ONE FULL-ATTENTION LAYER IS ENOUGH FOR HYBRID LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid language models replace most softmax attention with linear mixers such as Gated DeltaNet and keep about 25% of the layers as full attention, the recipe of Jamba, Samba, and Qwen3-Next. Those attention layers exist for one capability, long-context retrieval, and they are also the part of the model whose serving cost grows with context: at 32K the six attention layers of a 24-layer 25% hybrid hold 0.38 GB of KV cache per sequence. We ask whether the recipe can be replaced by a hybrid with a *single* full-attention layer. Using 65 pretrained hybrids at 403M and 1.27B (identical data and optimization, 2–6 seeds per cell) and over 80 short continued-pretraining runs, we find that it can, without giving anything up. Against the 25% recipe, one layer matches loss (2.619 vs. 2.616), downstream average, real-data recall (SWDE and FDA), and, after a few hundred needle-augmented continued-pretraining steps, real-text retrieval to 32K, with $1/6$ of the attention KV cache ($1/24$ of full attention). One layer is *necessary*: pure linear models collapse on real-data recall and cannot learn real-text needle retrieval even when trained on it. It is *sufficient* because what bounds real-text retrieval in every architecture we test, from one attention layer to twenty-four, is data and training length: under plain pretraining every model sits on a recency floor beyond 4K, and 500 needle-augmented steps (5% of pretraining tokens) lift a single layer from 0.16 to 0.96 at 8K while 12 or 24 layers learn no better. The remaining choices are secondary knobs with mapped effects: placement (mid-stack for loss and FDA, top for a reliably forming pretrained circuit), looping the layer with shared weights (matches three stacked layers everywhere and beats them when needle data is scarce, at $3\times$ the single-layer KV cache), and inference-time looping (a free unlock for an unformed circuit that becomes harmful once the circuit is trained).

1 INTRODUCTION

Long context is where transformers are expensive: KV cache and attention compute grow with sequence length, layer by layer. Linear-attention hybrids—stacks in which most softmax-attention layers are replaced by linear sequence mixers (Katharopoulos et al., 2020) such as Mamba (Gu & Dao, 2024; Dao & Gu, 2024), RWKV (Peng et al., 2023), RetNet (Sun et al., 2023), GLA (Yang et al., 2024b), or (Gated) DeltaNet (Yang et al., 2024c; 2025b)—exist largely to make long context affordable, and are rapidly becoming the default recipe for efficient LLMs (Lieber et al., 2024; Ren et al., 2025; De et al., 2024; Waleffe et al., 2024; MiniMax, 2025; NVIDIA, 2025; Qwen Team, 2025). The recipe keeps 20–25% of the layers as full attention, interleaved through the stack, a choice justified by perplexity sweeps and by synthetic needle-in-a-haystack probes (Kamradt, 2023; Hsieh et al., 2024; Liu et al., 2024; Bai et al., 2024). Each retained attention layer buys back the one capability linear mixers lack—precise long-context retrieval—at a serving cost that compounds with context length (Shazeer, 2019; Ainslie et al., 2023): the linear layers hold a constant-size state, so at long context the attention layers *are* the memory cost of the model. The implicit assumption is that retrieval is bought by *stacking* attention layers: more layers, more retrieval.

This paper tests that assumption directly and finds it false. A hybrid with a single full-attention layer replaces the 25%-attention recipe on every axis we measure, and cuts the attention KV cache by $6\times$ ($24\times$ against full attention). Our question is not which configuration is strongest but whether the standard recipe can be replaced by one layer without giving anything up. We pretrain 65 hybrid

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.

models under identical data, ordering, and optimization—iso-parameter at 403M (24-layer mini-Qwen3 skeleton with Gated DeltaNet replacements) and 1.27B—varying the attention ratio, placement, head width, and the number of *serial attention steps* (weight-shared looping of one layer, in the spirit of Universal Transformers (Dehghani et al., 2019)), with 2–6 seeds per load-bearing cell. We probe retrieval with the standard passkey protocol, with a needle-in-a-haystack (NIAH) protocol whose haystack is *held-out real documents*, with a 2×2 cross of needle format $\times$ haystack format, and with real-data extraction tasks (SWDE, FDA) (Arora et al., 2024b). We then run over 80 short continued-pretraining (CPT) experiments that insert needles into real documents, at 4K and 16K, at both scales, across a data-scarcity grid.

One layer replaces the 25% recipe. Against the 6-attention-layer `ratio075` baseline (the loss optimum of the ratio sweep), a single attention layer placed mid-stack matches validation loss (2.619 vs. 2.616) and downstream average, is level on SWDE and higher on FDA (0.50 vs. 0.43), and after needle CPT retrieves from real text at 8K–32K in the same band; the reference recipe's own pretrained passkey retrieval (0.30 at 8K) is the weakest of the interleaved family. Attention KV cache at 32K falls from 0.38 to 0.06 GB (Section 4).

One layer is necessary and sufficient for retrieval. Necessary: pure linear models collapse on real-data recall (SWDE 0.27, FDA 0.09) and do not learn real-text NIAH beyond a recency ramp even when trained on it; one attention layer restores recall and learns NIAH fully. Sufficient: under plain pretraining *no* architecture retrieves from a real haystack beyond 4K—full attention, 12-attention-layer hybrids, and every single-layer circuit share one recency floor, and a 2×2 cross shows the haystack, not the needle, is the cause. Data and length lift the floor: 500 CPT steps with needles in real documents (5% of pretraining tokens) take a single-layer model from 0.16 to 0.96 at 8K while a needle-free control does not move; 300 steps at 16K take 16K to 0.92–1.00 and 32K to 0.70–0.96 at 1.27B; twelve and twenty-four attention layers learn no better from the same data (Section 5).

The remaining choices are knobs, and we map them. Placement: under plain pretraining the synthetic passkey circuit forms stochastically, in 1/4 of seeds for a single layer anywhere and in 3/4 when capacity is added at the *top*; mid placement gives the best loss and FDA and the least reliable circuit; bottom placement is useless. Looping versus stacking: one layer looped $\times 3$ with shared weights forms the circuit at the same rate as three stacked layers, reaches equal or higher passkey strength, matches stacked and 12-layer models on real text after CPT, and with 640 needle examples reaches 0.85 at 8K where a plain layer or three stacked layers reach 0.70: capacity inside one layer, not serial depth across layers, buys sample efficiency. Inference-time looping of a never-looped layer unlocks unformed circuits for free (0.42/0.20/0.15/0/0 $\rightarrow$ 1.00/0.96/0.70/0.40/0) and becomes harmful the moment the circuit is trained (Section 6).

Our contributions are: (i) evidence, across two scales and multi-seed cells, that a single full-attention layer replaces the standard 25%-attention hybrid recipe on loss, downstream, real-data recall, and post-CPT long-context retrieval at $1/6$ of its attention KV cache; (ii) the finding that one layer is necessary and sufficient for retrieval because the retrieval ceiling is set by data and training length, not attention count; (iii) a controlled map of the secondary knobs—placement, looping versus stacking, needle-data budget, inference-time looping—with their boundaries; and (iv) a methodological record: several conclusions drawn from 1–2 seeds ("loops form reliably", "stacked layers never fail", "mid-stack looping is best", "top-placed FDA tracks the passkey circuit") were each overturned by additional seeds or controls.

2 Related work

Hybrid linear-attention LLMs. Jamba (Lieber et al., 2024), Samba (Ren et al., 2025), Zamba (Glorioso et al., 2024), Griffin (De et al., 2024), Nemotron-H (NVIDIA, 2025), MiniMax-01 (MiniMax, 2025), and Qwen3-Next (Qwen Team, 2025) interleave full attention with linear mixers, typically at 20–25% attention; Waleffe et al. (2024) and Poli et al. (2024) sweep the ratio at scale and settle in the same range. Gated DeltaNet (Yang et al., 2025b) is the strongest linear mixer in our setting and the one Qwen3-Next adopts. These works select ratios by perplexity and synthetic probes; we show that the synthetic probe measures a circuit that does not transfer to real text and that real text needs one layer.

KV-cache reduction. Multi- and grouped-query attention (Shazeer, 2019; Ainslie et al., 2023), cross-layer KV sharing (Brandon et al., 2024), a single shared cache for the whole decoder (Sun et al., 2024), and cache eviction (Zhang et al., 2023) shrink the KV cache of dense transformers by sharing or dropping entries. We shrink it by keeping one attention layer in a hybrid; the two are orthogonal and compose.

Recall limits of linear attention. Arora et al. (2024a) and Arora et al. (2024b) formalize the recall–throughput tradeoff and introduce the recall-intensive evaluations (SWDE, FDA) we adopt; Jelassi et al. (2024) show copying separations between attention and state-space models. We confirm both on real haystacks—pure linear models cannot learn real-text retrieval even when trained on it—and show that the separation is closed by a single attention layer.

Long-context data engineering. Xiong et al. (2024), Fu et al. (2024), and Gao et al. (2025) show that continued pretraining on longer sequences, with position-encoding adjustments (Su et al., 2024; Peng et al., 2024), extends the usable context of dense transformers. Our CPT experiments transfer this to hybrids with a single attention layer and add that the needle data, not the length alone, is what unlocks retrieval: length-only CPT leaves the recency floor untouched.

Recurrence in depth and test-time compute. Adaptive computation time (Graves, 2016), Universal Transformers (Dehghani et al., 2019), and their recent descendants *train* weight-shared recurrence: looped transformers are shown to emulate programs (Giannou et al., 2023), to learn in-context algorithms with fewer parameters (Yang et al., 2024a), to generalize in length (Fan et al., 2025), and to trade depth for iterations in reasoning (Saunshi et al., 2025); recursive (Bae et al., 2025) and depth-recurrent (Geiping et al., 2025) LLMs compare shared-weight recurrence with deeper stacks on language modeling. All loop the whole block stack. We loop one layer, the only attention in a hybrid, compare it with stacking on retrieval, and add the inference-only variant.

Emergence and training stochasticity. Induction heads form in phase transitions (Elhage et al., 2021; Olsson et al., 2022); fine-tuning is notoriously seed-sensitive (Dodge et al., 2020). We document an analogous phenomenon in pretraining hybrids and show that a synthetic probe can ignite while the real-text capability stays absent.

3 Experimental setup

Models. The base skeleton is a 24-layer mini-Qwen3 (Yang et al., 2025a) (hidden 1024, 16Q/4KV heads (Ainslie et al., 2023) of dim 128 with per-head QK-norm, SwiGLU $3\times$, RoPE $\theta{=}10^6$ (Su et al., 2024)), in which designated layers are replaced by Gated DeltaNet (5 heads $\times$ 128, expand-v 2). Attention and GDN layers are iso-parameter (5.24M vs. 5.26M), so all 403M variants share 402.7–403.2M parameters. The 1.27B variants scale hidden to 2048 and FFN to 6144 (GDN 7 heads; iso-parameter 10.5M vs. 11.0M). `top`/`mid`/`bot` denote 23 GDN layers plus one full-attention layer at layer 23/11/0; `loop`k (*looping*) re-enters that pre-norm block k times with shared weights; `3A` (*stacking*) uses three independent attention layers (top: 21–23, mid: 10–12); `wide3x` is one layer with 48 heads; `ratio`r is the standard interleaved recipe with r% GDN (`ratio050`: 12 attention layers; `d48`: 48 half-width layers, 12 attention).

Pretraining. FineWeb-Edu (Penedo et al., 2024), pre-tokenized with the SmolLM2 tokenizer (Allal et al., 2025), identical order across runs; 403M: 9,500 steps $\times$ 0.52M tokens $=$ 5B tokens, AdamW (Loshchilov & Hutter, 2019), LR 10^{-3} with a warmup–stable–decay schedule (Hu et al., 2024), sequence length 4096; 1.27B: 19,000 steps $=$ 10B tokens, LR 6×10^{-4}. 54 runs at 403M and 11 at 1.27B.

Needle-augmented continued pretraining (CPT). Starting from a pretrained checkpoint with a fresh optimizer, a fraction f of training blocks is rewritten into a NIAH example: a needle sentence stating a five-digit code is inserted at a random depth of the real-document block and a question plus the answer is appended at the block's end; the loss is the ordinary next-token loss on all tokens. Four paraphrase templates $\times$ eight objects are used for training; the evaluation template is excluded. Default: $f{=}0.25$, 500 steps at 4K (5% of pretraining tokens) or 300 steps at 16K, LR 10^{-4}. Controls use $f{=}0$.

Evaluation. (i) validation loss; (ii) six- to nine-task downstream averages via lm-eval-harness (Gao et al., 2023) (HellaSwag, ARC-e/c, PIQA, Winogrande, LAMBADA, SciQ, OpenBookQA, BoolQ;

Table 1: Capability divergence at 403M (identical data and optimization). Passkey cells are per-length success rates; "extrap." is the 32K minus 4K per-position loss gap.

GDN share	val loss	extrap.	passkey 2K/4K/8K/16K/32K				
0% (full attn)	2.680	+0.709	1.00	1.00	0.58	0.20	0
25%	2.638	+0.068	1.00	1.00	0.62	0.50	0.30
50%	2.622	+0.041	0.80	1.00	0.40	0.40	0.20
75%	**2.616**	+0.022	0.60	0.40	0.30	0.18	0
87.5%	2.620	+0.002	0.96	0	0	0	0
100%	2.637	−0.017	0.44	0.20	0.02	0	0

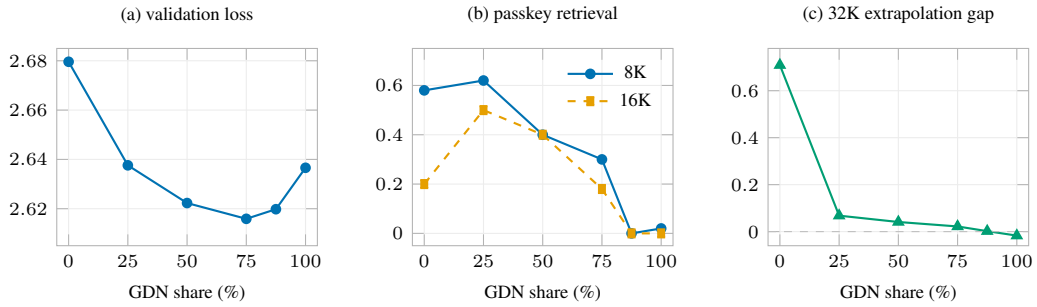

Figure 1: Capability divergence across the attention/GDN ratio (403M, identical data and optimization). The three axes optimize at *opposite* ends: loss is lowest at 75% GDN (a), retrieval collapses above 50% (b), and length extrapolation improves monotonically with GDN share (c). Perplexity-only selection picks a retrieval-impaired model; the reference recipe (`ratio075`, 75% GDN) sits at that optimum.

acc_norm where defined); (iii) length extrapolation (per-position loss to 32K); (iv) *passkey*: a five-digit key in repetitive filler, 5 lengths (2K–32K) × 5 depths × 30 keys, greedy decoding, reported as per-length means "2K/4K/8K/16K/32K"; (v) *real-text NIAH*: the same grid with held-out FineWeb-Edu documents as haystack and a natural-language needle and question (10 keys per cell); (vi) the 2 × 2 cross of needle format (passkey vs. natural) × haystack (filler vs. real); (vii) strict RULER-style multi-key NIAH and 4-hop variable tracking; (viii) real-data recall: SWDE and FDA (Arora et al., 2024b), 300 samples, contains-match. "Formed" means passkey success ≥ 0.55 at 8K, which in every case coincides with full-depth hits at 2K–4K.

4 One attention layer replaces the recipe

4.1 The reference recipe

Capabilities diverge across the ratio sweep. Loss is lowest at 75% GDN (2.616 vs. 2.680 for full attention); length extrapolation improves monotonically with GDN share (full attention degrades by $+0.709$ at 32K, all $\geq$25% GDN configurations by ≤ 0.07); but passkey retrieval falls off a cliff above 50% GDN (Figure 1, Table 1). Depth analysis shows the failure mode above the cliff is *recency*: only keys hidden in the final 25% of the context are recovered, i.e., the GDN state answers, not a retrieval circuit. Selection by perplexity alone picks a retrieval-impaired model; the 25% recipe sits exactly at that optimum, and the sweep suggests that its retrieval is bought by its six stacked attention layers. The rest of this section shows that one layer buys the same.

4.2 One layer against the recipe on every axis

Table 2 places the single-layer hybrids against the reference recipe and the denser interleaved recipes on every axis a practitioner would use to compare them. Where more than one seed exists we report the range; the reference recipe has one seed.

Table 2: One attention layer against the interleaved recipes at 403M; `ratio075` (25% attention) is the reference. Downstream is the 6-task average of Table 8 (noise $\approx \pm 0.004$); SWDE/FDA are real-data recall (contains, 300 samples); passkey@8K is the pretrained synthetic probe; NIAH@32K is real-text retrieval after 16K needle CPT; KV is the attention KV cache at 32K, bf16. Ranges span seeds where more than one was run (seed count in parentheses).

Config	attn	val loss	downstream	SWDE	FDA	passkey@8K	NIAH@32K	KV
pure GDN (`ratio100`)	0	2.637	0.526	0.27	0.09	0.02	(0.16 at 8K)	0
full attention	24	2.680	0.521	0.60	0.54	0.58	0.48	1.54 GB
25% GDN (`ratio025`)	18	2.638	0.529	**0.65**	**0.65**	0.62	–	1.15 GB
50% GDN (`ratio050`)	12	2.622	0.533	0.58	0.53	0.40	0.58	0.77 GB
d48 (Qwen3-Next-like)	12 (half-width)	**2.615**	0.531	0.61	**0.65**	0.40	0.48	0.38 GB
75% GDN (`ratio075`), reference	6	**2.616**	0.529	0.58	0.43	0.30	–	0.38 GB
mid $\times$1 (ours, 4)	**1**	2.619–2.626	0.521–0.528	0.55–0.62	0.44–0.54	0.00–0.60	0.56	**0.06 GB**
top $\times$1 (ours, 4)	**1**	2.629–2.631	0.525–0.527	0.57–0.59	0.28–0.47	0.15–0.60	0.54–0.68	**0.06 GB**
top loop3 (ours, 4)	**1** ($\times$3)	2.630–2.634	0.518–0.526	0.56–0.61	0.34–0.52	0.40–0.96	**0.58–0.72**	0.19 GB
mid3A (ours, 4)	3	2.616–2.623	0.528–**0.536**	0.53–0.58	0.53–0.70	0.20–0.72	0.50	0.19 GB

Loss and downstream. Mid placement matches the reference loss (2.619–2.626 vs. 2.616; three mid layers 2.616–2.623) and top placement pays 0.013; the loss optimum of the whole family is the 48-layer deep-thin `d48` at 2.615, which we match but do not advance. Downstream averages of every single-layer configuration are within the ± 0.004 noise band of the reference and of the 12- to 24-layer recipes; the three extra tasks of Section 3 (ARC-c, OpenBookQA, BoolQ) move every row by less than 0.01.

Real-data recall. SWDE is level across every configuration with at least one attention layer (0.53–0.66) and collapses without one (0.27). FDA separates the family by placement, not by attention count: top-placed single layers sit at 0.28–0.52, level with the reference (0.43); mid placement is higher (0.44–0.54); three mid layers or 12–18 interleaved layers reach 0.53–0.70. FDA prompts place the queried field a median of $1{,}800$ tokens before the question (600 for SWDE), so it is a 2K-range retrieval over dense text; Section 6.1 maps it over all 53 checkpoints.

Retrieval. On the pretrained synthetic probe the reference is the weakest interleaved recipe (0.30 at 8K), inside the seed band of a single layer (0.00–0.60); on real text every architecture, reference included, is on the same recency floor beyond 4K under plain pretraining, and after 300 needle-augmented steps at 16K the single-layer models reach 0.54–0.72 at 32K, the same band as the 12- and 24-layer recipes (0.48–0.58; Section 5). Length extrapolation, the axis on which linear mixers dominate, is preserved: every configuration with $\geq 25\%$ GDN degrades by ≤ 0.07 in per-position loss from 4K to 32K, against $+0.71$ for full attention.

Memory. Each attention layer holds $2\,KV\,d\,L$ bytes of cache in bf16, 0.064 GB at 32K for our 4 KV heads of width 128, while a GDN layer holds a fixed 0.3 MB state. The reference recipe's six layers cost 0.38 GB per sequence at 32K and 1.5 GB at 128K; one layer costs 0.06 and 0.26 GB, a looped layer $3\times$ that. At a serving batch of 64 sequences of 128K, the reference holds 96 GB of KV cache and the single layer 16 GB. This is the one axis on which the reference cannot catch up: the gap is $6\times$ at every length.

What the recipe is for. Nothing in Table 2 is bought by the five extra attention layers. The next section shows why: the retrieval those layers were meant to provide is bounded by data and training length, and one layer reaches the same ceiling.

5 Why one layer is enough: retrieval is bounded by data and length, not attention count

No architecture retrieves from a real haystack as pretrained. Table 3 crosses the needle format with the haystack. With held-out documents as haystack, every configuration falls to the same floor beyond 4K: only keys at depth ≥ 0.75 are recovered, and the 25-cell grids of nine top-placed checkpoints (loops, stacked layers, both scales, four seeds) differ from the circuit-free $\times 1$ baseline in at most two cells. Full attention and the 12-attention-layer hybrids are on the same floor. With repetitive filler as haystack, a natural-language needle is retrieved at 32K even by the pure linear

Table 3: Needle format × haystack format (real-text NIAH grid, 2K/4K/8K/16K/32K means; 403M seed 42 unless noted). Passkey needle + filler is the standard probe.

	passkey needle + filler	passkey needle + real text	natural needle + filler	natural needle + real text
top loop3	1.00/1.00/0.83/0.40/0.20	0.32/0.14/0.04/0.04/0	1.00/1.00/1.00/0.66/0.40	0.56/0.44/0.16/0.12/0.02
top 3A	1.00/1.00/0.65/0.40/0.24	0.14/0/0.04/0/0	1.00/1.00/0.68/0.40/0.22	0.56/0.42/0.14/0.12/0.02
1.27B loop3	1.00/1.00/0.97/0.55/0.35	0.48/0.22/0.10/0.06/0	1.00/1.00/0.96/0.58/0.38	0.56/0.46/0.16/0.14/0.02
ratio050 (12A)	0.80/1.00/0.40/0.40/0.20	0.36/0.20/0.02/0/0	1.00/1.00/0.92/0.56/0.38	0.74/0.66/0.28/0.14/0.06
full attention	1.00/1.00/0.58/0.20/0	0.52/0.38/0.08/0/0	1.00/1.00/0.56/0.20/0	0.70/0.54/0.22/0.04/0
pure GDN	0.44/0.20/0.02/0/0	0.06/0/0/0/0	1.00/1.00/1.00/1.00/0.76	0.38/0.24/0.08/0.04/0

Table 4: Real-text NIAH at 8K/16K/32K after needle-augmented CPT. 4K: $f{=}0.25$, 500 steps; 16K: $f{=}0.25$, 300 steps. After CPT every row is ≥ 0.98 at 2K and 4K except pure GDN ($0.56/0.38$ after 4K CPT). Ranges span seeds where more than one was run. "×1 ctrl" is CPT with $f{=}0$.

Start checkpoint	before CPT 8K / 16K / 32K	after 4K CPT 8K / 16K / 32K	after 16K CPT 8K / 16K / 32K
top ×1 (4 seeds; 3 at 16K)	0.16 / 0.12 / 0.02	0.86–0.96 / 0.46–0.56 / 0.32–0.36	0.96–1.00 / 0.94–0.96 / 0.54–0.68
top ×1 ctrl ($f{=}0$)	same	unchanged	unchanged
top loop3 (4 seeds; 2 at 16K)	same	0.98 / 0.46–0.72 / 0.20–0.38	1.00 / 0.96–0.98 / 0.58–0.72
top 3A	same	0.92 / 0.50 / 0.24	1.00 / 0.98 / 0.56
mid ×1	0.20 / 0.16 / 0.04	0.72 / 0.42 / 0.22	0.98 / 0.92 / 0.56
mid 3A	0.24 / 0.18 / 0.10	0.84 / 0.50 / 0.22	1.00 / 0.92 / 0.50
ratio050 (12A)	0.28 / 0.14 / 0.06	0.92 / 0.56 / 0.28	1.00 / 0.98 / 0.58
d48 (12A, 48 layers)	0.28 / 0.18 / 0.12	0.80 / 0.44 / 0.24	0.96 / 0.92 / 0.48
full attention (24A)	0.22 / 0.04 / 0	0.56 / 0.26 / 0	0.94 / 0.94 / 0.48
pure GDN	0.08 / 0.04 / 0	0.16 / 0.10 / 0	–
1.27B top ×1 (2 seeds at 16K)	same floor	0.92 / 0.50 / 0.28	0.98–1.00 / 0.98–1.00 / 0.70–0.96
1.27B top loop3 (2 seeds at 16K)	same floor	0.98 / 0.66 / 0.32	1.00 / 1.00 / 0.60–0.68
1.27B mid 3A	0.22 / 0.14 / 0.02	0.82 / 0.46 / 0.20	1.00 / 1.00 / 0.66

model (0.76). Repetitive filler barely writes into the linear state, so the passkey probe measures the ability to locate an anomalous span in an undisturbed state; real documents overwrite the state, and no amount of stacked attention learned to match through that interference during pretraining.

Within the training length, an untrained single layer looped at inference is the best zero-data retriever. On real text at 2K–4K the ranking is: a never-looped single layer executed ×3 at inference ($0.84/0.80$ and $0.92/0.82$ for two 403M seeds; $0.92/0.80$ and $0.98/0.96$ at 1.27B) ≈ mid placement ×3 (0.66–0.90) > 12–18-layer hybrids (≈ $0.75/0.65$) > full attention ($0.70/0.54$) > every *trained* top circuit, looped or stacked ($0.56/0.42$, equal to the ×1 baseline) > pure GDN ($0.38/0.24$).

Data and length lift the floor; attention count does not. Table 4 reports real-text NIAH after needle-augmented CPT. **(a)** 500 steps with $f{=}0.25$—5% of the pretraining tokens—take every hybrid with at least one attention layer to full-depth retrieval at 2K–4K and to 0.72–0.98 at 8K (full attention, whose RoPE does not extrapolate past 4K, reaches 0.56); the needle-free control does not move a single cell. **(b)** 300 steps at 16K take 16K to 0.92–0.98 for every 403M configuration and 32K to 0.48–0.72; needle-free CPT at 16K leaves the floor untouched. **(c)** Pure GDN does not learn the task from the same data (its 4K grid remains a recency ramp), while the single-layer model learns it fully; 12 attention layers (`ratio050`, `d48`) and 24 (full attention) learn it no better, and full attention is the weakest of the 16K group at 32K. At 1.27B the 32K band rises to 0.60–0.96 with no ordering by attention count; the single-layer model reaches 0.96 in one seed with $1/24$ of full attention's KV cache. One full-attention layer is thus necessary (pure GDN fails) and sufficient (more layers add nothing). **(d)** Passkey and real-text retrieval dissociate: the passkey-dead ×1 seed is the strongest plain-layer real-text retriever after 16K CPT (1.00/1.00/1.00/0.94/0.68) while its passkey score stays at 0.20 at 8K, and the strongest passkey circuit in the study (1.27B loop3) is on the recency floor before CPT.

Table 5: Passkey circuit formation under plain pretraining at the loss-optimal LR (403M unless noted): looped and stacked layers form the circuit at the same rate. "Formed" counts seeds with passkey ≥ 0.55 at 8K; the 8K values of every seed are listed. All runs at LR 10^{-3} land in a 2.61–2.65 (403M) or 2.404–2.413 (1.27B) loss band regardless of formation.

Configuration	Placement	Formed	8K passkey per seed
$\times 1$	top	1/4	0.15 / 0.60 / 0.20 / 0.40
$\times 1$	mid	1/4	0.09 / 0.20 / 0.00 / 0.60
loop $\times 2$	top	3/4	0.75 / 0.40 / 0.59 / 0.81
loop $\times 3$	top	3/4	0.83 / 0.96 / 0.40 / 0.60
independent 3A	top	3/4	0.65 / 0.53 / 0.78 / 0.40
loop $\times 2$	75% depth	2/3	0.65 / 0.74 / 0.40
loop $\times 4$	top	2/6	0.07 / 0.42 / 0.96 / 0 / 0.60 / 0.40
independent 3A	mid	1/4	0.72 / 0.37 / 0.20 / 0.40
loop $\times 2$ / $\times 3$	mid	1/3 / 1/4	0.76 / 0.51 / 0.40; 0.57 / 0.15 / 0.02 / 0.20
$3\times$ head width	top	1/2	0.68 / 0.41
$\times 1$, LR halved	top	1/1	0.80 ($+0.026$ loss)
any	bottom	0/2	0 / 0.29
$\times 1$, 1.27B	top	3/4	0.60 / 0.60 / 0.76 / 0.51 (none dead)
loop $\times 3$, 1.27B	top	2/2	0.55 / 0.97
independent 3A, 1.27B	top	0/1	0.35
independent 3A, 1.27B	mid	0/2	0.43 / 0.43

6 THE REMAINING KNOBS: PLACEMENT, LOOPING, DATA

With the replacement established, the choices left to a practitioner are where to put the layer, whether to loop it, and how much needle data to spend. Each has a mapped effect.

6.1 PLACEMENT SETS THE PRETRAINED CIRCUIT, THE LOSS, AND FDA

Looped and stacked layers form the circuit at the same rate; placement sets it. Table 5 aggregates circuit formation over all multi-seed cells at the loss-optimal LR. A single attention layer forms the passkey circuit in about one seed in four wherever it is placed, the other seeds spanning a continuum from partial to dead. Adding capacity at the *top* of the stack raises the rate to 3/4—and it does so identically whether the capacity is a loop ($\times 2$: 3/4, $\times 3$: 3/4) or three stacked layers (3/4); pooled with a 75%-depth looped placement, 11 of 15 seeds form. The same additions in the middle of the stack leave it at $\approx$1/4 (3 of 11 pooled). No configuration reaches 4/4. Over-looping ($\times 4$) collapses 2/6 seeds into pure recency (loss unaffected; loop-wise RMSNorm does not rescue it, another seed does), and bottom placement never forms a circuit. Mid placement gives the best loss at both scales (403M 2.615–2.626; 1.27B 2.4039–2.4065 vs. 2.4089–2.4105 for top) and the least stable circuit.

Controls. Halving the LR makes the $\times 1$ circuit form at a $+0.026$ loss cost, and looping adds nothing at the low LR: capacity buys robustness of formation under loss-optimal optimization pressure. Scale acts on the floor: at 1.27B the weakest of four $\times 1$ seeds is at 0.51 (403M: 0.15). Perplexity carries no information about formation: formed and unformed siblings differ by $|\Delta\text{loss}| \leq 0.007$ and downstream averages within noise. In passkey strength the loop is at least the equal of the stack: $\times 3$ seeds reach $0.83/0.96/0.40/0.60$ at 8K against $0.65/0.53/0.78/0.40$ for three stacked layers, and at 1.27B a $\times 3$ seed holds the strongest passkey result in the study (1.00/1.00/0.97/0.55/0.35).

Placement and FDA. Over all 53 pretrained 403M checkpoints with recall evaluations (Appendix A.8), FDA responds to placement depth and to the number of independent layers and to nothing else: the 27 top-placed checkpoints span 0.26–0.52 (mean 0.39) regardless of capacity (loops, width, three layers) and regardless of whether the passkey circuit formed (formed vs. unformed seeds: 0.39 vs. 0.39); 75%-depth placement averages 0.44, mid placement 0.50 over 11 single or looped checkpoints, three mid layers 0.62; interleaved recipes rise with count (0.43, 0.53, 0.65 for 6, 12, 18 layers); bottom placement is at the pure-GDN floor (0.07–0.12). The attention layer needs GDN layers below it to form queries and layers above it to use what it retrieves, and for

Table 6: Looping vs. stacking when needle data is scarce (403M, 4K CPT, LR 10^{-4}). Cells are real-text NIAH at 8K/16K, mean of seeds 42 and 43 (the 16,000 row averages all available seeds: four for $\times 1$ and loop $\times 3$, one for 3A); needle examples $= f\times$ steps $\times$ 128.

Needle examples	top $\times 1$	top 3A	loop $\times 2$	wide $3\times$	loop $\times 3$
16,000 (f=.25, 500 st.)	0.92 / 0.52	0.92 / 0.50	–	–	**0.98** / 0.59
3,200 (f=.25, 100 st.)	0.90 / 0.52	0.81 / 0.47	0.87 / 0.44	**0.98** / **0.64**	0.96 / 0.61
3,200 (f=.05, 500 st.)	0.78 / 0.41	0.76 / 0.42	0.82 / 0.45	**0.93** / 0.52	0.88 / **0.60**
640 (f=.05, 100 st.)	0.77 / 0.42	0.79 / 0.42	0.81 / 0.42	0.83 / 0.51	**0.87** / **0.59**
640 (f=.01, 500 st.)	0.62 / 0.36	0.62 / 0.37	0.63 / 0.38	**0.80** / 0.43	**0.82** / **0.52**
130 (f=.01, 100 st.)	0.16 / 0.13	0.20 / 0.13	0.23 / 0.13	0.24 / 0.14	**0.26** / 0.15

this task neither looping nor width substitutes for depth after the layer. Needle CPT at 4K or 16K leaves SWDE and FDA unchanged for every configuration ($|\Delta| \leq 0.04$, identical to the needle-free control), so the CPT that unlocks long-context retrieval does not touch short-context extraction; the gap between top and mid narrows with scale (1.27B: top 0.52–0.57, mid 0.66–0.71).

6.2 Looping matches stacking

We compare one layer executed $\times 2$–3 with shared weights against three stacked independent layers (same serial depth, $3\times$ the attention parameters) and against the 12-layer interleaved recipe.

Formation and passkey strength. Equal formation rate (3/4 vs. 3/4, Table 5); equal or higher passkey strength (Section 6.1).

Real text with ample needle data. After 4K CPT with 16,000 needle examples, four looped seeds reach 0.98 at 8K against 0.92 for three stacked layers and 0.92 for the 12-layer recipe; at 16K and 32K the looped, stacked, plain, and 12-layer models overlap (Table 4). After 16K CPT the looped layer holds the top of the 32K band (0.58–0.72) and the 12- and 24-layer recipes the bottom (0.48–0.58).

6.3 Scarce needle data: looping beats stacking

Table 6 sweeps the needle budget. As examples become scarce the configurations separate: with 640 needle examples the looped layer reaches 0.82–0.87 at 8K and 0.52–0.59 at 16K, a plain layer 0.62–$0.77/0.36$–0.42, and three stacked layers 0.62–$0.79/0.37$–0.42; two loops track the plain layer. A single layer with $3\times$ heads behaves like the loop (0.80–0.83 at 8K). With 130 examples only the looped and wide layers move at 2K–4K in both seeds. The quantity that buys sample efficiency is therefore capacity inside *one* layer—iterations of the same weights, or heads—not serial depth from stacked independent layers. Re-executing a plain layer $\times 3$ at inference after scarce CPT gives no gain ($0.80 \rightarrow 0.80$, $0.70 \rightarrow 0.72$) or harm ($0.70 \rightarrow 0.36$): the advantage is trained capacity, not test-time compute.

Cost. Three stacked layers cost $+2.6\%$ parameters and buy 0.005–0.01 lower validation loss than the loop (403M: 2.623–2.629 vs. 2.630–2.634); the 12-layer recipe buys ≈ 0.01 lower loss and ≈ 0.01 higher downstream average. The loop's KV cache is that of three layers (0.19 GB at 32K, bf16), $1/4$ of the 12-layer recipe and $1/8$ of full attention; a plain single layer is 0.06 GB. On retrieval the loop matches or beats every stacked alternative; on perplexity it pays a small, fixed price.

6.4 Inference-time looping: the zero-data version

Table 7 shows passkey accuracy when the attention layer of a *trained, frozen* model is re-executed k times at inference (KV cache $\times k$; no other change). A 403M $\times 1$ model whose circuit did not form is unlocked to near loop-trained levels at k=3; the intervention rescues every weak-formation single layer in the study, including an unformed stacked-layer seed (top 3A seed 45: $0.40 \rightarrow 0.74$ at 8K), does nothing for bottom-placed models, and on real text within the training length is the strongest zero-data intervention (Section 5).

Table 7: Inference-time looping (passkey, 2K–32K). Top: 403M $\times$1-trained (unformed seed). Bottom: 1.27B, formed seed.

	2K	4K	8K	16K	32K
403M $\times$1 (as trained)	0.42	0.20	0.15	0	0
$\rightarrow k$=2	0.88	0.82	0.60	0.40	0
$\rightarrow k$=3	**1.00**	**0.96**	**0.70**	**0.40**	0
1.27B $\times$1 (as trained)	1.00	1.00	0.60	0.40	0.20
$\rightarrow k$=3	1.00	1.00	**0.88**	**0.50**	**0.28**

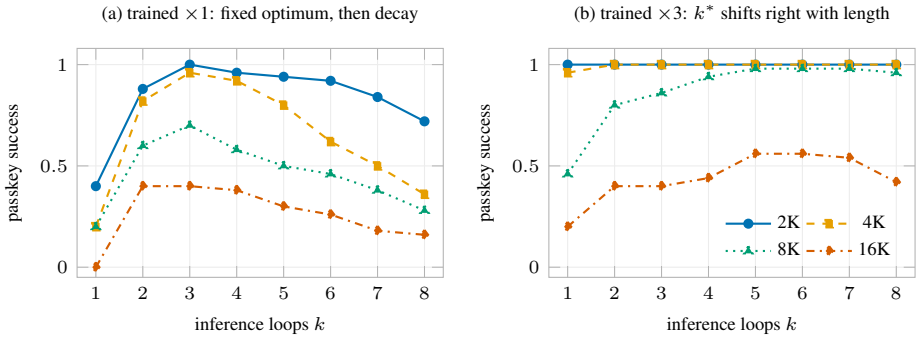

Figure 2: Two regimes of inference-time looping (403M, passkey success vs. loop count k; series are context lengths, identical assignment in both panels). A model never trained with loops (a) has a fixed sweet spot $k^* \approx 3$ at every length and degrades beyond it; a loop-trained model (b) is stable to k=8 and its optimum moves right with context length (8K: k^*=5–7; 16K: 5–6), consistent with a learned convergent iterative operator whose useful iteration count scales with the haystack.

Two regimes of $k^*(L)$. Sweeping $k \in [1, 8]$ against context length on passkey (Figure 2) separates models by training: never-looped models have a fixed sweet spot $k^* \approx 3$ and degrade beyond it; loop-trained models are stable to k=8 with the optimum shifting right by roughly +1–2 loops per doubling of context. A trained loop is a convergent iterative operator; an untrained loop is a fixed-budget trick.

Boundaries. The inference switch is for *unformed single layers only*. It does nothing or harms once the circuit is present, whether the circuit came from seed luck (formed $\times$1 seeds: no change), from stacked parameters (`mid3A` $\times$3: 1.00 $\rightarrow$ 0.24 at 2K; the 12-layer hybrid $\times$3 executes 36 serial steps and loses retrieval), or from CPT (1.27B $\times$1 after needle CPT: real-text 1.00 $\rightarrow$ 0.04 at 2K and 0.92 $\rightarrow$ 0 at 8K; `ratio050` $\rightarrow$ 0.06; full attention $\rightarrow$ 0). It does not transfer to short-context real-data recall (SWDE 0.57 $\rightarrow$ 0.48, FDA 0.30 $\rightarrow$ 0.11), costs standard-benchmark accuracy while enabled (BoolQ 0.60 $\rightarrow$ 0.53), and cannot repair $\times$4 collapse or bottom placement. It is a per-request switch for the zero-data regime; once a few hundred CPT steps are affordable, it should be turned off.

7 Conclusion

The standard hybrid recipe keeps 25% of its layers as full attention to buy retrieval. One layer buys the same. It is necessary—pure linear mixers do not learn real-text retrieval even when trained on it—and it is sufficient, because what bounds retrieval in every architecture we tested, from one attention layer to twenty-four, is the data and the training length, not the attention count: a few hundred steps of needle-augmented continued pretraining at the target length lift a single layer to the same ceiling as full attention at $1/24$ of the KV cache. Hybrids should keep one attention layer and spend the saved budget on context length.

Reproducibility statement

All runs share one self-contained training/evaluation codebase (configs included); evaluation protocols are fully seeded; the CPT needle templates and the evaluation template are listed in the appendix. Code and per-run JSON results will be released.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query Transformer models from multi-head checkpoints. In *Empirical Methods in Natural Language Processing*, 2023.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When smol goes big—data-centric training of a small language model. In *Conference on Language Modeling*, 2025.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024a.

Simran Arora, Sabri Eyuboglu, Michael Zhang, et al. Simple linear attention language models balance the recall-throughput tradeoff. In *International Conference on Machine Learning*, 2024b.

Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. Relaxed recursive Transformers: Effective parameter sharing with layer-wise LoRA. In *International Conference on Learning Representations*, 2025.

Yushi Bai, Xin Lv, Jiajie Zhang, et al. LongBench: A bilingual, multitask benchmark for long context understanding. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

William Brandon, Mayank Mishra, Aniruddha Nrusimha, Rameswar Panda, and Jonathan Ragan-Kelley. Reducing Transformer key-value cache size with cross-layer attention. In *Advances in Neural Information Processing Systems*, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, 2024.

Soham De, Samuel L Smith, Anushan Fernando, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal Transformers. In *International Conference on Learning Representations*, 2019.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for Transformer circuits. *Transformer Circuits Thread*, 2021.

Ying Fan, Yilun Du, Kannan Ramchandran, and Kangwook Lee. Looped Transformers for length generalization. In *International Conference on Learning Representations*, 2025.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. In *International Conference on Machine Learning*, 2024.

Leo Gao, Jonathan Tow, Baber Abbasi, et al. A framework for few-shot language model evaluation. `https://github.com/EleutherAI/lm-evaluation-harness`, 2023.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Jonas Geiping, Sean McLeish, Neel Jain, et al. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, 2025.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped Transformers as programmable computers. In *International Conference on Machine Learning*, 2023.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, et al. Zamba: A compact 7B SSM hybrid model. *arXiv preprint arXiv:2405.16712*, 2024.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, et al. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024.

Shengding Hu, Yuge Tu, Xu Han, et al. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *Conference on Language Modeling*, 2024.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning*, 2024.

Gregory Kamradt. Needle in a haystack: Pressure testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`, 2023.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive Transformers with linear attention. In *International Conference on Machine Learning*, 2020.

Opher Lieber, Barak Lenz, Hofit Bata, et al. Jamba: A hybrid Transformer-Mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2024.

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

MiniMax. MiniMax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

NVIDIA. Nemotron-H: A family of accurate and efficient hybrid Mamba-Transformer models. *arXiv preprint arXiv:2504.03624*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Bo Peng, Eric Alcaide, Quentin Anthony, et al. RWKV: Reinventing RNNs for the Transformer era. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

Michael Poli, Armin W Thomas, Eric Nguyen, et al. Mechanistic design and scaling of hybrid architectures. In *International Conference on Machine Learning*, 2024.

Qwen Team. Qwen3-Next: towards ultimate training and inference efficiency. https://qwen.ai/blog, 2025.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *International Conference on Learning Representations*, 2025.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped Transformers. In *International Conference on Learning Representations*, 2025.

Noam Shazeer. Fast Transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with rotary position embedding. *Neurocomputing*, 2024.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to Transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. You only cache once: Decoder-decoder architectures for language models. In *Advances in Neural Information Processing Systems*, 2024.

Roger Waleffe, Wonmin Byeon, Duncan Riach, et al. An empirical study of Mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *North American Chapter of the Association for Computational Linguistics*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped Transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *International Conference on Machine Learning*, 2024b.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear Transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, 2024c.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025b.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, et al. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, 2023.

A Appendix

A.1 Scale and cost

KV cache is what stacking never buys back. A single attention layer holds $1/6$ of the reference recipe's KV cache and $1/24$ of full attention's at every length; the other columns of Table 2 show nothing on the other side of the ledger.

Recipe. One full-attention layer plus a few hundred steps of needle-augmented continued pretraining at the target context length (f=0.05–0.25 and 100–500 steps recover within-length retrieval; 300 steps at 16K extend it). Place it mid-stack for loss and FDA, or at the top if a reliably forming pretrained circuit matters more (3/4 vs. 1/4 with capacity added). Loop it $\times 3$ only when needle data is scarce (a few hundred examples) or no CPT is possible, at $3\times$ the single-layer KV cache and 0.005–0.01 loss. Use inference-time looping only on a never-looped layer and only before CPT.

1.27B replication. The findings hold at 1.27B/10B tokens: the $\times 1$ circuit forms in 3/4 seeds with no dead seed, the looped layer in 2/2, and the stacked top layers in 0/1 (0.35 at 8K)—looping matches or exceeds stacking here too; mid placement takes the best loss (2.4039–2.4065) and the highest downstream (0.594–0.597) with partial circuits (0.40–0.43 at 8K, looping harmful); looped and stacked top circuits are equal on the real-text floor before CPT and above 0.98 at 16K after 16K CPT, with 32K in 0.60–0.96 and no ordering by attention count. Strict multi-key NIAH rises from the 403M guess floor to 0.43–0.82 at 2K; 4-hop variable tracking stays at zero for every architecture, full attention included.

A.2 Limitations

Scales are 403M/1.27B with 5B/10B tokens; formation rates are estimated from 2–6 seeds per cell and CPT cells from 1–4 seeds; the looping-versus-stacking comparison is made at three serial steps and three stacked layers, not across a full depth sweep; the CPT recipe is evaluated with a held-out template that shares its form (a five-digit code plus a question) with the training templates, so transfer to other retrieval formats is untested, and only up to 32K; a single linear mixer (Gated DeltaNet) and data distribution are studied; the loss frontier is matched, not advanced, relative to the best deep-thin baseline; a top-placed single layer trails mid placement on FDA by ≈ 0.1 and three mid-placed or 12+ interleaved layers by ≈ 0.25 (the gap narrows at 1.27B and is untouched by CPT); multi-hop tracking is frozen at these scales for all architectures.

A.3 Per-run results for all pretraining runs

Table 8 lists every evaluated pretraining run: final validation loss, the downstream average over the six core tasks (HellaSwag, ARC-e, PIQA, Winogrande, LAMBADA, SciQ; acc_norm where defined), passkey (30 keys per cell) and real-text NIAH (10 keys per cell) as per-length means (2K/4K/8K/16K/32K) over five depths. Run names: `s_` = 403M, `b_` = 1.27B; `g23a1` = 23 GDN + 1 attention layer at the named position; `loop`k = looped k times; `3a` = three stacked layers; `ratio`r = interleaved recipe with r% GDN; `d48` = 48 half-width layers; `wide3x` = 48 heads; `l17` = layer 17 (75% depth); `lr5e4` = LR halved; `norm` = loop-wise RMSNorm; `_s`n = seed n (default 42).

Table 8: All pretraining runs.

run	val loss	downstream	passkey 2K/4K/8K/16K/32K	real-text NIAH 2K/4K/8K/16K/32K
s_d48_ratio075	2.6148	0.5311	1.00/0.84/0.40/0.46/0.30	0.76/0.64/0.28/0.18/0.12
s_g23a1_bot	2.6481	0.5248	0.00/0.00/0.00/0.00/0.00	–
s_g23a1_bot_s43	2.6456	0.5258	0.25/0.33/0.29/0.15/0.00	–
s_g23a1_l17_loop2	2.6218	0.5318	0.59/0.60/0.65/0.24/0.20	–
s_g23a1_l17_loop2_s43	2.6188	0.5276	1.00/0.80/0.74/0.40/0.20	0.64/0.48/0.20/0.14/0.02
s_g23a1_l17_loop2_s44	2.6232	0.5357	1.00/0.60/0.40/0.20/0.00	0.56/0.44/0.16/0.14/0.02
s_g23a1_loop2	2.6292	0.5321	1.00/1.00/0.75/0.39/0.20	0.56/0.40/0.16/0.12/0.02
s_g23a1_loop2_s43	2.6266	0.5248	0.53/0.60/0.40/0.20/0.20	0.56/0.42/0.16/0.10/0.02
s_g23a1_loop2_s44	2.6310	0.5226	1.00/0.92/0.59/0.40/0.16	0.60/0.42/0.16/0.12/0.02
s_g23a1_loop2_s45	2.6336	0.5283	1.00/1.00/0.81/0.39/0.20	0.56/0.42/0.16/0.14/0.02
s_g23a1_loop3	2.6326	0.5257	1.00/1.00/0.83/0.40/0.20	0.56/0.44/0.16/0.12/0.02
s_g23a1_loop3_lr5e4	2.6558	0.5180	1.00/0.97/0.59/0.40/0.20	–
s_g23a1_loop3_s43	2.6301	0.5258	1.00/0.99/0.96/0.55/0.34	0.56/0.42/0.16/0.12/0.02
s_g23a1_loop3_s44	2.6330	0.5183	0.82/0.54/0.40/0.20/0.00	0.56/0.40/0.16/0.10/0.02
s_g23a1_loop3_s45	2.6344	0.5254	1.00/1.00/0.60/0.39/0.20	0.56/0.48/0.16/0.14/0.02
s_g23a1_loop4	2.6327	0.5205	0.15/0.20/0.07/0.00/0.00	–
s_g23a1_loop4_norm	2.6413	0.5182	0.59/0.40/0.20/0.20/0.00	–
s_g23a1_loop4_s43	2.6314	0.5235	1.00/1.00/0.42/0.40/0.20	–
s_g23a1_loop4_s44	2.6362	0.5199	0.99/0.99/0.96/0.40/0.19	0.56/0.42/0.16/0.14/0.02
s_g23a1_loop4_s45	2.6355	0.5189	0.00/0.11/0.00/0.00/0.00	–
s_g23a1_loop4_s46	2.6320	0.5276	1.00/1.00/0.60/0.40/0.20	–
s_g23a1_loop4_s47	2.6344	0.5263	1.00/1.00/0.40/0.20/0.00	–
s_g23a1_mid	2.6190	0.5278	0.20/0.08/0.09/0.00/0.00	0.68/0.58/0.20/0.16/0.04
s_g23a1_mid_loop2	2.6160	0.5274	1.00/1.00/0.76/0.43/0.07	0.68/0.56/0.18/0.14/0.06
s_g23a1_mid_loop2_s43	2.6213	0.5323	0.34/0.53/0.51/0.31/0.20	–
s_g23a1_mid_loop2_s44	2.6221	0.5206	0.79/0.60/0.40/0.20/0.00	–
s_g23a1_mid_loop3	2.6149	0.5309	1.00/0.97/0.57/0.39/0.19	0.80/0.66/0.22/0.10/0.04
s_g23a1_mid_loop3_s43	2.6221	0.5312	0.40/0.22/0.15/0.19/0.00	–

run	val loss	downstream	passkey	real-text NIAH
s_g23a1_mid_loop3_s44	2.6206	0.5251	0.02/0.00/0.00/0.00/0.00	–
s_g23a1_mid_loop3_s45	2.6265	0.5246	0.80/0.40/0.20/0.18/0.00	–
s_g23a1_mid_s43	2.6236	0.5226	0.41/0.39/0.20/0.00/0.00	–
s_g23a1_mid_s44	2.6249	0.5284	0.08/0.00/0.00/0.00/0.00	0.60/0.44/0.16/0.12/0.02
s_g23a1_mid_s45	2.6260	0.5208	1.00/1.00/0.60/0.40/0.20	0.66/0.58/0.22/0.18/0.08
s_g23a1_top	2.6287	0.5262	0.42/0.20/0.15/0.00/0.00	0.56/0.44/0.16/0.12/0.02
s_g23a1_top_lr5e4	2.6548	0.5166	1.00/1.00/0.80/0.40/0.20	–
s_g23a1_top_s43	2.6287	0.5246	1.00/1.00/0.60/0.40/0.20	0.56/0.42/0.16/0.12/0.02
s_g23a1_top_s44	2.6301	0.5252	0.57/0.40/0.20/0.15/0.00	0.56/0.42/0.16/0.12/0.02
s_g23a1_top_s45	2.6305	0.5272	1.00/0.80/0.40/0.20/0.03	0.56/0.42/0.16/0.14/0.02
s_g23a1_wide3x	2.6243	0.5251	0.95/0.86/0.68/0.33/0.20	–
s_g23a1_wide3x_s43	2.6224	0.5345	0.61/0.65/0.41/0.30/0.19	–
s_mid3a	2.6155	0.5362	1.00/1.00/0.72/0.40/0.38	0.74/0.62/0.24/0.18/0.10
s_mid3a_s43	2.6211	0.5312	0.97/0.60/0.37/0.20/0.00	0.68/0.60/0.18/0.10/0.02
s_mid3a_s44	2.6213	0.5310	0.94/0.60/0.20/0.19/0.00	0.66/0.58/0.20/0.10/0.02
s_mid3a_s45	2.6227	0.5282	1.00/0.80/0.40/0.20/0.00	0.76/0.62/0.20/0.06/0.00
s_ratio000	2.6796	0.5205	1.00/1.00/0.58/0.20/0.00	0.70/0.54/0.22/0.04/0.00
s_ratio025	2.6376	0.5293	1.00/1.00/0.62/0.50/0.30	0.74/0.66/0.30/0.18/0.12
s_ratio050	2.6223	0.5327	0.80/1.00/0.40/0.40/0.20	0.74/0.66/0.28/0.14/0.06
s_ratio075	2.6159	0.5293	0.60/0.40/0.30/0.18/0.00	0.70/0.66/0.22/0.08/0.02
s_ratio087	2.6198	0.5292	0.96/0.00/0.00/0.00/0.00	–
s_ratio100	2.6366	0.5255	0.44/0.20/0.02/0.00/0.00	0.38/0.24/0.08/0.04/0.00
s_top3a	2.6226	0.5282	1.00/1.00/0.65/0.40/0.24	0.56/0.42/0.14/0.12/0.02
s_top3a_s43	2.6262	0.5224	1.00/0.95/0.53/0.37/0.20	0.58/0.44/0.16/0.12/0.02
s_top3a_s44	2.6291	0.5241	1.00/1.00/0.78/0.39/0.20	0.56/0.42/0.16/0.12/0.02
s_top3a_s45	2.6263	0.5278	1.00/0.74/0.40/0.20/0.00	0.58/0.46/0.16/0.12/0.02
b_g23a1_loop2	2.4102	0.5949	1.00/0.79/0.40/0.20/0.20	0.56/0.44/0.16/0.12/0.02
b_g23a1_loop3	2.4128	0.5850	1.00/0.81/0.55/0.33/0.19	–
b_g23a1_loop3_s43	2.4092	0.5945	1.00/1.00/0.97/0.55/0.35	0.56/0.46/0.16/0.14/0.02
b_g23a1_mid	2.4065	0.5943	1.00/0.93/0.40/0.20/0.00	0.64/0.48/0.16/0.12/0.02
b_g23a1_top_s42	2.4098	0.5907	1.00/1.00/0.60/0.29/0.20	0.56/0.42/0.16/0.12/0.02
b_g23a1_top_s43	2.4105	0.5883	1.00/1.00/0.60/0.40/0.20	0.56/0.42/0.16/0.14/0.02
b_g23a1_top_s44	2.4089	0.5931	1.00/1.00/0.76/0.38/0.20	0.60/0.46/0.16/0.12/0.02
b_g23a1_top_s45	2.4093	0.5851	1.00/1.00/0.51/0.34/0.15	0.56/0.42/0.16/0.10/0.02
b_mid3a	2.4054	0.5896	0.99/0.83/0.43/0.28/0.18	0.68/0.56/0.22/0.14/0.02
b_mid3a_s43	2.4039	0.5966	1.00/0.81/0.43/0.24/0.09	0.86/0.70/0.28/0.18/0.10
b_top3a	2.4061	0.5890	0.90/0.44/0.35/0.18/0.00	0.60/0.46/0.16/0.12/0.02

A.4 INFERENCE-LOOP COUNT × CONTEXT LENGTH GRIDS

Passkey success (10 keys per cell) for $k = 1..8$ inference loops at each length, for the four checkpoints of Figure 2.

403M top ×1 (seed 42, unformed)	2K	4K	8K	16K	32K
$k = 1$	0.40	0.20	0.20	0.00	0.00
$k = 2$	0.88	0.82	0.60	0.40	0.00
$k = 3$	1.00	0.96	0.70	0.40	0.00
$k = 4$	0.96	0.92	0.58	0.38	0.00
$k = 5$	0.94	0.80	0.50	0.30	0.00
$k = 6$	0.92	0.62	0.46	0.26	0.00
$k = 7$	0.84	0.50	0.38	0.18	0.00
$k = 8$	0.72	0.36	0.28	0.16	0.00

403M top loop3 (seed 42)	2K	4K	8K	16K	32K
$k = 1$	1.00	0.96	0.46	0.20	0.00
$k = 2$	1.00	1.00	0.80	0.40	0.20
$k = 3$	1.00	1.00	0.86	0.40	0.20
$k = 4$	1.00	1.00	0.94	0.44	0.20
$k = 5$	1.00	1.00	0.98	0.56	0.20
$k = 6$	1.00	1.00	0.98	0.56	0.20
$k = 7$	1.00	1.00	0.98	0.54	0.20
$k = 8$	1.00	1.00	0.96	0.42	0.20

1.27B top ×1 (seed 43)	2K	4K	8K	16K	32K
$k = 1$	1.00	1.00	0.60	0.40	0.20
$k = 2$	1.00	1.00	0.88	0.40	0.20
$k = 3$	1.00	1.00	0.88	0.50	0.28
$k = 4$	1.00	1.00	0.80	0.44	0.26
$k = 5$	1.00	1.00	0.74	0.42	0.20
$k = 6$	1.00	1.00	0.68	0.40	0.18
$k = 7$	0.98	1.00	0.66	0.36	0.18
$k = 8$	0.98	0.96	0.60	0.34	0.16

1.27B top loop3 (seed 42)	2K	4K	8K	16K	32K
$k = 1$	0.88	0.46	0.24	0.20	0.02
$k = 2$	1.00	0.80	0.52	0.24	0.20
$k = 3$	1.00	0.80	0.54	0.34	0.20
$k = 4$	1.00	0.80	0.64	0.36	0.20
$k = 5$	1.00	0.78	0.76	0.40	0.20
$k = 6$	1.00	0.72	0.82	0.40	0.20
$k = 7$	1.00	0.70	0.80	0.40	0.20
$k = 8$	1.00	0.66	0.80	0.42	0.20

A.5 Over-looping collapse: depth-resolved passkey

Passkey success per depth (0.1/0.25/0.5/0.75/0.9) at 2K, 4K and 8K for every top loop4 seed and the loop-wise-RMSNorm control (seed 42). Collapsed seeds recover only the deepest positions (pure recency); loss is unaffected.

run	val loss	2K depths	4K depths	8K depths
loop4	2.6327	0.00/0.00/0.00/0.00/0.77	0.00/0.00/0.00/0.00/1.00	0.00/0.00/0.00/0.00/0.33
loop4_s43	2.6314	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	0.00/0.00/0.10/1.00/1.00
loop4_s44	2.6362	1.00/1.00/0.97/1.00/1.00	0.97/1.00/1.00/1.00/1.00	0.83/1.00/0.97/1.00/1.00
loop4_s45	2.6355	0.00/0.00/0.00/0.00/0.00	0.00/0.00/0.00/0.00/0.53	0.00/0.00/0.00/0.00/0.00
loop4_s46	2.6320	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	0.00/0.00/1.00/1.00/1.00
loop4_s47	2.6344	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	0.00/0.00/0.00/1.00/1.00
loop4_norm	2.6413	0.00/0.00/0.97/1.00/1.00	0.00/0.00/0.00/1.00/1.00	0.00/0.00/0.00/0.00/1.00

A.6 Protocols

Passkey. Prefix: "*There is an important piece of information hidden inside a lot of irrelevant text. Find it and memorize it. I will quiz you about it.*" Filler, repeated to fill the length: "*The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.*" Needle at the target depth: "*The pass key is* {*key*}*. Remember it.* {*key*} *is the pass key.*" Query: "*What is the pass key? The pass key is*". Keys are five-digit integers seeded per (length, depth, sample); greedy decoding of 10 tokens; a hit requires the key string in the output.

Real-text NIAH. Prefix: "*Read the following documents carefully. Somewhere inside them there is one important fact you will be asked about.*" Haystack: a contiguous slice of held-out FineWeb-Edu validation tokens, a different slice per (length, sample). Needle: "*Remember this: the special access code for the archive room is* {*key*}*. Keep it safe.*" Query: "*Question: what is the special access code for the archive room? Answer: the special access code for the archive room is*". Same grid, keys, decoding and scoring as passkey. The 2×2 cross swaps needle/query text and haystack independently.

Needle-augmented CPT. A training block (4K or 16K tokens of real text) is rewritten with probability f: a needle is inserted at a uniform depth in $[0.05, 0.95]$ and the question, a space, the key and EOS are appended at the end, truncating the real text to fit. Templates (object $\in$ {vault, server room, storage locker, garden gate, lab door, supply cabinet, back office, tool shed}):

- "Note this down: the secret code for the {obj} is {key}. Keep it in mind." / "Question: what is the secret code for the {obj}? Answer: the secret code for the {obj} is"
- "Important: the unlock number for the {obj} is {key}." / "Question: what is the unlock number for the {obj}? Answer: the unlock number for the {obj} is"
- "For the record, the pin for the {obj} is {key}." / "Question: what is the pin for the {obj}? Answer: the pin for the {obj} is"
- "The entry password of the {obj} is {key}, remember it." / "Question: what is the entry password of the {obj}? Answer: the entry password of the {obj} is"

The evaluation phrase "special access code for the archive room" appears in none of them. Loss is next-token on all tokens; optimizer is fresh AdamW, LR 10^{-4}, warmup 10% of steps (max 20), cosine decay over the second half.

RULER v2. Multi-key NIAH with six needles of the same form and distinct keys; a sample counts only if the queried key is emitted and no distractor key is; 3 depths $\times$ 20 samples per length. Variable tracking: three parallel four-hop assignment chains; a sample counts only if the full chain value is emitted. A first version (four needles; single chain) was discarded after an audit showed guess floors of 25% and a copy-the-only-number shortcut.

A.7 ALL CONTINUED-PRETRAINING RUNS

Table 9 lists every CPT run: needle fraction f / sequence length / steps, final validation loss (at the CPT sequence length), real-text NIAH per length (2K/4K/8K/16K/32K), the same with inference loops k=3, and passkey at 8K (10 keys). Run names: `ft_` = f=0.25, 500 steps at 4K (`_ctrl`: f=0); `ft16k_` = 300 steps at 16K; `sc_`⟨cfg⟩`_f`⟨f⟩`_s`⟨steps⟩ = scarcity grid; `_b_` = 1.27B. The start checkpoint is the pretraining run named after the prefix (seed 42 unless a `_s`n suffix is given), e.g. `ft16k_loop3_s43` starts from `s_g23a1_loop3_s43`.

Table 9: All continued-pretraining runs.

run	f/len/steps	val	NIAH	NIAH, k=3	pk@8K
ft_b_loop3	0.25/4096/500	2.4068	1.00/1.00/0.98/0.66/0.32	1.00/1.00/0.98/0.66/0.32	1.00
ft_b_mid	0.25/4096/500	2.4041	1.00/1.00/0.84/0.42/0.20	1.00/1.00/0.84/0.42/0.20	0.80
ft_b_mid3a	0.25/4096/500	2.4027	1.00/1.00/0.82/0.46/0.20	1.00/1.00/0.70/0.42/0.20	0.72
ft_b_top	0.25/4096/500	2.4068	1.00/1.00/0.92/0.50/0.28	0.04/0.04/0.00/0.04/0.00	0.76
ft_d48	0.25/4096/500	2.6064	1.00/1.00/0.80/0.44/0.24	0.90/0.72/0.64/0.34/0.18	0.98
ft_loop3	0.25/4096/500	2.6238	1.00/1.00/0.98/0.72/0.38	1.00/1.00/0.98/0.72/0.38	1.00
ft_loop3_s43	0.25/4096/500	2.6219	1.00/1.00/0.98/0.56/0.26	1.00/1.00/0.98/0.56/0.26	0.98
ft_loop3_s44	0.25/4096/500	2.6242	1.00/1.00/0.98/0.46/0.20	1.00/1.00/0.98/0.46/0.20	0.52
ft_loop3_s45	0.25/4096/500	2.6254	1.00/1.00/0.98/0.62/0.36	1.00/1.00/0.98/0.62/0.36	0.94
ft_mid	0.25/4096/500	2.6105	1.00/1.00/0.72/0.42/0.22	1.00/0.98/0.68/0.42/0.22	0.70
ft_mid3a	0.25/4096/500	2.6067	1.00/0.98/0.84/0.50/0.22	0.98/0.92/0.68/0.44/0.20	1.00
ft_ratio000	0.25/4096/500	2.6702	1.00/1.00/0.56/0.26/0.00	0.00/0.00/0.00/0.00/0.00	0.60
ft_ratio050	0.25/4096/500	2.6134	1.00/1.00/0.92/0.56/0.28	0.06/0.02/0.00/0.00/0.00	0.60
ft_ratio100	0.25/4096/500	2.6284	0.56/0.38/0.16/0.10/0.00	0.56/0.38/0.16/0.10/0.00	0.40
ft_top	0.25/4096/500	2.6200	1.00/1.00/0.96/0.50/0.32	0.96/0.96/0.90/0.48/0.32	0.26
ft_top3a	0.25/4096/500	2.6141	1.00/1.00/0.92/0.50/0.24	1.00/1.00/0.80/0.42/0.22	1.00
ft_top_ctrl	0.0/4096/500	2.6198	0.56/0.44/0.16/0.12/0.02	0.60/0.56/0.24/0.18/0.04	0.20
ft_top_s43	0.25/4096/500	2.6198	1.00/1.00/0.86/0.46/0.32	0.90/0.92/0.58/0.34/0.20	0.72
ft_top_s44	0.25/4096/500	2.6218	1.00/1.00/0.96/0.56/0.36	0.96/0.94/0.86/0.42/0.24	0.40
ft_top_s45	0.25/4096/500	2.6212	1.00/1.00/0.90/0.54/0.34	0.88/0.98/0.64/0.38/0.20	0.76
ft16k_b_loop3	0.25/16384/300	2.3830	0.98/1.00/1.00/1.00/0.68	0.98/1.00/1.00/1.00/0.68	1.00
ft16k_b_loop3_s42	0.25/16384/300	2.3864	1.00/1.00/1.00/1.00/0.60	1.00/1.00/1.00/1.00/0.60	1.00
ft16k_b_mid3a	0.25/16384/300	2.3785	1.00/1.00/1.00/1.00/0.66	0.98/1.00/0.94/0.82/0.64	0.66
ft16k_b_top	0.25/16384/300	2.3831	1.00/1.00/0.98/0.98/0.96	0.16/0.10/0.20/0.14/0.08	0.80
ft16k_b_top_s43	0.25/16384/300	2.3840	1.00/1.00/1.00/1.00/0.70	0.22/0.50/0.42/0.16/0.26	1.00
ft16k_d48	0.25/16384/300	2.5832	1.00/1.00/0.96/0.92/0.48	0.86/0.72/0.72/0.42/0.24	1.00
ft16k_loop3	0.25/16384/300	2.6010	1.00/1.00/1.00/0.96/0.58	1.00/1.00/1.00/0.96/0.58	1.00
ft16k_loop3_s43	0.25/16384/300	2.5992	1.00/1.00/1.00/0.98/0.72	1.00/1.00/1.00/0.98/0.72	1.00
ft16k_mid	0.25/16384/300	2.5879	1.00/1.00/0.98/0.92/0.56	1.00/1.00/1.00/0.84/0.48	0.72
ft16k_mid3a	0.25/16384/300	2.5842	1.00/0.98/1.00/0.92/0.50	0.92/0.96/0.94/0.76/0.38	1.00
ft16k_ratio000	0.25/16384/300	2.6480	1.00/1.00/0.94/0.94/0.48	0.00/0.00/0.00/0.00/0.00	0.80
ft16k_ratio050	0.25/16384/300	2.5900	1.00/1.00/1.00/0.98/0.58	0.18/0.02/0.00/0.00/0.00	0.60
ft16k_top	0.25/16384/300	2.5980	1.00/1.00/1.00/0.94/0.68	0.98/1.00/1.00/0.92/0.58	0.20
ft16k_top3a	0.25/16384/300	2.5908	1.00/1.00/1.00/0.98/0.56	1.00/1.00/1.00/0.96/0.54	1.00
ft16k_top_s43	0.25/16384/300	2.5964	1.00/1.00/0.96/0.94/0.54	0.86/0.90/0.86/0.74/0.42	0.62
ft16k_top_s43_ctrl	0.0/16384/300	2.5963	0.56/0.42/0.18/0.12/0.02	0.84/0.80/0.28/0.22/0.10	0.60
ft16k_top_s44	0.25/16384/300	2.5991	1.00/1.00/1.00/0.96/0.56	0.96/0.98/0.98/0.92/0.54	0.40
sc_loop2_f01_s100	0.01/4096/100	2.6271	0.70/0.64/0.30/0.14/0.04	–	0.76
sc_loop2_f01_s500	0.01/4096/500	2.6205	1.00/1.00/0.68/0.46/0.20	–	0.80
sc_loop2_f05_s100	0.05/4096/100	2.6271	1.00/1.00/0.76/0.44/0.24	–	1.00
sc_loop2_f05_s500	0.05/4096/500	2.6206	1.00/1.00/0.78/0.46/0.22	–	0.80
sc_loop2_f25_s100	0.25/4096/100	2.6273	1.00/1.00/0.80/0.42/0.22	–	1.00
sc_loop2_s43_f01_s100	0.01/4096/100	2.6242	0.56/0.44/0.16/0.12/0.02	–	0.40
sc_loop2_s43_f01_s500	0.01/4096/500	2.6177	1.00/0.98/0.58/0.30/0.18	–	0.40
sc_loop2_s43_f05_s100	0.05/4096/100	2.6242	1.00/1.00/0.86/0.40/0.20	–	0.40
sc_loop2_s43_f05_s500	0.05/4096/500	2.6176	1.00/1.00/0.86/0.44/0.20	–	0.40
sc_loop2_s43_f25_s100	0.25/4096/100	2.6243	1.00/1.00/0.94/0.46/0.20	–	0.40
sc_loop3_f01_s100	0.01/4096/100	2.6303	0.78/0.66/0.32/0.16/0.04	0.78/0.66/0.32/0.16/0.04	0.98
sc_loop3_f01_s500	0.01/4096/500	2.6236	1.00/0.98/0.82/0.58/0.24	1.00/0.98/0.82/0.58/0.24	1.00
sc_loop3_f05_s100	0.05/4096/100	2.6303	1.00/1.00/0.84/0.66/0.30	–	1.00
sc_loop3_f05_s500	0.05/4096/500	2.6236	1.00/1.00/0.82/0.62/0.28	–	1.00
sc_loop3_f25_s100	0.25/4096/100	2.6304	1.00/1.00/0.96/0.68/0.38	–	1.00
sc_loop3_s43_f01_s100	0.01/4096/100	2.6282	0.66/0.54/0.20/0.14/0.02	–	0.94
sc_loop3_s43_f01_s500	0.01/4096/500	2.6218	1.00/1.00/0.82/0.46/0.20	–	0.96
sc_loop3_s43_f05_s100	0.05/4096/100	2.6283	1.00/1.00/0.90/0.52/0.20	–	0.98
sc_loop3_s43_f05_s500	0.05/4096/500	2.6218	1.00/1.00/0.94/0.58/0.24	–	0.96
sc_loop3_s43_f25_s100	0.25/4096/100	2.6285	1.00/1.00/0.96/0.54/0.22	–	0.98
sc_top3a_f01_s100	0.01/4096/100	2.6204	0.58/0.50/0.16/0.14/0.02	–	0.96
sc_top3a_f01_s500	0.01/4096/500	2.6139	1.00/0.96/0.58/0.34/0.14	1.00/0.96/0.58/0.30/0.14	1.00
sc_top3a_f05_s100	0.05/4096/100	2.6204	1.00/0.98/0.82/0.40/0.18	1.00/1.00/0.76/0.40/0.16	1.00
sc_top3a_f05_s500	0.05/4096/500	2.6140	1.00/0.98/0.78/0.42/0.18	–	1.00
sc_top3a_f25_s100	0.25/4096/100	2.6206	1.00/1.00/0.88/0.48/0.24	–	1.00
sc_top3a_s43_f01_s100	0.01/4096/100	2.6239	0.70/0.54/0.24/0.12/0.04	–	0.62
sc_top3a_s43_f01_s500	0.01/4096/500	2.6175	1.00/1.00/0.66/0.40/0.20	–	0.72
sc_top3a_s43_f05_s100	0.05/4096/100	2.6240	1.00/1.00/0.76/0.44/0.22	–	0.84
sc_top3a_s43_f05_s500	0.05/4096/500	2.6176	1.00/1.00/0.74/0.42/0.20	–	0.76
sc_top3a_s43_f25_s100	0.25/4096/100	2.6242	1.00/1.00/0.74/0.46/0.24	–	0.86
sc_top_f01_s100	0.01/4096/100	2.6263	0.60/0.48/0.16/0.14/0.02	0.72/0.60/0.34/0.22/0.10	0.20
sc_top_f01_s500	0.01/4096/500	2.6198	1.00/0.98/0.70/0.38/0.18	0.94/0.98/0.72/0.40/0.20	0.20
sc_top_f05_s100	0.05/4096/100	2.6264	1.00/1.00/0.80/0.46/0.20	0.92/0.98/0.80/0.46/0.24	0.58
sc_top_f05_s500	0.05/4096/500	2.6199	1.00/1.00/0.86/0.46/0.22	0.92/0.96/0.88/0.48/0.26	0.20
sc_top_f25_s100	0.25/4096/100	2.6266	1.00/1.00/0.96/0.50/0.36	–	0.64
sc_top_s43_f01_s100	0.01/4096/100	2.6263	0.58/0.42/0.16/0.12/0.02	0.92/0.78/0.38/0.24/0.10	0.60
sc_top_s43_f01_s500	0.01/4096/500	2.6197	1.00/1.00/0.54/0.34/0.18	0.88/0.88/0.38/0.30/0.16	0.60

run	f/len/steps	val	NIAH	NIAH, k=3	pk@8K
sc_top_s43_f05_s100	0.05/4096/100	2.6263	1.00/1.00/0.74/0.38/0.22	0.90/0.98/0.60/0.38/0.20	0.88
sc_top_s43_f05_s500	0.05/4096/500	2.6197	1.00/1.00/0.70/0.36/0.20	0.80/0.84/0.36/0.30/0.12	0.60
sc_top_s43_f25_s100	0.25/4096/100	2.6265	1.00/1.00/0.84/0.54/0.32	–	1.00
sc_wide3x_f01_s100	0.01/4096/100	2.6220	0.74/0.68/0.24/0.14/0.08	–	0.82
sc_wide3x_f01_s500	0.01/4096/500	2.6155	1.00/1.00/0.84/0.46/0.18	–	0.80
sc_wide3x_f05_s100	0.05/4096/100	2.6221	1.00/1.00/0.86/0.52/0.26	–	1.00
sc_wide3x_f05_s500	0.05/4096/500	2.6155	1.00/1.00/0.98/0.52/0.26	–	0.80
sc_wide3x_f25_s100	0.25/4096/100	2.6224	1.00/1.00/0.96/0.64/0.30	–	1.00
sc_wide3x_s43_f01_s100	0.01/4096/100	2.6201	0.66/0.60/0.24/0.14/0.04	–	0.50
sc_wide3x_s43_f01_s500	0.01/4096/500	2.6136	1.00/1.00/0.76/0.40/0.22	–	0.78
sc_wide3x_s43_f05_s100	0.05/4096/100	2.6202	1.00/1.00/0.80/0.50/0.28	–	0.90
sc_wide3x_s43_f05_s500	0.05/4096/500	2.6136	1.00/1.00/0.88/0.52/0.32	–	0.84
sc_wide3x_s43_f25_s100	0.25/4096/100	2.6204	1.00/1.00/1.00/0.64/0.36	–	1.00

A.8 Real-data recall before and after needle CPT

SWDE / FDA (contains-match, 300 samples) of the pretrained checkpoint and of the same checkpoint after needle-augmented CPT (4K: f=0.25, 500 steps; 16K: 300 steps). Needle CPT leaves both unchanged ($|\Delta| \leq 0.04$; the needle-free control moves identically), so the FDA gap of top-placed single layers is not a data effect. Pretrained 1.27B baselines: top $\times$1 0.63 / 0.52, top loop3 0.63 / 0.57, mid $\times$1 0.63 / 0.66, mid3A 0.59 / 0.71.

checkpoint	CPT	before	after	ΔFDA
403M top $\times$1	4K	0.57 / 0.30	0.56 / 0.33	$+0.03$
ctrl (f=0)	4K	0.57 / 0.30	0.56 / 0.33	$+0.03$
	16K	0.57 / 0.30	0.61 / 0.33	$+0.03$
403M top loop3	4K	0.57 / 0.36	0.54 / 0.35	-0.01
	16K	0.57 / 0.36	0.59 / 0.38	$+0.02$
403M top 3A	4K	0.66 / 0.41	0.63 / 0.40	-0.01
403M mid $\times$1	4K	0.62 / 0.54	0.59 / 0.54	0
403M mid 3A	4K	0.54 / 0.64	0.55 / 0.65	$+0.01$
403M ratio050 (12A)	4K	0.58 / 0.53	0.59 / 0.55	$+0.02$
403M full attention	4K	0.60 / 0.54	0.60 / 0.57	$+0.03$
1.27B top $\times$1	4K	0.63 / 0.52	0.59 / 0.53	$+0.01$
1.27B top loop3	4K	0.63 / 0.57	0.60 / 0.57	0

Placement. SWDE / FDA of every 403M pretrained checkpoint (all but `ratio087`), grouped by the depth of the attention layer, with each seed's passkey-8K score. In the FDA prompts the queried field lies a median of 1,816 tokens (p90 2,086) before the question, against 577 (p90 1,017) for SWDE; no prompt exceeds the 4K window. Over the 27 top-placed checkpoints, FDA does not track circuit formation (Pearson r=0.13 with passkey 8K; formed vs. unformed mean 0.39 vs. 0.39) and does not respond to capacity; it responds to placement depth and to the number of independent layers.

placement	configuration	SWDE / FDA per seed	passkey 8K per seed	mean FDA
layer 0 (bot)	$\times$1	0.27 / 0.07; 0.24 / 0.12	0 / 0.29	0.10
layer 11 (mid)	$\times$1	0.62 / 0.54; 0.55 / 0.44; 0.56 / 0.48; 0.58 / 0.52	0.09 / 0.20 / 0.00 / 0.60	0.50
	loop $\times$2	0.57 / 0.43; 0.57 / 0.55; 0.57 / 0.58	0.76 / 0.51 / 0.40	0.52
	loop $\times$3	0.58 / 0.59; 0.59 / 0.45; 0.57 / 0.46; 0.57 / 0.47	0.57 / 0.15 / 0.02 / 0.20	0.49
	3A	0.54 / 0.64; 0.57 / 0.70; 0.53 / 0.53; 0.58 / 0.61	0.72 / 0.37 / 0.20 / 0.40	0.62
layer 17	loop $\times$2	0.60 / 0.40; 0.60 / 0.48; 0.58 / 0.43	0.65 / 0.74 / 0.40	0.44
layer 23 (top)	$\times$1	0.57 / 0.30; 0.58 / 0.45; 0.57 / 0.28; 0.59 / 0.47	0.15 / 0.60 / 0.20 / 0.40	0.38
	$\times$1, LR halved	0.56 / 0.26	0.80	0.26
	loop $\times$2	0.61 / 0.41; 0.57 / 0.43; 0.57 / 0.34; 0.57 / 0.36	0.75 / 0.40 / 0.59 / 0.81	0.39
	loop $\times$3	0.57 / 0.36; 0.59 / 0.52; 0.56 / 0.34; 0.61 / 0.37	0.83 / 0.96 / 0.40 / 0.60	0.40
	loop $\times$3, LR halved	0.58 / 0.35	0.59	0.35
	loop $\times$4 (+norm)	0.64 / 0.39; 0.57 / 0.36; 0.55 / 0.34; 0.58 / 0.40; 0.57 / 0.45; 0.57 / 0.34; 0.66 / 0.41	0.07 / 0.42 / 0.96 / 0 / 0.60 / 0.40 / 0.20	0.38
	wide 3$\times$	0.53 / 0.45; 0.57 / 0.44	0.68 / 0.41	0.45
	3A	0.66 / 0.41; 0.60 / 0.43; 0.60 / 0.43; 0.57 / 0.43	0.65 / 0.53 / 0.78 / 0.40	0.43
interleaved	6A / 12A / 18A / 24A / d48	0.58 / 0.43; 0.58 / 0.53; 0.65 / 0.65; 0.60 / 0.54; 0.61 / 0.65	–	
none	pure GDN	0.27 / 0.09	–	0.09

A.9 Depth-resolved real-text NIAH before and after CPT

Per-depth success (0.1/0.25/0.5/0.75/0.9) at 8K, 16K and 32K. Before CPT every configuration recovers only the deepest positions; after CPT the early depths fill in up to the CPT length.

checkpoint	8K depths	16K depths	32K depths
403M top $\times$1 s42, pretrained	0.0/0.0/0.0/0.3/0.5	0.0/0.0/0.0/0.2/0.4	0.0/0.0/0.0/0.0/0.1
after 4K CPT	0.8/1.0/1.0/1.0/1.0	0.0/0.2/0.3/1.0/1.0	0.0/0.0/0.0/0.6/1.0
after 16K CPT	1.0/1.0/1.0/1.0/1.0	0.7/1.0/1.0/1.0/1.0	0.4/0.3/0.8/0.9/1.0
403M top loop3 s42, pretrained	0.0/0.0/0.0/0.3/0.5	0.0/0.0/0.0/0.2/0.4	0.0/0.0/0.0/0.0/0.1
after 4K CPT	0.9/1.0/1.0/1.0/1.0	0.3/0.3/1.0/1.0/1.0	0.0/0.0/0.0/0.9/1.0
after 16K CPT	1.0/1.0/1.0/1.0/1.0	0.8/1.0/1.0/1.0/1.0	0.0/0.1/0.8/1.0/1.0
403M full attention, pretrained	0.0/0.0/0.0/0.3/0.8	0.0/0.0/0.0/0.0/0.2	0.0/0.0/0.0/0.0/0.0
after 4K CPT	0.0/0.0/0.8/1.0/1.0	0.0/0.0/0.0/0.4/0.9	0.0/0.0/0.0/0.0/0.0
after 16K CPT	0.7/1.0/1.0/1.0/1.0	0.8/1.0/0.9/1.0/1.0	0.2/0.0/0.5/0.7/1.0
403M pure GDN, pretrained	0.0/0.0/0.0/0.0/0.4	0.0/0.0/0.0/0.0/0.2	0.0/0.0/0.0/0.0/0.0
after 4K CPT	0.0/0.0/0.0/0.3/0.5	0.0/0.0/0.0/0.1/0.4	0.0/0.0/0.0/0.0/0.0
1.27B top $\times$1 s42, pretrained	0.0/0.0/0.0/0.3/0.5	0.0/0.0/0.0/0.2/0.4	0.0/0.0/0.0/0.0/0.1
after 4K CPT	0.6/1.0/1.0/1.0/1.0	0.0/0.2/0.3/1.0/1.0	0.0/0.0/0.1/0.3/1.0
after 16K CPT	0.9/1.0/1.0/1.0/1.0	0.9/1.0/1.0/1.0/1.0	0.9/0.9/1.0/1.0/1.0