Q-Delta: Beyond Key–Value Associative State Evolution

Reading discussion: query feedback in recurrent memory

Paper authors: Sumin Park, Seojin Kim, Noseong Park
KAIST

arXiv:2606.08804v1 | Reading: October 8, 2026

Use the query to help decide what the state should remember.

A reading presentation, not an author presentation. Paper: ICML 2026; arXiv v1, June 7, 2026.

The readout direction can shape future memory

The gap

Conventional delta updates correct a key-based prediction.
The query reads the result but does not enter that correction.

$$\underbrace{S_{t-1}k_t}_{\text{key prediction}} \qquad \underbrace{S_{t-1}q_t}_{\text{query prediction}}$$

- Both are value-space projections of the same memory.
- Q-Delta uses both to form the write residual.
- The recurrent state keeps the same dimensions.

Paper, Secs. 1 and 3, PDF pp. 1–5; Appendix B, p. 13.

Delta rules correct an existing key–value association

Additive linear attention

$$S_t = S_{t-1} + v_t k_t^\top, \qquad o_t = S_t q_t$$

DeltaNet: write a prediction residual

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

Gated DeltaNet: decay, then correct

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

The query is absent from every write residual above.

Paper, Sec. 2.1, Eqs. (1)–(3), PDF p. 2. Single head: $S \in \mathbb{R}^{d_v \times d_k}$.

A query reads another mixture of past values

For $S_t = S_{t-1}P_t + \eta_t v_t k_t^\top$ and $S_0 = 0$:

$$S_{t-1} = \sum_{\tau<t} v_\tau b_{\tau,t-1}^\top, \qquad b_{\tau,t} = P_t^\top b_{\tau,t-1}.$$

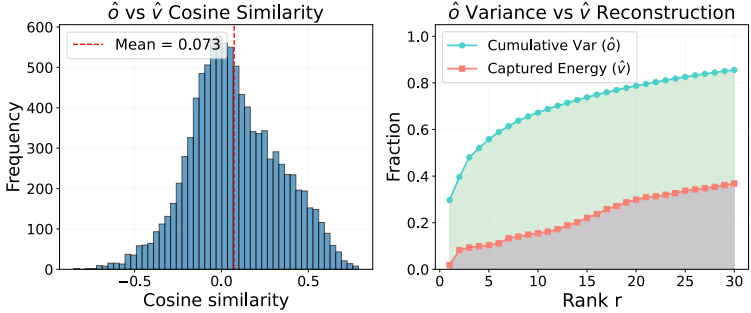
Query-conditioned value aggregation

$$\hat{o}_t = S_{t-1}q_t = \sum_{\tau<t} (b_{\tau,t-1}^\top q_t) v_\tau$$

- State transitions reshape the effective historical keys.
- Weights are unnormalized and may be signed.
- The identity holds for any probe; q_t is the actual readout direction.

Paper, Sec. 3.1, Eqs. (8)–(15), PDF pp. 3–4; proof in Appendix C.1, pp. 13–14.

The two readouts emphasize different value directions



- Mean cosine similarity: **0.073**; 340M model, layers 5/10/15.
- Principal directions of $\hat{o}$ capture much less energy of $\hat{v}$.

Paper, Fig. 2, PDF p. 5; Sec. 3.2, pp. 4–5. Author figure; 10,000 sampled timesteps.

Q-Delta adds query feedback to the write residual

$$\hat{v}_t = S_{t-1}k_t, \qquad \hat{o}_t = S_{t-1}q_t$$

Without state decay

$$S_t = S_{t-1} + \beta_t\big(v_t - \hat{v}_t - \lambda_t\hat{o}_t\big)k_t^\top$$

$$x_t = k_t + \lambda_t q_t \quad \Longrightarrow \quad S_t = S_{t-1}(I - \beta_t x_t k_t^\top) + \beta_t v_t k_t^\top$$

- Predict along x_t; write along k_t; output along q_t.
- This is not a gradient step on $\frac{1}{2}\|v_t - Sx_t\|^2$: that gradient would write along x_t.

Paper, Eqs. (17)–(18), PDF p. 5; Sec. 3.3.3, p. 6.

The deployed rule first decays the old state

$$\bar{S}_{t-1} = \alpha_t S_{t-1}, \qquad x_t = k_t + \lambda_t q_t$$

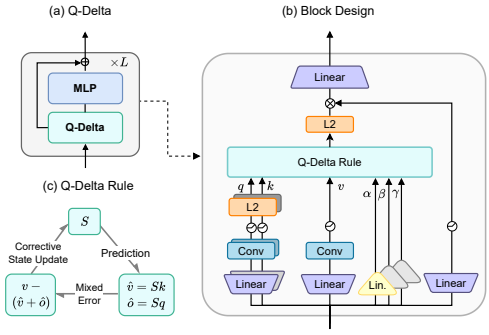
One recurrent step

$$S_t = \bar{S}_{t-1} + \beta_t \big(v_t - \bar{S}_{t-1} x_t\big) k_t^\top, \qquad o_t = S_t q_t$$

Gate	Role	Special case
α_t	Retain the previous state	$\alpha_t = 1$: no decay
β_t	Scale the write residual	$\beta_t = 0$: no write
λ_t	Scale query feedback	$\lambda_t = 0$: Gated DeltaNet

Paper, Eqs. (19)–(20), PDF p. 5. The fresh value term is not multiplied by α_t.

One extra projection controls each head's feedback



$$\lambda_t = \sigma(W_\lambda h_t + b)$$

$$W_\lambda \in \mathbb{R}^{H \times d_{\text{model}}}$$

- Per token, per head.
- Scalar bias initialized to -0.8.
- No extra recurrent state.
- Same q, k, v transformations as the underlying block.

Paper, Fig. 1, PDF p. 2; Appendix B, p. 13. Author architecture figure.

One-step contraction depends on directional alignment

Lemma 3.1 assumes $\alpha_t = 1$. Define

$$a_t = k_t^\top x_t, \qquad \widetilde{e}_t = v_t - S_{t-1}x_t, \qquad e_t = v_t - S_t x_t.$$

Right-multiplying the update by x_t gives the exact identity:

The central algebraic fact

$$e_t = (1 - \beta_t a_t)\widetilde{e}_t$$

- $0 < \beta_t a_t < 2$ implies contraction of this mixed error.
- It does not establish contraction of the entire state norm.
- A uniform rate requires $\sup_t |1 - \beta_t a_t| < 1$.

Paper, Lemma 3.1, PDF p. 6; Appendix C.2 proof, p. 14.

Global tracking also requires bounded residual drift

$$r_t = \underbrace{v_t - v_{t-1}}_{\text{target drift}} - \underbrace{S_{t-1}(x_t - x_{t-1})}_{\text{prediction drift}}$$

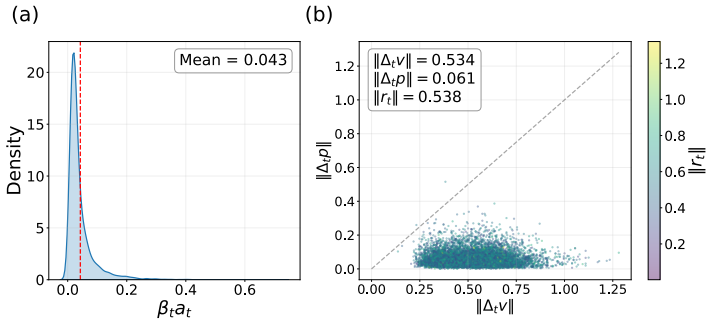
Since $\widetilde{e}_t = e_{t-1} + r_t$, if $\|r_t\| \le r$ and $\rho < 1$:

$$\boxed{\|e_t\| \le \rho^t \|e_0\| + \frac{\rho(1-\rho^t)}{1-\rho}\, r}$$

- Initial error fades; persistent drift sets a nonzero error radius.
- Pointwise contraction alone does not supply a uniform margin.
- With $\alpha_t < 1$, an extra term $(1-\alpha_t)S_{t-1}x_t$ enters tracking (backup slide).

Paper, Theorem 3.2, PDF pp. 6–7; proof pp. 14–16. Decay extension: reader derivation.

Measured dynamics fall within the contraction regime



- Mean $\beta_t a_t$: **0.043**; mean residual drift norm: **0.538**.
- Prediction drift is smaller than target drift in these samples.

Paper, Fig. 3, PDF p. 6; discussion p. 7. Sample means do not specify worst-case ρ or r.

Chunking turns token dependencies into triangular solves

Stack X, K by rows. The interaction is

$$(XK^\top)_{ri} = x_r^\top k_i, \qquad x_r = k_r + \lambda_r q_r.$$

With $B = \mathrm{diag}(\beta)$ and causal decay ratios H_{ri}:

$$L = \mathrm{sl}(BXK^\top), \qquad L_\gamma = \mathrm{sl}\big(B(H \odot XK^\top)\big)$$

$$W = (I + L)^{-1}BX, \qquad \widetilde{U} = (I + L_\gamma)^{-1}BV$$

- Chunk-local matrix operations; recurrent state passed between chunks.
- Same state size; linear scaling in length for fixed chunk/head sizes.
- Appendix D.2 prints $KX^\top$; the recurrence requires $XK^\top$.

Paper, Eqs. (23)–(27), PDF pp. 5–6; Appendix D, pp. 16–18. sl: strict lower triangle.

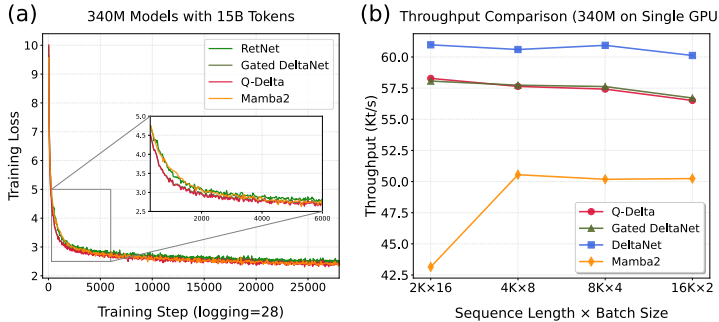
Comparisons match data and optimization settings

Model size	FineWeb-Edu tokens	Peak learning rate
340M	15B	10^{-3}
1.3B	30B	4×10^{-4}

- FLA framework; 4 RTX Pro 6000 Blackwell GPUs; bfloat16.
- AdamW, cosine schedule, gradient clipping; baselines retrained.
- Baselines: RetNet, Mamba, Mamba2, DeltaNet, Gated DeltaNet.
- Language: both scales. Synthetic retrieval: 1.3B at 1K–4K.
- Real-world recall: 340M, inputs capped at 2K, cloze prompts.

Paper, Secs. 4.1–4.2, PDF pp. 7–8; Appendix A, p. 13.

Training throughput stays close to Gated DeltaNet



- 340M; single GPU; 50 training steps per throughput setting.
- Fixed 32,768 tokens/step; lengths 2K–16K. DeltaNet is faster.

Paper, Fig. 4, PDF p. 7. Left: loss over 28,600 steps. Right: training throughput, not decoding.

Average zero-shot accuracy improves at both scales

Model	340M / 15B tokens	1.3B / 30B tokens
RetNet	44.55	50.31
Mamba	46.64	52.39
Mamba2	46.27	52.46
DeltaNet	43.82	52.53
Gated DeltaNet	46.01	52.77
Q-Delta	**47.24**	**53.47**

- Over Gated DeltaNet: +**1.23 pp** at 340M, +**0.70 pp** at 1.3B.
- At 340M, the gain over the best other average is +0.60 pp.

Paper, Table 2, PDF p. 8. Unweighted mean of 8 accuracy tasks; higher is better.

Perplexity gains depend on scale and comparison

Scale	Model	LAMBADA ppl. ↓	WikiText ppl. ↓
340M	Mamba2	**30.35**	**26.60**
	Gated DeltaNet	36.27	27.82
	Q-Delta	32.67	26.89
1.3B	Mamba2	17.40	19.47
	Gated DeltaNet	15.32	19.61
	Q-Delta	**15.19**	**19.21**

- Q-Delta improves both perplexities over Gated DeltaNet.
- Mamba2 leads at 340M; Q-Delta leads all listed baselines at 1.3B.

Paper, Table 2, PDF p. 8. Selected rows; the accuracy average excludes perplexity.

The largest synthetic retrieval gains occur at 4K

Model	4K accuracy (%) Pass-key	Number	UUID	Mean over all 9 settings
Mamba2	99.0	57.0	11.6	76.58
DeltaNet	100.0	49.6	25.4	81.29
Gated DeltaNet	100.0	76.6	21.4	83.51
Q-Delta	**100.0**	**94.2**	**48.0**	**90.02**

Relative to Gated DeltaNet, in percentage points

Number at 4K: $+$**17.6** UUID at 4K: $+$**26.6**
Mean over 3 tasks $\times$ 3 lengths: $+$**6.51**

Paper, Table 3, PDF p. 8. 1.3B models; RULER S-NIAH at 1K, 2K, and 4K.

UUID retrieval still degrades sharply with length

Context	DeltaNet	Gated DeltaNet	Q-Delta
1K	87.4	83.8	**94.6**
2K	**75.8**	70.0	74.0
4K	25.4	21.4	**48.0**

- At 2K, DeltaNet retains the best UUID result.
- At 4K, Q-Delta improves substantially but succeeds only 48% of the time.
- Quality is evaluated up to 4K; the 16K experiment measures throughput.

Paper, Table 3, PDF p. 8, accuracy in percent; throughput settings in Fig. 4, p. 7.

Real-world recall shows gains and regressions

Task	Gated DeltaNet	Q-Delta	Difference
SWDE	29.1	31.6	+2.5
SQuAD	56.0	52.3	−3.7
FDA	18.9	22.0	+3.1
TriviaQA	48.9	48.6	−0.3
NQ	18.2	18.4	+0.2
DROP	34.2	33.2	−1.0

Reader-computed unweighted means: 34.22 vs. 34.35.

The average advantage is only 0.13 score points.

Paper, Table 4, PDF p. 8. 340M, 2K cap, cloze prompts. Original table scores; mean not printed there.

Fixed feedback captures most of the accuracy gain

340M configuration	Wiki ppl. ↓	Lamb. ppl. ↓	Avg. acc. ↑
Learnable λ_t	26.89	**32.67**	**47.24**
Fixed $\lambda = 0.2$	26.96	35.39	46.99
Fixed $\lambda = 0.5$	26.86	33.31	47.20
Fixed $\lambda = 0.8$	26.61	33.58	46.42
No decay: $\alpha_t = 1$	**26.52**	32.97	45.86
Full feedback: $\lambda_t = 1$	26.55	35.21	46.36

- Learnable vs. fixed 0.5: only **0.04 pp** in mean accuracy.
- No-decay Q-Delta (45.86) still exceeds DeltaNet (43.82).

Paper, Table 5, PDF p. 9; DeltaNet from Table 2, p. 8. "No gating" means $\lambda = 1$, not 0.

Query feedback is a useful, compact memory modification

1. **Design:** separate the mixed prediction direction, the write address, and the output query.
2. **Evidence:** better average language scores and stronger 4K synthetic retrieval, at training throughput close to Gated DeltaNet.
3. **Trade-offs:** real-world recall is mixed; adaptive feedback barely improves over a well-chosen fixed coefficient.
4. **Theory:** conditional mixed-error tracking; the decay gate needs an additional drift argument.

Synthesis of Sec. 3.3, Tables 2–5, and Fig. 4; PDF pp. 5–9.

Paper and directions for discussion

Sumin Park, Seojin Kim, Noseong Park.
Q-Delta: Beyond Key–Value Associative State Evolution.
ICML 2026; arXiv:2606.08804v1, June 7, 2026.

arxiv.org/abs/2606.08804v1
Author-listed code: github.com/psmiz/Q-Delta

- Extend tracking theory to the actual decay-gated recurrence.
- Evaluate retrieval beyond 4K and measure recurrent decoding cost.
- Identify tasks where token-dependent λ_t matters.

All empirical figures and results in this deck come from v1. Supplemental algebra is labeled as reader derivation.

Backup: a small example separates prediction and write

Reader-created example; $\alpha = 1$**:**

$$S_{t-1} = [1\ 2], \quad k = [1\ 0]^\top, \quad q = [0\ 1]^\top, \quad v = 3, \quad \lambda = \beta = \tfrac{1}{2}.$$

$$x = [1\ \tfrac{1}{2}]^\top, \qquad \widetilde{e} = 3 - S_{t-1}x = 1$$

$$S_t = [1\ 2] + \tfrac{1}{2} \cdot 1 \cdot [1\ 0] = [1.5\ 2]$$

- Mixed error becomes $3 - S_t x = 0.5$: contraction factor $1 - \beta k^\top x = 0.5$.
- The immediate output remains $S_t q = 2$ because $k^\top q = 0$.
- The memory has changed and can affect later queries.

Teaching example from Eq. (17), PDF p. 5, and Lemma 3.1, p. 6; not a paper experiment.

Backup: a consistent chunk output formula

Row stacks: $Q, K, W \in \mathbb{R}^{C\times d_k}$, $\widetilde{U} \in \mathbb{R}^{C\times d_v}$, $S_{\text{in}} \in \mathbb{R}^{d_v\times d_k}$.

$$\gamma_i = \prod_{j\le i} \alpha_j, \quad D_\gamma = \mathrm{diag}(\gamma), \quad H_{ij} = 1_{\{j\le i\}}\, \gamma_i/\gamma_j$$

$$Z = \widetilde{U} - D_\gamma W S_{\text{in}}^\top, \qquad K_{\text{out}} = \mathrm{diag}(\gamma_C/\gamma_i) K$$

$$S_{\text{out}} = \gamma_C S_{\text{in}} + Z^\top K_{\text{out}}$$

$$O = D_\gamma Q S_{\text{in}}^\top + (H \odot QK^\top) Z$$

In v1 Eq. (26): the entrance-state product needs $S^\top$ under row stacking, and intra-chunk output needs the decay ratios.

Reader derivation from Eq. (25), PDF p. 5. Compare Eq. (26), p. 6; WY proof, pp. 16–17.

Backup: decay changes the cross-time tracking residual

For the actual gated recurrence:

$$e_t = (1 - \beta_t a_t)\big(v_t - \alpha_t S_{t-1} x_t\big)$$

Using the theorem's original e_{t-1} and r_t:

$$v_t - \alpha_t S_{t-1} x_t = e_{t-1} + r_t + \underbrace{(1 - \alpha_t) S_{t-1} x_t}_{\text{additional drift}}.$$

- A gated tracking bound must control this term too.
- For unit-norm q, k, $a_t \in [1 - \lambda_t, 1 + \lambda_t]$.
- This gives pointwise contraction for $0 < \lambda_t < 1$ and $0 < \beta_t \leq 1$, but not a uniform margin away from 0 and 2.

Reader derivation from Eq. (19), PDF p. 5; normalization in Fig. 1, p. 2; theorem pp. 6, 14–16.

Backup: zero-shot accuracy has task-level trade-offs

	340M		1.3B	
Task	GDN	Q-Delta	GDN	Q-Delta
ARC-Easy	60.02	59.51	68.60	68.27
ARC-Challenge	25.94	28.50	33.28	36.60
HellaSwag	40.25	41.61	52.60	53.46
LAMBADA	31.52	33.90	43.80	43.28
PIQA	67.30	67.63	70.84	71.44
WinoGrande	51.54	52.88	54.78	54.93
BoolQ	57.13	59.48	59.42	61.41
OpenBookQA	34.40	34.40	38.80	38.40

Paper, Table 2, PDF p. 8. Accuracy in percent; GDN = Gated DeltaNet.

Backup: a surrogate objective explains the update

A mixed squared error has gradient

$$\ell(S) = \tfrac{1}{2}\|v_t - Sx_t\|^2, \qquad -\nabla_S \ell = (v_t - Sx_t)x_t^\top.$$

Q-Delta instead writes along k_t. Fix the previous state and set $\delta_t = v_t - \alpha_t S_{t-1} x_t$. Then

$$J_t(S) = \|S - \alpha_t S_{t-1}\|_F^2 - 2\beta_t \langle Sk_t, \delta_t \rangle$$

has the minimizer

$$\arg\min_S J_t(S) = \alpha_t S_{t-1} + \beta_t \delta_t k_t^\top.$$

A per-step surrogate is compatible with the update; direct gradient descent on the mixed squared loss is a different rule.

Paper, Table 1, PDF p. 3; Sec. 3.3.3, p. 6. Algebra of the surrogate expanded for teaching.