Triadic Linear Attention

Three-Dimensional Recurrent States for Long-Context Sequence Modeling

Paper by Oliver Sieberling, Bharat Runwal, David Jin,
Ryan Chin, Rameswar Panda, and Yoon Kim

MIT / MIT-IBM Computing Research Lab

Independent research paper discussion

arXiv:2609.36529v1 · 9 October 2026

Two short addresses create a larger memory with few new parameters.

Memory capacity is a separate scaling axis

$$S_t = S_{t-1} + k_t v_t^\top, \qquad o_t = \sum_{s \le t} (q_t^\top k_s) v_s$$

- A $d \times d$ matrix stores superposed associations.
- Overlapping keys produce retrieval interference; d dimensions allow at most d mutually orthogonal keys.
- More heads or wider values also enlarge projections, competing with the MLP for a fixed parameter budget.

Research question

How can recurrent memory grow much faster than model parameters?

Paper, Secs. 1–2.1, PDF pp. 1–2; budget comparison in Sec. 3.2, p. 7.

A second key expands the joint address space

$$\underbrace{S_t}_{d\times E\times d} = S_{t-1} + \underbrace{k_t}_{d} \otimes \underbrace{k'_t}_{E} \otimes \underbrace{v_t}_{d}$$

$$o_t = S_t \times_1 q_t \times_2 q'_t = \sum_{s\le t} (q_t^\top k_s)(q_t'^\top k'_s) v_s$$

- The effective key is $\kappa_t = k_t \otimes k'_t \in \mathbb{R}^{dE}$.
- State size grows by E; only two short address projections are added, plus small gates and convolutions.
- $E = 1$ with $k'_t = q'_t = 1$ recovers ordinary linear attention.

Paper, Sec. 2.2, Eq. (2), PDF p. 3. $\times_n$ contracts tensor axis n.

Two records can share one key without colliding

One address

Both records use $k = e_1$:

$$S = e_1 a^\top + e_1 b^\top$$

Querying with e_1 returns

$$S^\top e_1 = a + b.$$

A joint address

Use orthogonal f_1, f_2:

$$S = e_1 \otimes f_1 \otimes a + e_1 \otimes f_2 \otimes b.$$

Querying with (e_1, f_1) returns a, since $f_1^\top f_2 = 0$.

There are dE orthogonal address pairs in an ideal product basis. Learned addresses need not be orthogonal.

Teaching example derived from Eq. (2), PDF p. 3; not a paper experiment.

The delta rule updates the two-key association jointly

$$\widetilde{S}_t[:, e, :] = \alpha_{t,e} S_{t-1}[:, e, :], \qquad \widehat{v}_t = \widetilde{S}_t \times_1 k_t \times_2 k'_t$$

$$S_t = \widetilde{S}_t + \beta_t k_t \otimes k'_t \otimes (v_t - \widehat{v}_t)$$

- Forget each slice, then correct the residual at the joint address.
- With unit-norm keys, the same-address read becomes $(1 - \beta_t)\widehat{v}_t + \beta_t v_t$.
- All slices contribute to one residual: this couples their updates.
- sGLA keeps the write strength but drops the erase term: $S_t = \widetilde{S}_t + \beta_t k_t \otimes k'_t \otimes v_t$.

Paper, Eqs. (3)–(5), PDF p. 4; Algorithm 2, p. 20. Interpolation is a derived identity.

The construction fits inside a modern GDN block

1. RMSNorm the input; project and causally convolve.
2. Apply SiLU to q, k, v and softplus to q', k'.
3. Normalize q, k, q', k' individually to unit length.
4. Apply slice-wise forgetting and the joint delta update.
5. Read out, apply RMSNorm and an output gate, then project.
6. Add the residual and the standard SwiGLU MLP block.

$$\alpha_{t,e} = \exp[-\exp(A_e)\,\mathrm{softplus}(W_{\alpha,e}u_t + b_e)], \quad \beta_t = \sigma(W_\beta u_t)$$

At $E = 1$, positivity plus normalization makes $q' = k' = 1$ exactly.

Paper, Appendix D / Algorithm 2, PDF pp. 19–20. Projections are fused; Appendix A, p. 18.

Factorization avoids an E-fold block-attention cost

For a chunk of C tokens, without forgetting or the delta rule:

$$A = \mathrm{tril}\Big[(QK^\top) \odot (Q'K'^\top)\Big]$$

$$O = \sum_e \mathrm{diag}(q'_e) Q S_{[i],e} + AV$$

$$S_{[i+1],e} = S_{[i],e} + K^\top \mathrm{diag}(k'_e) V$$

Where the saving occurs

Forming A: $O(C^2 dE) \longrightarrow O(C^2(d+E))$.
State reads and writes still grow with E.

Paper, Sec. 2.5, Eqs. (6)–(8), PDF pp. 4–5. q'_e, k'_e are columns over a chunk.

Tiling the value axis makes the large state practical

One head, $d = 128$, $E = 8$

$$128 \times 8 \times 128 \times 4$$

$$= 512 \text{ KiB in FP32}$$

Too large for one Hopper SM's register file.

One tile, 32 value columns

$$128 \times 8 \times 32 \times 4$$

$$= 128 \text{ KiB in FP32}$$

Each tile keeps all E slices for its value columns.

- CuTe kernels; 64-token chunks; independent value columns.
- Training saves BF16 chunk-boundary states for backward.
- Fixed inference state does not imply fixed training activation memory.

Paper, Sec. 2.5, PDF p. 5; Appendix A, pp. 17–18.

All language models receive long-context extension

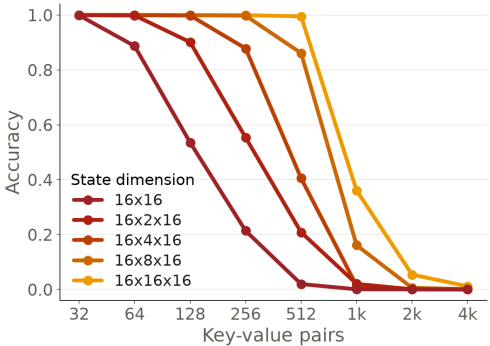
Setting	400M scale	1.3B scale
Layers / width / heads	24 / 1024 / 8	24 / 2048 / 16
Key and value dimension	128	128
Pretraining tokens	20B	65B
Pretraining context	4K	4K
Extension context	64K	64K

- FineWeb-Edu pretraining; extension adds 5 tokens per parameter using web, books, and scientific PDFs.
- Transformer: GQA-8, QK-norm, RoPE base $10\mathrm{K} \rightarrow 2\mathrm{M}$.
- Report PPL, six-task recall, ten-task zero-shot, and RULER NIAH.

Paper, Sec. 3.1, PDF pp. 5–6; Appendix D, pp. 19–20. All evaluations follow extension.

MQAR isolates a substantial increase in capacity

Multi-query associative recall: read N key–value pairs, then answer N queries.



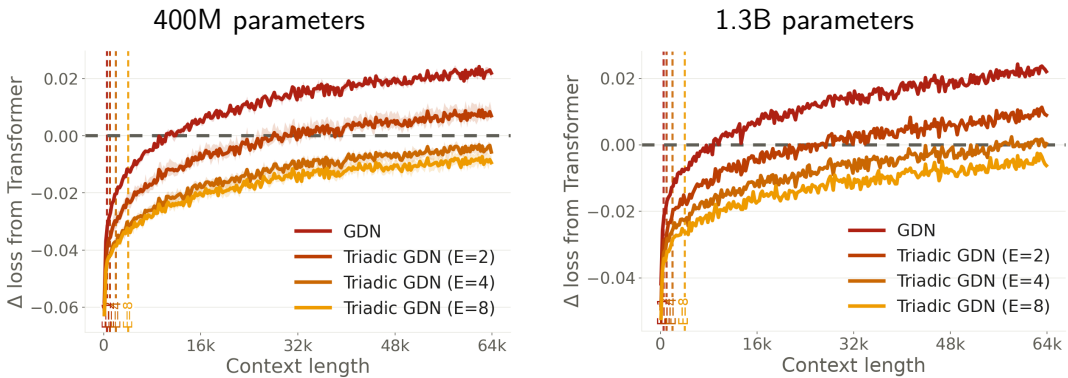
Two layers, four heads, $d = 16$; no forgetting, convolution, or delta rule in the mixer.

Increasing E moves the accuracy curve to the right.

$E = 16$: roughly $16\times$ more associations with only $1.08\times$ non-embedding parameters.

Paper, Fig. 1 / Sec. 2.3, PDF p. 3; Appendix C, p. 19. 25 seeds per (N, E) cell; read-only queries.

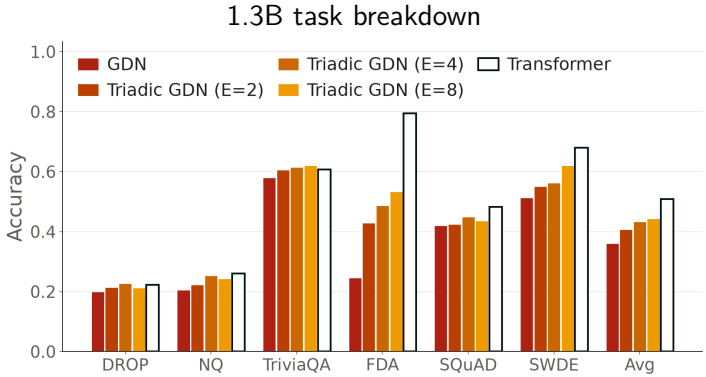
Larger states extend the long-book prediction advantage



Negative loss difference favors the recurrent model. $E = 8$ remains below the Transformer at 64K in both panels.

Paper, Fig. 2 left, PDF p. 5. Vertical dashed lines mark memory-size crossovers, not loss crossovers.

Recall improves strongly, especially on copying tasks



GDN → Triadic, $E = 8$

400M recall:
26.2 → 33.1
+6.9 percentage points

1.3B recall:
36.1 → 44.4
+8.3 percentage points

FDA and SWDE gain most. Individual tasks need not improve monotonically with E; the Transformer remains stronger on average.

Paper, Fig. 2 right, PDF p. 5; exact averages from Table 2, p. 7. 400M: three seeds; 1.3B: one seed.

At equal state size, how it is expanded matters

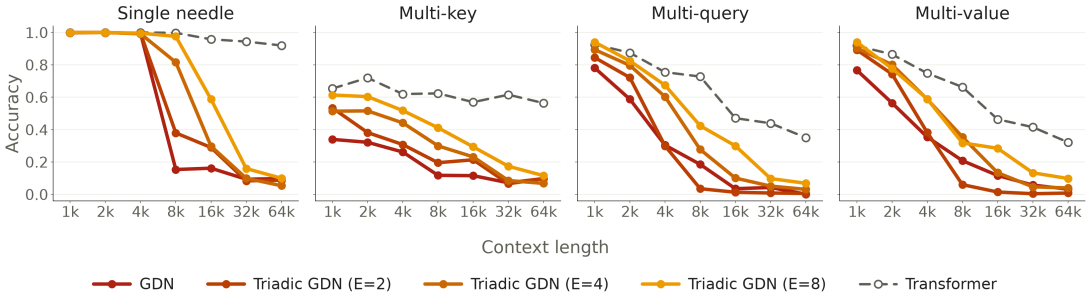
400M GDN family; 4× state (25.2 MB), approximately matched parameters.

Expansion	Wiki PPL ↓	PG19 PPL ↓	Recall ↑
Larger heads, $d = 512$	11.25	14.24	28.5
Four value heads / key	11.35	14.36	27.3
Wider values, $d_v = 512$	11.41	14.44	27.5
Triadic, $E = 4$	**10.93**	**13.79**	**31.1**

- Wider/grouped values shrink the MLP from 2816 to 640 to fit the budget; Triadic preserves it.
- Larger heads preserve projection width and still underperform.
- sGLA shows similar PPL gains; its 2× recall has a small exception (28.4 vs. 28.5).

Paper, Table 1 / Sec. 3.2, PDF pp. 6–7. PG19 column: positions 16K–64K.

Larger states delay retrieval failure, but do not remove it



- Increasing E improves retrieval at intermediate context lengths.
- At 64K, all four pure Triadic curves remain low; the Transformer retains a clear advantage.
- Book PPL and exact retrieval expose different memory demands.

Paper, Fig. 3 / Sec. 3.2, PDF p. 7. Four representative RULER tasks at 1.3B.

Continued pretraining can recover much of the benefit

Copy forget gates from $E = 1$ to 8 slices; initialize new addresses and convolutions; run long-context extension.

Scale	GDN variant	PG19 PPL ↓	Recall ↑
400M	Base	14.15	26.2
	Upcycled to $E = 8$	13.86	29.7
	$E = 8$ from scratch	13.74	33.1
1.3B	Base	10.20	36.1
	Upcycled to $E = 8$	10.00	42.0
	$E = 8$ from scratch	9.94	44.4

400M: about 71% of the long-PPL improvement and 51% of the recall improvement are recovered. sGLA also benefits.

Paper, Table 2, PDF p. 7. PG19: 16K–64K. Recovery ratios calculated from table values.

Hybrid models benefit more from expanding GDN state

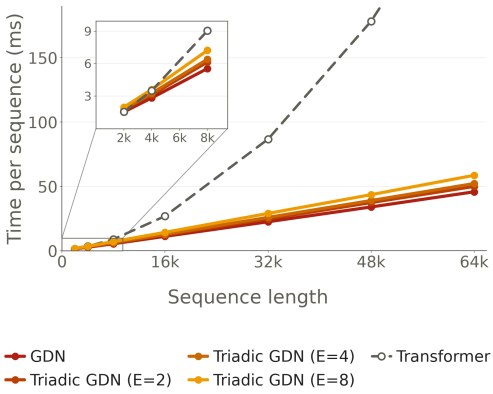
400M, 3:1 recurrent/softmax blocks, NoPE. Compare a larger KV cache with $E = 4$ recurrent state.

Hybrid	MB @64K	PG19 ↓	Recall ↑	NIAH ↑
GDN / GQA-8	206	13.50	43.7	51.5
GDN / GQA-4	407	13.47	**45.3**	53.0
Triadic / GQA-8	220	**13.39**	44.5	**56.3**

- At 64K: 46% less state memory than the larger-KV alternative, with better PPL and NIAH.
- At 4K: 31 MB vs. 30 MB; memory crossover is about 4.6K.
- Recall favors larger KV by 0.8 points (within seed spread, per authors).

Paper, Table 3 / Sec. 3.4, PDF p. 8. PG19: 4K–64K. Three-seed averages. Memory excludes weights.

Eight times the state adds about 30% block training time



Measured setting

1.3B block; H100; batch 4; forward + backward.

Overhead over GDN:

$E = 2$: 9–11%

$E = 4$: 14–15%

$E = 8$: 28–30%

$E = 8$ vs. Transformer:

3% slower at 4K;

5.1× faster at 64K.

Paper, Fig. 4 / Sec. 3.5, PDF p. 8. CuTe vs. FlashQLA/TileLang vs. FlexAttention; not decoding throughput.

Both address factorization and activation affect quality

Fixed $d_k E = 1024$
One scalar forget gate per head.

$d_k \times E$	Wiki	Long PPL
128×8	10.80	13.70
64×16	10.84	13.73
32×32	10.90	13.78

Balanced factors save projection parameters but slightly worsen PPL; MLP width is adjusted.

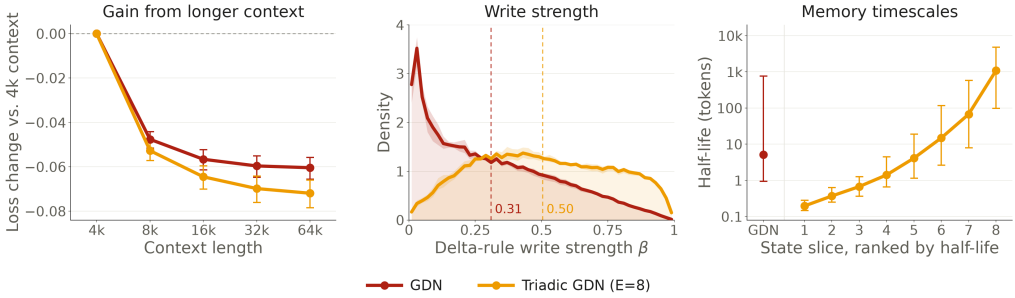
Second-key/query activation
$E = 8$, separate activation study.

Activation	Wiki	Long PPL
Softplus	10.86	13.77
Sigmoid	10.86	13.77
SiLU	10.98	13.84
Linear	11.06	13.98

Nonnegative activations win on PPL. Signed cancellation is the authors' explanation.

Paper, Table 4 / Sec. 3.6, PDF p. 9. Long PPL: PG19 16K–64K; lower is better.

The added slices learn a wider range of timescales



- Fixed final 4K targets: Triadic uses more preceding context.
- Mean write strength β: 0.31 $\rightarrow$ 0.50.
- Ranked-slice median half-lives: 0.20–1080 tokens; GDN: 5.1.

Paper, Fig. 5 / Appendix B, PDF pp. 18–19. Half-life measures forget-gate decay, not complete recall lifetime.

The gain is a better capacity–parameter trade-off

What the evidence supports

Structured product addresses enlarge recurrent memory cheaply in parameters and improve long-context PPL and recall.

- $E = 8$: about 1.2–1.3% more parameters, 8× recurrent state, and about 30% more measured block training time.
- Upcycling and hybrid results make the idea relevant beyond training a pure model from scratch.
- Fixed memory still limits exact long-context retrieval; hardware efficiency depends on the supplied kernels.

For memory research: compare state size and compute alongside parameters.

Paper, Tables 1–3, Figs. 3–4, PDF pp. 6–8; Discussion and Conclusion, pp. 9–10.

Paper and discussion directions

Oliver Sieberling, Bharat Runwal, David Jin, Ryan Chin, Rameswar Panda, and Yoon Kim.

Triadic Linear Attention: Three-Dimensional Recurrent States for Long-Context Sequence Modeling. Preprint, 2026.

arXiv:2609.36529v1 Author-provided code link

- Higher-order memories predate this work; the contribution combines modern state management, scale, and efficient kernels.
- Promising extension: grow state during training as contexts grow.
- Open design issue: allocate a fixed budget between recurrent state and exact KV storage.

Figures reproduced from the paper's author source under CC BY 4.0. Discussion based on Secs. 3–5.

Backup: the delta correction shares one triangular solve

For a $C = 64$ chunk, define $\gamma_e^r = \prod_{j=1}^{r} \alpha_e^j$ and

$$(\Gamma_e)_{rs} = \begin{cases} \gamma_e^r / \gamma_e^s, & r \geq s, \\ 0, & r < s. \end{cases}$$

$$R = \sum_e (k'_e k_e^\top) \odot \Gamma_e, \qquad R' = \sum_e (q'_e k_e^\top) \odot \Gamma_e$$

$$T = \Big[I + \mathrm{tril}\Big(\mathrm{diag}(\beta)(KK^\top \odot R), -1\Big)\Big]^{-1} \mathrm{diag}(\beta)$$

$T \in \mathbb{R}^{C \times C}$ is shared across all slices. The strictly lower triangle encodes dependence on previous writes.

Paper, Appendix A, Eqs. (9)–(10), PDF p. 17. k'_e, q'_e, γ_e are length-C columns.

Backup: the complete gated chunk recurrence

$$U = T\left[V - \sum_e \operatorname{diag}(k'_e \odot \gamma_e) K S_{[i],e}\right]$$

$$O = \sum_e \operatorname{diag}(q'_e \odot \gamma_e) Q S_{[i],e} + (QK^\top \odot R')U$$

$$S_{[i+1],e} = \gamma_e^C S_{[i],e} + K^\top \operatorname{diag}\left(k'_e \odot \frac{\gamma_e^C}{\gamma_e}\right) U$$

- U contains delta-corrected values after the shared solve.
- Slice-specific decay prevents a single precomputed TK from serving every slice; apply T to the residual instead.

Paper, Appendix A, Eq. (11), PDF p. 17. Products/divisions within diagonals are elementwise.

Backup: stable decay factors preserve near-diagonal terms

$$\frac{\gamma_e^r}{\gamma_e^s} = \exp(\log \gamma_e^r - \log \gamma_e^s), \quad r \geq s$$

- Causal log differences are nonpositive; evaluating them directly avoids positive-exponent overflow.
- Factoring around a middle reference reduces exponentials from C^2E to $2CE$, but individual factors may overflow.
- Clamping those factors can distort even near-diagonal entries.

Within-chunk total decay	e^{-200}	e^{-600}
Relative output error (clamp simulation)	0.05	1.5

Paper, Appendix A, PDF p. 17. Authors' FP64 simulation isolates the clamp; deployed kernels form terms directly.

Backup: evaluation and memory accounting

- **PG19:** roughly 1700 books, 150M tokens, 64K windows. Table 1/2 long bin: 16K–64K; Table 3: 4K–64K.
- **Recall:** six tasks; an answer is correct if the generation contains it. This is distinct from RULER NIAH.
- **NIAH average:** eight tasks and tested lengths from 1K to 64K. Figure 3 displays four tasks separately.
- **Uncertainty:** main 400M comparisons average three seeds; Table 2's 1.3B results use one seed. MQAR uses 25 seeds per cell.
- **Memory:** tables count state entries at 2 bytes each. FP32 kernel registers, weights, optimizer state, and saved training activations have separate costs.

Paper, Tables 1–3, PDF pp. 6–8; Appendices A, C, D, pp. 17–20.