The Query Knows What to Forget: A Second Erase Direction for Linear Attention

Query-derived Erase Direction (QED)

Dhruman Gupta, Aritra Das, Debayan Gupta
Truth Audit Labs

arXiv:2608.13668v1 Submitted 2026-08-13

Publication not found as of 2026-10-30.
Public code not found as of 2026-10-30.

Research group discussion, prepared 2026-10-30; not an author presentation.

The gain is retrieval at a fixed state size

Idea

Use the query to select old content for an erase correction, while keeping a single rank-one state update.

- 340M parameters, 15B training tokens, a 2K training window.
- At 8K, S-NIAH-1 improves by 27.2 / 33.6 points across two seeds.
- Maximum gain: 38.6 points at 16K, seed 1.
- Validation loss and SQuAD-completion remain nearly unchanged.

Paper, Secs. 3–6; Tables 2–3, PDF pp. 5, 7. Differences are percentage points.

Many associations share the same recurrent state

$$S_t \in \mathbb{R}^{K \times V}, \qquad o_t = S_t^\top q_t = \sum_{j \le t} (q_t^\top \tilde{k}_j) \tilde{v}_j$$

- More context adds associations without enlarging the state.
- The query determines which stored associations enter the read.
- Existing delta rules select old content through a key-derived erase vector.

Design question

Can the read direction help determine the state correction?

Paper, Secs. 1–3.1, Eq. (5), PDF pp. 1–3. Tildes denote transformed stored pairs.

GDN-2 separates decay, erase, and write

$$\bar{S}_t = \mathrm{Diag}(\exp g_t) S_{t-1}, \qquad a_t = b_t \odot k_t$$

$$S_t = \bar{S}_t + k_t \underbrace{\left[w_t \odot v_t - \bar{S}_t^\top a_t \right]^\top}_{\text{residual to write}}, \qquad o_t = S_t^\top q_t$$

- $q_t, k_t, b_t, g_t \in \mathbb{R}^K$; $v_t, w_t \in \mathbb{R}^V$.
- q_t, k_t are unit vectors; b_t, w_t are channel-wise gates in $(0, 1)$.
- a_t selects old content; k_t is the direction of the state edit.

Paper, Sec. 2 and Sec. 3.2, Eqs. (4), (7)–(8), PDF pp. 2–3.

QED adds a query term to the erase vector

$$
\begin{aligned}
u_t &= b_t \odot q_t, \\
d_t &= u_t - k_t(k_t^\top u_t) = P_{k_t}^{\perp}(b_t \odot q_t), \\
a_t &= b_t \odot k_t + \lambda_h d_t.
\end{aligned}
$$

- Query: selects content associated with the read.
- Gate: controls the correction channel by channel.
- Projection: enforces $d_t^\top k_t = 0$ for a unit key.

The state update still has the form $\bar{S}_t + k_t r_t^\top$.

Paper, Sec. 3.1, Eq. (6), PDF p. 3. Gate first, then project.

Orthogonal probe responses remain unchanged

For any residual r_t and any probe x,

$$\Delta S_t = k_t r_t^\top \quad \Longrightarrow \quad \Delta S_t^\top x = (x^\top k_t) r_t.$$

If $x \perp k_t$, the edit cannot change that response

This holds for both GDN-2 and QED.

With $q_t = \rho_t k_t + q_t^\perp$ and $\rho_t = q_t^\top k_t$,

$$o_t^{\mathrm{GDN2}} = \rho_t \Big[w_t \odot v_t + \bar{S}_t^\top ((\mathbf{1} - b_t) \odot k_t) \Big] + \bar{S}_t^\top q_t^\perp.$$

Paper, Sec. 3.2, Eqs. (9)–(10), PDF p. 4. The last term bypasses direct editing.

QED routes a negative measurement through the key

For a shared incoming state and shared token-level quantities,

$$S_t^{\mathrm{QED}} - S_t^{\mathrm{GDN2}} = -\lambda_h k_t (\bar{S}_t^\top d_t)^\top,$$
$$o_t^{\mathrm{QED}} - o_t^{\mathrm{GDN2}} = -\lambda_h \rho_t \bar{S}_t^\top d_t.$$

1. Measure old content along d_t.
2. Write its negative, scaled value along k_t.
3. Reach the current read through $\rho_t = q_t^\top k_t$.

If $\rho_t = 0$, the current read is unchanged; later reads can still change.

Paper, Eqs. (11)–(12), PDF p. 4. Last sentence: direct implication of the identities.

Projection preserves a specific spectral condition

Let $M_t = I - k_t a_t^\top$. Its eigenvalues are

$$1 \ \ (K-1 \text{ times}), \qquad e_t = 1 - a_t^\top k_t.$$

Because $d_t^\top k_t = 0$,

$$e_t = 1 - \sum_i b_{t,i} k_{t,i}^2 \in (0,1).$$

- The added correction leaves the nontrivial eigenvalue unchanged.
- This statement concerns the erase factor M_t.
- For $D_t = \mathrm{Diag}(\exp g_t)$, the full transition is $M_t D_t$; norm contraction does not follow from this spectrum alone.

Paper, Secs. 3.1 and 6, PDF pp. 3, 7. Last two bullets clarify the scope by algebra.

QED learns one strength per layer and head

$$\lambda_h = \lambda_{\max} \operatorname{sigmoid}(\theta_h), \qquad \lambda_{\max} = 0.25, \qquad \theta_h^{\text{init}} = -2$$

- Initial strength ≈ 0.0298; one scalar per layer and head.
- The strength parameter uses a $10\times$ learning-rate multiplier.
- Same $K \times V$ state; $O(K)$ extra direction construction.
- The rank-one recurrence and GDN-2 kernel structure are retained.

What the cost claim supports

Small additional vector work. The paper gives no measured throughput, latency, or memory benchmark.

Paper, Appendix A, Eq. (13), PDF p. 9. Parameter count: 24 layers $\times$ 6 heads $=$ 144 scalars.

Matched training isolates the recurrence change

Training

340M parameters
24 layers, hidden size 1024
6 heads, head dimension 128
FineWeb-Edu, 15B tokens
Llama-2 tokenizer, 2048 length
Five arms, two seeds each

Evaluation

S-NIAH-1: 1K–32K
500 examples per length
4K–32K are extrapolation

FDA, SWDE, SQuAD-completion
Short-document recall/control

Validation loss in nats

Paper, Sec. 4, PDF pp. 4–5. S-NIAH-2/3 are listed in setup but no scores are reported in v1.

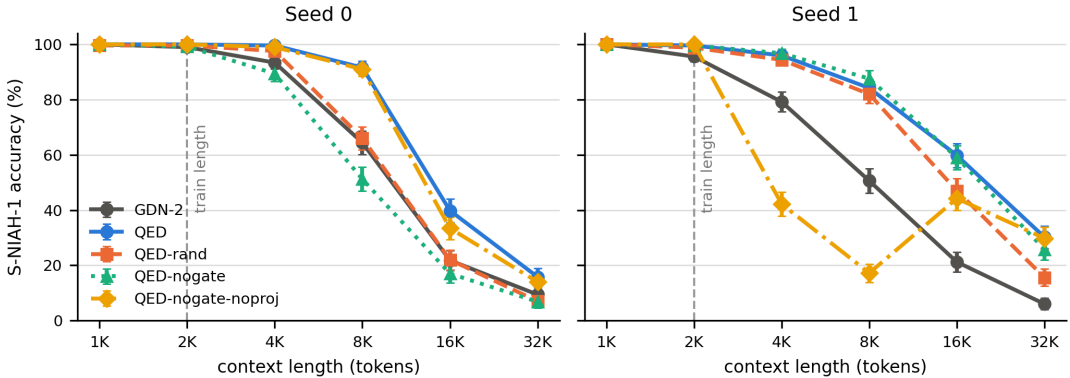
Five arms test direction, gate, and projection

Arm	Added erase term
GDN-2	0
QED-nogate-noproj	$\lambda_h q_t$
QED-nogate	$\lambda_h P^{\perp}_{k_t} q_t$
QED (full)	$\lambda_h P^{\perp}_{k_t}(b_t \odot q_t)$
QED-rand	$\lambda_h P^{\perp}_{k_t}(b_t \odot z_h)$

- z_h: a fixed random unit direction per head.
- "No gate" removes the gate from the correction only.
- Projection is compared with the correction gate absent; the design is not a complete factorial experiment.

Paper, Table 1, PDF p. 4. All arms retain the baseline term $b_t \odot k_t$.

QED improves every reported extrapolation length



Blue vs. gray: QED beats GDN-2 at 4K, 8K, 16K, and 32K in both seeds.

Paper, Fig. 1, PDF p. 6. Vertical line: 2K training length. Error bars: 95% binomial intervals.

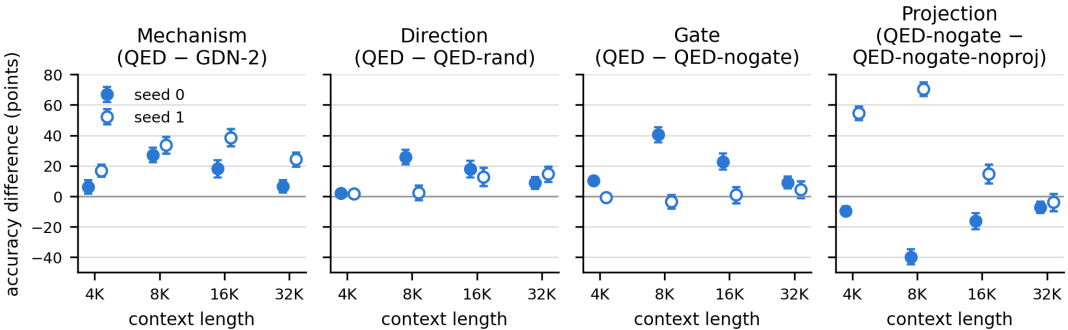
The improvements are large, but accuracy still falls

	Seed 0			Seed 1		
Length	GDN-2	QED	Δ pp	GDN-2	QED	Δ pp
4K	93.4	99.6	+6.2	79.2	96.0	+16.8
8K	64.4	91.6	+27.2	50.6	84.2	+33.6
16K	21.8	39.8	+18.0	21.2	59.8	+38.6
32K	9.4	15.8	+6.4	6.0	30.2	+24.2

- "About 2× usable context" has no stated accuracy threshold.
- Seed 1: QED at 16K (59.8%) exceeds baseline at 8K (50.6%).
- Seed 0: the corresponding comparison is 39.8% vs. 64.4%.

Paper, Table 3 and Sec. 6, PDF pp. 6–7. Scores: accuracy %; gains: percentage points.

Query direction helps; gate and projection vary by seed



- Full minus random is positive at every extrapolation length.
- At 8K, gate effects are $+40.4$ / -3.4 points.
- At 8K, projection effects without the gate are -39.8 / $+70.4$ points.

Paper, Fig. 2 and Table 3, PDF p. 7. Filled/open markers: seeds 0/1; original 95% intervals.

Random directions can help, but query does better

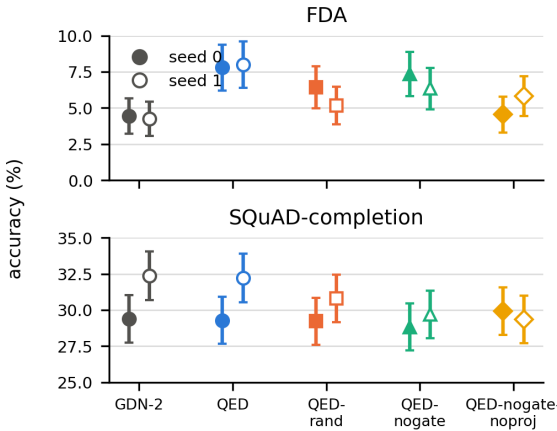
Seed 1 at 16K: S-NIAH-1 accuracy

GDN-2: 21.2% QED-rand: 47.0% QED: 59.8%

- Random-direction control does not always return to baseline.
- The consistent finding is QED's advantage over this control.
- Gate and projection are plausible design choices, but their individual necessity is not established by the two runs.

Paper, Table 3 and Sec. 6, PDF pp. 6–7. Interpretation follows the observed contrasts.

Document recall improves selectively



Full QED minus GDN-2

FDA: +3.35 / +3.73 pp
Final accuracy: 7.80% / 7.99%

SQuAD-completion:
−0.10 / −0.16 pp

SWDE: +2.43 / 0.00 pp

Short documents test recall, not context extrapolation.

Paper, Sec. 5.4, Table 3 and Fig. 3, PDF pp. 6–8. Paired numbers: seeds 0/1.

Language modeling loss is nearly unchanged

Seed	GDN-2	QED	QED minus GDN-2
0	2.3714	2.3719	$+0.0005$
1	2.3746	2.3735	-0.0011

- Validation loss in nats; lower is better.
- Across all five arms and both seeds, the total spread is 0.0034 nats.
- Together with the recall results, this supports an improvement in access to compressed memory.

Paper, Table 2, PDF p. 5; discussion, Sec. 6, PDF p. 6.

Inference strength affects the two seeds differently

$$\lambda_h^{(s)} = s\lambda_h, \qquad \text{all trained weights fixed}$$

S-NIAH-1 at 16K	-1	0	0.5	1	1.5	2
Seed 0	39.8	41.4	42.0	41.8	41.6	43.6
Seed 1	44.8	55.0	61.0	65.6	69.8	73.0

- Seed 1: larger correction helps at all four extrapolation lengths.
- Seed 0: small changes with no consistent direction.
- Scale is applied throughout prefill and decoding, changing the trajectory.

Paper, Appendix A.1, Table 4, PDF pp. 9–10. Scores: accuracy %; $s = 1$ is the trained setting.

The intervention has a narrower interpretation

- $s = 0$ removes QED from a QED-trained model; it is not the independently trained GDN-2 baseline.
- The incoming state also depends on s, so this is not just a final-read intervention.
- Table 4's $s = 1$ differs from Table 3: seed 1 at 16K is 65.6% vs. 59.8%. The paper does not explain the discrepancy.
- A gain at $s = 2$ motivates a training-time strength sweep; it does not establish that $\lambda_{\max} = 0.25$ constrained training.

Paper, Tables 3–4, PDF pp. 7, 10; Appendix A.1, PDF p. 9. Compare strengths within Table 4.

The contribution is query-aware state maintenance

1. **Mechanism:** measure old content using the query; write a correction along the key, keeping a rank-one edit.
2. **Evidence:** better S-NIAH-1 extrapolation and FDA recall in two matched 340M runs, with nearly unchanged LM loss.
3. **Open issue:** component effects and inference-time reliance vary across seeds.

Useful next experiments

More seeds; training-time $\lambda_{\max}$ sweeps; gated/no-projection control; larger scales and measured kernel overhead.

Paper, Secs. 6–7, PDF pp. 6–8. The missing control and kernel benchmark are discussion suggestions.

References and paper status

Primary reading

D. Gupta, A. Das, D. Gupta.
The Query Knows What to Forget: A Second Erase Direction for Linear Attention.
arXiv:2608.13668v1, 2026.

Status checked 2026-10-30

Formal publication not found; public code not found.
Current arXiv record.

Task background

S. Arora et al., *Simple linear attention language models balance the recall-throughput tradeoff.*
arXiv:2402.18668v2: FDA, SWDE, and SQuAD task definitions.

All experimental numbers in this deck are from QED v1. Original figure numbers are retained.

Backup: the progression to GDN-2

Model	Decay	Erase / write
DeltaNet	None	Shared scalar β_t
GDN	Scalar α_t	Shared scalar β_t
KDA	Channel-wise	Shared scalar β_t
GDN-2	Channel-wise	Separate b_t and w_t
QED	Channel-wise	Query term in erase vector

$$S_t^{\mathrm{DeltaNet}} = (I - \beta_t k_t k_t^\top) S_{t-1} + \beta_t k_t v_t^\top$$

- GDN-2 is the directly trained baseline in this study.
- EDA is discussed as prior work with separate erase/write addresses.

Paper, Sec. 2, Eqs. (1)–(4), PDF pp. 2–3.

Backup: a two-dimensional example

$$k = \begin{pmatrix}1\\0\end{pmatrix}, \quad q = \begin{pmatrix}0.6\\0.8\end{pmatrix}, \quad b = \begin{pmatrix}0.5\\0.5\end{pmatrix}, \quad \bar{S} = \begin{pmatrix}2\\3\end{pmatrix}, \quad wv = 1, \quad \lambda = 0.2$$

$$r^{\mathrm{GDN2}} = 1 - \bar{S}^\top (b \odot k) = 0, \qquad o^{\mathrm{GDN2}} = 3.6,$$

$$d = (0, 0.4)^\top, \qquad \bar{S}^\top d = 1.2,$$

$$\Delta S = (-0.24, 0)^\top, \qquad \Delta o = -0.144.$$

The second state coordinate stays at 3

Its content determines a correction to the first coordinate. The new output is 3.456.

Teaching example derived from Eqs. (6), (11)–(12); not an experiment from the paper.

Backup: eigenvalue preservation is not norm contraction

Choose $k = (1, 0)^\top$, $q = (0, 1)^\top$, $b = (0.5, 0.5)^\top$, and $\lambda = 0.2$. Then

$$a = (0.5, 0.1)^\top, \qquad M = I - ka^\top = \begin{pmatrix} 0.5 & -0.1 \\ 0 & 1 \end{pmatrix}.$$

$$\operatorname{spec}(M) = \{0.5, 1\}, \qquad \left\| M \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\|_2 = \sqrt{1.01} > 1.$$

The projection preserves $1 - a^\top k$, but a non-normal erase factor can amplify some vectors. Channel-wise decay adds another matrix factor.

Independent algebraic clarification of the claim in Sec. 3.1; not a measured model instability.

Backup: interpreting the projection diagnostic

- Without projection, the paper gives the possible range $e = 1 - a^\top k \in [-0.25, 1.25]$.
- In 1.18M token-head samples from two unprojected models, the observed range is $[-0.011, 0.985]$.
- The paper also says counterfactual projection changes e by less than 10^{-3}, without specifying the statistic.

A numerical point to clarify

A uniform per-sample change below 10^{-3} would conflict with a minimum of -0.011 and projected $e \in (0, 1)$. An average change would not imply that uniform bound.

Paper, Sec. 6, PDF p. 7. The consistency observation is this reading's analysis.

Backup: optimization and intervention details

- AdamW, learning rate 3×10^{-4}; WSD decay begins at 13.5B of 15B training tokens.
- Global batch: 524,288 tokens; BF16 with FP32 master weights.
- RMSNorm, SwiGLU, tied embeddings.
- Strength sweep: same 500 deterministic examples per length for every s, with an exact McNemar test against $s = 1$ specified.
- Paired outcome counts and p-values are not reported; marginal accuracies alone cannot reconstruct that test.

Paper, Sec. 4.2 and Appendix A.1, PDF pp. 5, 9–10.