the `all_boundaries_nonempty=true` check. Table 8 reports all metrics; the accuracy aggregates correspond to the integer document counts in Table 9.

Table 8: Complete L-scale zero-shot results. Accuracy is percentage ± item-sampling standard error. Accuracy differences are percentage points; the LAMBADA PPL difference is relative.

Task	Metric	Block	RD	RD − Block
HellaSwag	Accuracy	37.8909 ± 0.4841	39.1755 ± 0.4871	$+1.2846$
HellaSwag	Accuracy norm.	48.1777 ± 0.4986	49.6714 ± 0.4989	$+1.4937$
PIQA	Accuracy	70.2938 ± 1.0659	72.9053 ± 1.0367	$+2.6115$
PIQA	Accuracy norm.	69.4777 ± 1.0741	71.7628 ± 1.0500	$+2.2851$
ARC-Easy	Accuracy	61.6162 ± 0.9977	64.0993 ± 0.9841	$+2.4832$
ARC-Easy	Accuracy norm.	55.5556 ± 1.0194	57.4916 ± 1.0142	$+1.9360$
LAMBADA	Accuracy	35.9208 ± 0.6683	37.2793 ± 0.6736	$+1.3584$
LAMBADA	PPL	23.937792	23.075031	-3.6042%

LAMBADA scores 7,705 continuation tokens. Its task PPL is a document-level evaluator aggregate and is distinct from FineWeb-Edu validation PPL. The task files and measured runtimes are:

Exact SHA-256 identifiers for both checkpoints, the tokenizer, evaluator, runner, and task splits, together with the harness commit and per-task runtime metadata, are recorded in `data/l_scale_reproducibility.json`.

C.4 Training efficiency

Both archived runs use one NVIDIA RTX PRO 6000 Blackwell Server Edition, PyTorch 2.12.1+cu130, CUDA 13.0, and BF16 autocast. The timing and memory values characterize these runs and are not extrapolated across hardware.

Table 9: L-scale zero-shot audit counts. Each accuracy metric is the mean of one binary outcome per document.

Task	Metric	Documents	Requests	Block correct	RD correct
HellaSwag	Accuracy	10,042	40,168	3,805	3,934
HellaSwag	Accuracy norm.	10,042	40,168	4,838	4,988
PIQA	Accuracy	1,838	3,676	1,292	1,340
PIQA	Accuracy norm.	1,838	3,676	1,277	1,319
ARC-Easy	Accuracy	2,376	9,501	1,464	1,523
ARC-Easy	Accuracy norm.	2,376	9,501	1,320	1,366
LAMBADA	Accuracy	5,153	5,153	1,851	1,921

Table 10: L-scale downstream task inputs and measured evaluation times. Token maxima are measured after truncation. The suite total is an independent wall-clock record that includes runner overhead.

Task	Max input	Max cont.	Block (s)	RD (s)	RD/Block
HellaSwag	154	91	98.257	139.469	1.420
PIQA	258	250	6.363	9.617	1.511
ARC-Easy	158	32	15.236	22.944	1.506
LAMBADA	225	6	12.525	17.312	1.382
Suite wall-clock	—	—	132.394	189.351	1.430

Table 11: Archived L-scale training footprint. Duration is derived from the common token budget and recorded throughput, which is used as the timing authority. Memory units follow the archived summaries.

Metric	Block	RD	RD change
Throughput (tokens/s)	9,253.743	9,115.990	-1.489%
Implied duration (h)	300.179	304.715	$+1.511\%$
Peak allocated (GB)	55.609	57.235	$+1.626$ GB
Peak reserved (GB)	94.283	94.102	-0.181 GB

C.5 Layerwise route diagnostics

At the final L-scale step, layer 0 has only one source and therefore a constant route. QK/V Jensen–Shannon divergence is nonzero in each of the remaining 23 layers. Following Appendix E.4, the equal-weight mean over these non-degenerate layers is 0.171838 nats and the maximum is 0.617517 nats.

Table 12: L-scale route diagnostics at step 305,176. Layers are zero-indexed; each value micro-averages 16,384 contexts and uses natural logarithms.

Layer	Sources	QK/V JS (nats)	Layer	Sources	QK/V JS (nats)
0	1	0.000000	12	4	0.046144
1	2	0.617517	13	5	0.027424
2	2	0.351918	14	5	0.036338
3	2	0.021262	15	5	0.041931
4	2	0.381697	16	5	0.129662
5	3	0.089866	17	6	0.089363
6	3	0.033489	18	6	0.128683
7	3	0.006799	19	6	0.084185
8	3	0.232728	20	6	0.420487
9	4	0.101165	21	7	0.279749
10	4	0.072572	22	7	0.553594
11	4	0.079928	23	7	0.125772

D EXACT RECOVERY OF BLOCK ATTNRES

In real arithmetic, assume $q_\ell^{\mathrm{V}} = q_\ell^{\mathrm{QK}}$ for every layer and that both routes use the same source tensors, RMSNorm, and softmax. Their logits, route distributions p, and mixed states x are then equal elementwise. Starting from the common embedding, suppose all preceding writes agree between tied RD and Block. Their next source stacks agree, so Equations 6–7 produce the same attention write; the unchanged MLP read and write then agree as well. Induction over layers and blocks preserves every subsequent source, final mixture, and logit. The Tied control realizes this constraint with one shared route-query parameter and matches the parent within 0.000354 NLL at the 0.5B-token milestone.

D.1 DIRECT SCORE AND CONTENT EFFECTS

For one token-attention head, let $e_{t\tau} = Q_t^\top K_\tau/\sqrt{d_h}$, $\omega_{t\tau} = \mathrm{softmax}_{\tau\le t}(e_{t\tau})$, and $y_t = \sum_{\tau\le t}\omega_{t\tau}V_\tau$. On unmasked positions,

$$dy_t = \sum_{\tau\le t}\omega_{t\tau}\,dV_\tau + \sum_{\tau\le t}\omega_{t\tau}(V_\tau - y_t)\,de_{t\tau},$$
$$de_{t\tau} = \frac{(dQ_t)^\top K_\tau + Q_t^\top dK_\tau}{\sqrt{d_h}}. \tag{15}$$

The first term changes the content aggregated under fixed token weights; the second changes those weights through Q/K. The identity follows by differentiating the softmax and substituting $d\omega_{t\tau} = \omega_{t\tau}(de_{t\tau} - \sum_\nu \omega_{t\nu}de_{t\nu})$.

E FULL EXPERIMENTAL DETAILS

E.1 DATASET AND TOKENIZER

The source archive contains 14 FineWeb-Edu sample/10BT parquet shards totaling 28.518 GB and 9,672,101 documents. The deterministic document split contains 9,575,157 train documents (9,892,016,691 BPE tokens), 48,683 validation documents (50,328,166 tokens), and 48,261 test documents (49,144,462 tokens). The tokenizer is trained only on a deterministic one-in-20 sample of training documents: 478,729 documents and 2,274,481,521 UTF-8 bytes. It is byte-level BPE with vocabulary size 32,768 and special IDs PAD=0, UNK=1, BOS=2, EOS=3. Token streams append EOS per document and are stored as unsigned 16-bit integers. The test split was not consulted in method selection or reporting.

CodeParrot evaluation corpus. We evaluate the near-deduplicated CodeParrot validation corpus (CodeParrot, 2022). The repository has no named configuration (`name=None`) and exposes its native split as `train`; `validation` is the local split name assigned by our preparation pipeline. We pin revision `428766bc07a9e8699b7782e8557b7e07d32923f3`. Dataset preparation uses direct file downloads rather than `datasets-server`: it fetches three objects from this revision and packages them as the local split.

> Dataset: `codeparrot/codeparrot-valid-near-deduplication`
> Files:
> `data/train-00000-of-00002.parquet`
> `data/train-00001-of-00002.parquet`
> `data/train-00066-of-00070.parquet`

The pinned corpus contains 110,960 raw records. Evaluation reads the `content` field and applies a cap of 10,000,000 tokenizer tokens.

E.2 MODEL AND OPTIMIZATION HYPERPARAMETERS

Both models use RMSNorm $\epsilon = 10^{-6}$, RoPE base 10,000, zero dropout, and Gaussian initialization with standard deviation 0.02; attention projections are bias-free, and the input embedding and language-model head are tied. Route queries start at zero. The global batch is 32 sequences (32,768

Table 13: Model configurations. Parameter counts include tied token embeddings and output head.

Configuration	S	M
Layers / width	16 / 576	24 / 832
Attention heads / head dimension	36 / 16	52 / 16
MLP hidden width	2,880	4,160
Layers per residual block	4	4
Context length / vocabulary	1,024 / 32,768	1,024 / 32,768
Parent parameters	119,791,296	343,039,424
RD parameters	119,800,512	343,059,392
RD increase	9,216 (0.0077%)	19,968 (0.0058%)
Peak learning rate	3×10^{-4}	2×10^{-4}
Microbatch / accumulation	8 / 4	4 / 8

tokens per step). Warmup occupies 2% of 61,036 steps, followed by cosine decay to 10% of the peak learning rate; weight decay applies only to matrix parameters. Validation milestones are 250,019,840; 500,006,912; 1,000,013,824; and 2,000,027,648 training tokens, each evaluated on 10,485,760 validation tokens.

E.3 Controls

The *Tied* control shares one route query between QK and V. *Average* executes both route operators and averages their mixed states before the QKV projections. Writing its two queries as $q_1 = \mu - \delta/2$ and $q_2 = \mu + \delta/2$, the averaged state $\{x(q_1) + x(q_2)\}/2$ has zero first derivative with respect to δ at the symmetric point. With the recorded zero initialization, this control therefore allocates two queries and duplicates routing computation without exposing RD's role-assigned first-order difference direction at initialization. *Frozen-V* fixes the value query at zero while QK learns. All three use seed 123 and stop at 15,259 updates (500,006,912 tokens). The S RD endpoint in Section 5.4 is an independent ablation run (3.499386); its main-run milestone is 3.498424. Elsewhere, the M RD value is the main-run milestone because no separate completed M ablation record was archived.

E.4 Route-diagnostic protocol

Route probabilities are computed with an FP32 softmax along the source dimension. Distribution and state diagnostics use four batches of four 1,024-token sequences, giving 16,384 $(\text{batch}, \text{position})$ contexts per layer. Jensen–Shannon divergence is computed independently at each context using natural logarithms and is therefore reported in nats; probabilities passed to a logarithm are clamped below at 10^{-12} and are not renormalized after clamping. The per-layer value is the micro-average over all 16,384 contexts. For `mixed_state_cosine`, cosine similarity is first computed along the model-width dimension for each context and then averaged over the same contexts. The depth-routed mixed state spans the complete model width and is formed before the token-attention projections, so no reduction over token-attention heads is involved. A cross-layer summary excludes the first, single-source layer and gives every remaining layer equal weight.

The reported $\cos(a, v)$ uses one validation batch of four 1,024-token sequences. After backpropagation at the tied point, the first layer is excluded; all remaining QK-route gradients are concatenated into a, all remaining V-route gradients are concatenated into v, and one global cosine is computed. It is not an average of layerwise cosines. The machine-readable convention is recorded in `data/route_diagnostic_protocol.json`.

F Statistical Reporting

For each seed s, the paired effect is

$$\Delta_s = \mathrm{NLL}_{\mathrm{Block}}(s) - \mathrm{NLL}_{\mathrm{RD}}(s). \tag{16}$$

We report $\bar{\Delta}$, sample standard deviation across the five paired effects, and a two-sided 95% Student-t interval with four degrees of freedom. A deterministic paired bootstrap resamples the five seed-level

differences and uses percentile endpoints. The S mean, standard deviation, t interval, and bootstrap interval are 0.030118, 0.003036, $[0.026348, 0.033887]$, and $[0.028064, 0.032828]$. The corresponding M quantities are 0.024691, 0.011983, $[0.009812, 0.039570]$, and $[0.016267, 0.034742]$. No independence claim is made beyond the seed-level pairing. The two scale-level intervals are unadjusted, pointwise summaries. No seed-level interval is computed for the single L pair.

G Compute Environment and Runtime Measurements

The S/M experiments ran as independent single-GPU jobs on two NVIDIA RTX 5090 GPUs (32,607 MiB each), without distributed data parallelism. The archived environment reports Python 3.12.3, PyTorch 2.12.1+cu130, CUDA 13, cuDNN 92000, and driver 595.71.05; Appendix C.4 records the L environment.

In the archived seed-123 runs, S throughput is 56,158 tokens/s for Block and 50,505 for RD, with peak allocated memory of 16.005 and 17.647 GB. M throughput is 21,616 versus 19,697 tokens/s, with 21.015 and 23.132 GB peak allocated. These run records are separate from, and not averaged with, the three-round M timing benchmark in Section 5.7.