Kronecker Delta Rule for Long Context Linear Attention

KronLSTM: targeted associative memory revision

Paper by A.-R. Hartl, L. Zólyomi, M. List, G. Klambauer, S. Böck, and S. Hochreiter

Research group discussion 22 October 2026
arXiv:2610.26834v1

The question: can a bounded matrix remember more selectively?

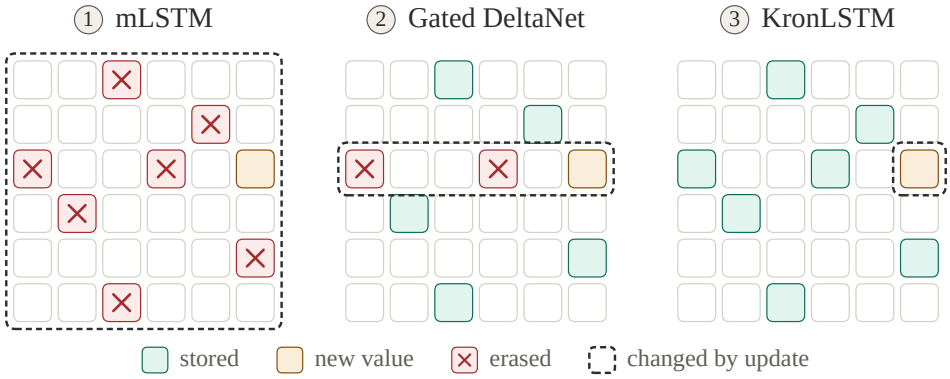
Context

Recurrent linear attention stores history in a fixed-size matrix S, so its recurrent state does not grow with context length.

- Additive writes can accumulate without control.
- Forget gates suppress old information, including useful long-range facts.
- Delta-rule corrections bound memory but can disturb earlier associations sharing a key direction.

Paper, Introduction and Sec. 2.1, PDF pp. 2–3.

Three updates differ in what they revise



Paper, Fig. 1, PDF p. 1. Conceptual illustration, not measured interference.

DeltaNet corrects a full vector; KronLSTM corrects a score

$$\text{DeltaNet:} \quad S^+ = S - \beta(Sk - v)k^\top,$$
$$\text{Kron rule:} \quad S^+ = S - \beta(v^\top Sk - c)vk^\top.$$

- $S \in \mathbb{R}^{d_v \times d_k}$; k is a key, v a value, c a scalar target.
- DeltaNet imposes $Sk \approx v$; the Kron rule imposes $v^\top Sk \approx c$.
- The Kron correction points along $vk^\top$, a direction specified by the input pair.

Paper, Eqs. (2)–(3), Sec. 2.1, PDF pp. 3–4.

The rule is gradient descent on one scalar association

$$\ell(S) = \tfrac{1}{2}(v^\top S k - c)^2, \qquad \nabla_S \ell = (v^\top S k - c) v k^\top.$$

Correction of the addressed pair

$e^+ = e\big(1 - \beta \|k\|^2 \|v\|^2\big)$, where $e = v^\top S k - c$.

- The error contracts for $0 < \beta \|k\|^2 \|v\|^2 < 2$.
- $\beta^* = 1/(\|k\|^2 \|v\|^2)$ hits the scalar target in one step.

Paper, Sec. 2.1 and Appendix A3.3, PDF pp. 3–4, 22–24.

Why "Kronecker": a delta rule in matrix space

$$a = k \otimes v = \mathrm{vec}(vk^\top), \quad s = \mathrm{vec}(S), \quad v^\top S k = a^\top s.$$

$$s^+ = (I - \beta a a^\top) s + \beta c a.$$

- a has $d_k d_v$ components, but no implementation needs to form it explicitly.
- $a^\top s = c$ is a hyperplane; at β^* the update is its orthogonal projection.
- That step is the smallest exact correction in Frobenius norm.

Paper, Eq. (4), Sec. 2.1, PDF p. 4; Appendix A3.4, PDF pp. 26–27.

Interference requires overlap in both key and value

$$\Delta(v_i^\top S k_i) = -\beta e_t \underbrace{(v_i^\top v_t)}_{\text{value overlap}} \underbrace{(k_t^\top k_i)}_{\text{key overlap}} .$$

- Orthogonality in either factor protects an old scalar association.
- At the exact-correction step, the paper proves minimum expected interference for independent isotropic probe keys and values.
- The formula is exact; the statistical optimality statement needs its probe assumptions.

Paper, Eq. (5), PDF p. 4; Appendix A3.4, PDF p. 27.

One key can support two orthogonal value directions

Teaching example

$k = e_1$, initially $S = e_1 e_1^\top$.

Write (k, e_2) with $c = \beta = 1$.

$$S^+_{\mathrm{Kron}} = (e_1 + e_2)e_1^\top, \quad S^+_{\mathrm{Delta}} = e_2 e_1^\top.$$

- Kron preserves both scalar probes: $e_1^\top S^+ e_1 = e_2^\top S^+ e_1 = 1$.
- Its vector read-out is a *sum*; downstream queries must disambiguate it.
- Capacity counts independent scalar constraints, not arbitrary full vectors.

Teaching example from Paper Sec. 2.1 and Appendix A3.3, PDF pp. 4, 24–25.

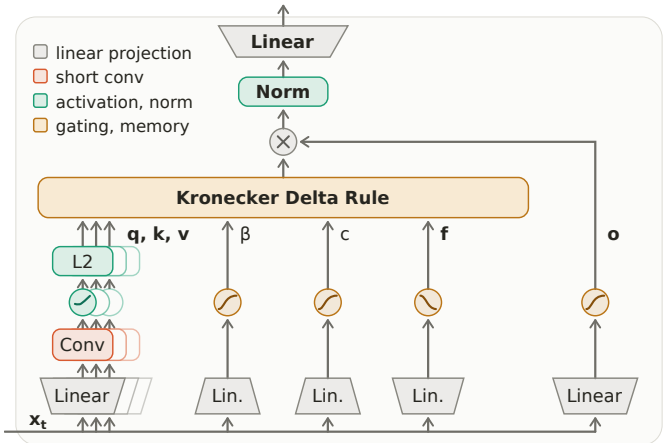
KronLSTM separates forgetting, erasing, and writing

$$S_{t+1} = \underbrace{S_t F_t}_{\text{forget}} - \underbrace{\beta_t (v_t^\top S_t F_t k_t) v_t k_t^\top}_{\text{erase}} + \underbrace{c_t v_t k_t^\top}_{\text{write}}.$$

- $F_t = \mathrm{diag}(f_t)$ decays each key channel separately.
- Sigmoid $\beta_t, c_t \in (0, 1)$ are per-head scalar gates.
- q, k, v are ℓ_2-normalized; read-out still uses an output gate and RMSNorm.

Paper, Eqs. (10)–(12), Sec. 2.2, PDF pp. 5–6.

The block keeps only matrix memories as recurrent state



Paper, Fig. 3, PDF p. 6. Projections, short causal convolutions, gates, and per-head memory.

The theorem gives a conditional second-moment bound

- Analyze the *ungated* rule with i.i.d. pairs, fixed β, and $\bar{A} = \mathbb{E}[aa^\top] \succ 0$.
- Mean memory approaches least squares; the state law approaches a unique stationary distribution.
- If $\kappa = \beta \max_i \|a_i\|^2 \le 1$ and $S_0 = 0$:

$$\mathbb{E}\|S_t\|_F^2 \le (1 + q_\beta^{t/2})^2 \frac{\mathbb{E}[c^2]}{\lambda_{\min}(\bar{A})} \le 4 \frac{\mathbb{E}[c^2]}{\lambda_{\min}(\bar{A})}.$$

$$q_\beta = 1 - \beta(2 - \kappa)\lambda_{\min}(\bar{A}) \in [0, 1).$$

Paper, Theorems 1–2, PDF pp. 4–5; Appendix A3.5–A3.6, PDF pp. 29–40.

Interpret the bound at its actual scope

- At fixed β, inconsistent associations leave stationary fluctuations of order $O(\beta)$ in covariance.
- The bound is on $\mathbb{E}\|S_t\|_F^2$, with size controlled by feature coverage $\lambda_{\min}(\bar{A})$.
- The trained network has data-dependent F_t, β_t, c_t; the theorem does not directly cover every such trajectory.
- The high-learning-rate stress test supplies separate empirical evidence of smoother gradients than mLSTM.

Paper, Fig. 2 and Theorems 1–2, PDF pp. 2, 4–5; Appendix A3.5.

Evaluation matches pretraining budgets across mixers

- 1.5B parameters, 24 layers, 100B FineWeb-Edu tokens, 8k training context.
- Same corpus, tokenizer, optimizer, batch size, and seed in each mixer comparison.
- Pure recurrent model; or 5:1 recurrent / sliding-window attention layers (window 2048).

Questions asked

Six language accuracies + LAMBADA perplexity; six real-world recall tasks at 2k; RULER at 4k, 8k, and 16k.

Paper, Sec. 3, PDF p. 6; Appendix A7, PDF p. 48.

Language modeling: highest average in both settings

Model	Linear		SWA hybrid	
	Avg-6 ↑	LMB ppl ↓	Avg-6 ↑	LMB ppl ↓
KDA	58.12	12.67	57.85	12.60
GDN-2	57.94	12.54	58.79	11.85
KronLSTM	**59.72**	**11.72**	**59.56**	**11.18**

- Linear Avg-6 lead over KDA: **+1.60 accuracy points**.
- KronLSTM's own hybrid Avg-6 is slightly lower than its linear Avg-6.

Paper, Table 1, PDF p. 7. Avg-6 averages five reasoning accuracies and LAMBADA accuracy.

Real-world recall: small average lead, mixed task results

2k recall Avg-6 ↑	Linear	SWA hybrid
GDN	29.81	35.27
Mamba-3	28.41	36.96
KronLSTM	**30.31**	**37.40**

- KronLSTM wins three of six tasks in each setting.
- Linear SWDE/SQuAD/FDA improve; linear DROP is 19.31 vs GDN's 21.18.
- Hybrid gains are substantial for all families, so setting matters.

Paper, Table 2, PDF p. 8. Average of SWDE, SQuAD, FDA, TriviaQA, NQ, DROP.

RULER reveals the benefit and a sharp S1 failure

Linear, 16k evaluation	S1	S3	MQ	MV
GDN	99.2	5.4	8.6	13.8
KronLSTM	11.6	**11.8**	**12.7**	**17.6**

- Models trained at 8k; 16k is length extrapolation.
- S1 repeats artificial noise sentences; the authors suggest recurring-key residual accumulation.
- Natural-text needles favor KronLSTM more consistently.

Paper, Table 3 and Sec. 3.2, PDF p. 9. All cells are accuracy percentages.

32k continuation improves natural-text retrieval

- Branch at 75B tokens; continue 25B at 32,768 context on a long-document mix.
- Token-matched comparison with GDN-2; architecture and tokenizer unchanged.

Linear RULER	16k evaluation		32k evaluation	
Model	S3	MV	S3	MV
GDN-2	7.4	8.9	0.8	7.2
KronLSTM	**41.8**	**30.1**	**8.6**	**11.3**

Paper, Sec. 3.3, PDF p. 9; Appendix Table A4, PDF p. 47. Accuracy %.

Length extension still leaves difficult 32k cases

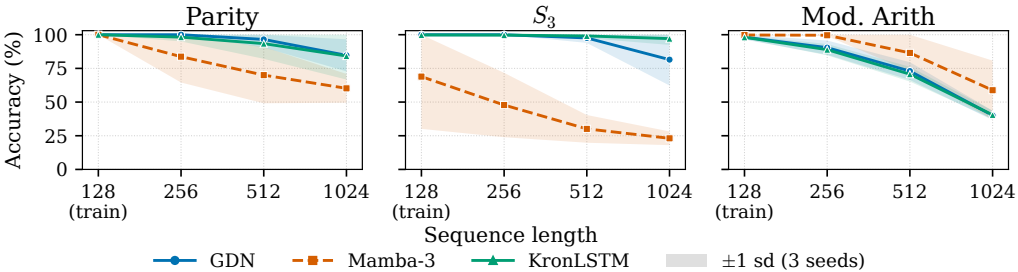
- At 32k, KronLSTM's five natural-text NIAH accuracies are only 3.0–11.3%.
- It leads GDN-2 on S2/S3/MK/MQ/MV, while GDN-2 wins artificial S1 (97.8 vs 81.4%).
- The long-document mixture changes along with sequence length, so the gain cannot be assigned to length alone.

Reading the claim

Better relative retention at long contexts; retrieval remains hard in absolute terms.

Paper, Fig. 4, PDF p. 10; Appendix Table A4 and pretraining details, PDF pp. 47–48.

State tracking is competitive on two of three tasks



Paper, Fig. 5, PDF p. 10. Trained up to length 128, evaluated to 1024; mean ± SD over 3 seeds. State-tracking KronLSTM uses $\beta \in (0, 2)$.

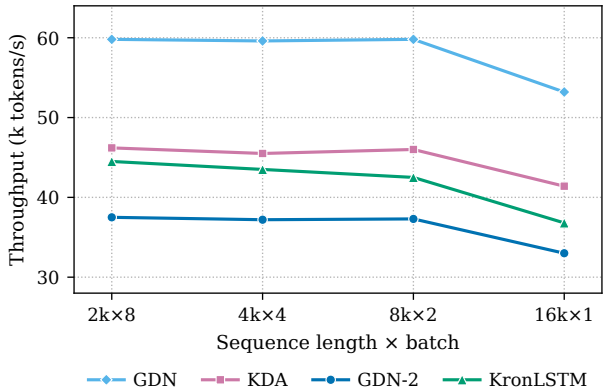
Write-gate ablation isolates the retrieval gain

1.5B linear	Avg-6	NIAH @4k	@8k	@16k
Base KronLSTM	59.72	**59.5**	30.5	**14.3**
No ℓ_2 norm on v	59.41	39.6	**32.7**	12.6
Coupled erase/write	**60.00**	46.9	29.4	11.8
Fixed unit target	58.17	41.9	26.2	10.1

- A learned, independent write target matters especially for retrieval.
- The coupled variant is slightly better on language Avg-6.

Paper, Appendix Table A3, PDF p. 47. NIAH values average six tasks at each length.

Throughput is above GDN-2, below GDN



Paper, Fig. 6 and Sec. 3.5, PDF p. 10. Single H100-80GB; 16,384 tokens/step; KronLSTM 36.8–44.5k tokens/s.

Take-home messages

1. A scalar key–value target makes each memory correction local in *both* key and value directions.
2. Conditional theory bounds the ungated rule's expected squared memory norm; the full network is supported by design constraints and experiments.
3. At matched pretraining budgets, KronLSTM improves average language and recall scores and often long natural-text retrieval.
4. Artificial repeated-noise S1 and low absolute 32k retrieval remain the main practical cautions.

Paper, Secs. 2–4; Tables 1–4 and Appendix Table A4.

Backup: chunkwise parallelization without explicit $k \otimes v$

- For chunk token r, write coefficient $u_r = c_r - \beta_r v_r^\top S_{r-1} F_r k_r$.
- Inside a chunk, earlier writes affect u_r through key and value Gram products.

$$\left[I + \mathrm{diag}(\beta)(G^{kk} \odot G^{vv})\right] u = c - \beta \odot p.$$

- G^{kk} is strictly lower triangular and decay-weighted; $G^{vv} = VV^\top$.
- The system is unit lower triangular; only matrix states pass between chunks.

Paper, Appendix A8, PDF pp. 49–51.

Backup: retention depends on the feature spectrum

$$\mathbb{E}[s_t] = (I - \beta\bar{A})^t s_0 + \left[I - (I - \beta\bar{A})^t\right] s_{\mathrm{LS}}.$$

- Initial-memory modes decay at rates $(1 - \beta\lambda_j(\bar{A}))^t$.
- If keys and values are independent, $\bar{A} = \mathbb{E}[kk^\top] \otimes \mathbb{E}[vv^\top]$.
- Value-direction diversity can slow forgetting; the result needs the stated independence and non-expansive step conditions.

Paper, Appendix A5, PDF pp. 44–46.