Erase-then-Delta Attention

Decoupling Erase and Write Addresses in Delta-Rule Linear Attention

Paper: Xiao Li et al.
Qwen Team, Nanjing University, Zhejiang University

arXiv:2606.26560v1 (25 June 2026)
Independent reading presentation — 22 September 2026

Erase elsewhere, then correct the current write

Main idea

Give recurrent memory an independent cleanup address before its usual delta-rule write.

- Reported average gains over KDA: +0.63 pp (dense), +1.07 pp (MoE) after pretraining.
- After long-context midtraining: +0.14 pp on the general suite.
- A conditional cleanup path; historical recall is not uniformly better.

Paper, Eq. (8), p. 4; Tables 1–2, p. 8; Sec. 6, p. 11.

Finite-state memory must manage interference

- Linear recurrence compresses history into a fixed-size state.
- Additive writes accumulate associations; delta writes correct the current key.
- A stale association may occupy a direction different from the new write key.
- Motivating examples: reassignment, fact correction, context shifts. These are not separately evaluated tasks here.

Paper, Secs. 1–2, pp. 1–3.

The delta rule corrects what the current key reads

$$S_t \in \mathbb{R}^{d_k \times d_v}, \quad k_t, q_t \in \mathbb{R}^{d_k}, \quad v_t \in \mathbb{R}^{d_v}, \quad o_t = S_t^\top q_t$$

$$\mathcal{L}_t(S) = \tfrac{1}{2}\|S^\top k_t - v_t\|^2$$

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

- $D_t = \mathrm{Diag}(\alpha_t)$ retains key channels.
- β_t controls correction at the same address used for writing.
- The state update is an online-gradient interpretation, not a separate model-parameter optimizer.

Paper, Eqs. (2), (5), (7), pp. 2–3.

Gate separation and address separation differ

Variant	Memory control
GDN	Scalar retention within each head; delta correction at k.
KDA	Channel-wise retention; delta correction at k.
GDN-2	Key-side erase and value-side write gates; edit still committed along k.
EDA	Independently generated e; erase along e, then correct and write along k.

Paper, Sec. 2.2, p. 3; Sec. 5, pp. 10–11.

EDA preserves the delta write after cleanup

$$\boxed{S_t = (I - \beta_t k_t k_t^\top)(I - \gamma_t e_t e_t^\top) D_t S_{t-1} + \beta_t k_t v_t^\top}$$

$$\widehat{S}_t = D_t S_{t-1} \qquad \text{retain coordinates}$$

$$\widetilde{S}_t = \widehat{S}_t - \gamma_t e_t (\widehat{S}_t^\top e_t)^\top \qquad \text{erase a direction}$$

$$S_t = \widetilde{S}_t + \beta_t k_t (v_t - \widetilde{S}_t^\top k_t)^\top \qquad \text{correct the write}$$

Unit-norm e_t, k_t; scalar gates $\gamma_t, \beta_t \in (0, 1)$.

Paper, Eqs. (8)–(13), p. 4.

Directional erasure suppresses a memory readout

$$\mathcal{L}_t^{\mathrm{erase}}(\widehat{S}_t) = \tfrac{1}{2}\|\widehat{S}_t^\top e_t\|^2$$

$$\widetilde{S}_t^\top q = \widehat{S}_t^\top q - \gamma_t (q^\top e_t)\widehat{S}_t^\top e_t$$

- At $q = e_t$, the readout is multiplied by $1 - \gamma_t$.
- At $q \perp e_t$, this erase step leaves the readout unchanged.
- Nonorthogonal queries can still change; this is not deletion of a symbolic database entry.

Paper, Eqs. (10)–(12), p. 4.

A two-dimensional example isolates the new capability

$$S_{t-1} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \quad e = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad k = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad v = 7$$

Take $D = I$ and the limiting gates $\beta = \gamma = 1$.

$$\text{EDA:} \quad \begin{bmatrix} 3 \\ 5 \end{bmatrix} \xrightarrow{\text{erase}} \begin{bmatrix} 0 \\ 5 \end{bmatrix} \xrightarrow{\text{delta}} \begin{bmatrix} 0 \\ 7 \end{bmatrix}$$

$$\text{Baseline:} \quad \begin{bmatrix} 3 \\ 5 \end{bmatrix} \xrightarrow{\text{delta}} \begin{bmatrix} 3 \\ 7 \end{bmatrix}$$

Expressivity illustration, not an experiment or a semantic deletion guarantee.

Teaching example derived from Paper, Eq. (8), p. 4.

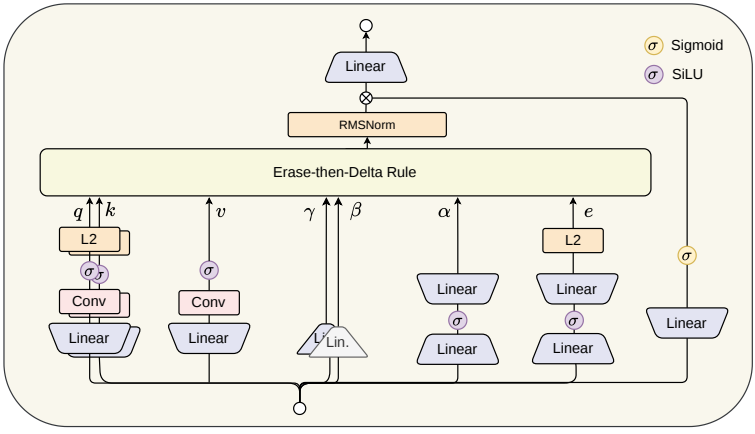
Order matters when the two addresses overlap

$$
\begin{aligned}
&(I - \beta k k^\top)(I - \gamma e e^\top) \\
&\quad = I - \gamma e e^\top - \beta k k^\top + \beta\gamma (k^\top e) k e^\top
\end{aligned}
$$

- The cross-term depends on address overlap.
- Erasing after the write can also suppress newly written content.
- Diagonal decay and directional erasure generally do not commute.
- Reader derivation: each old-state factor is nonexpansive under normalized addresses and gates in $[0, 1]$; this is not a full training-stability proof.

Paper, Eq. (16), pp. 5–6; final bullet: reader analysis.

The extra address is generated inside the layer



Low-rank erase projection; normalized q, k, e; output normalization and gating.

Paper, Fig. 1, p. 5; Appendix A, p. 13.

EDA also changes the safe decay parameterization

$$A = e^{A_{\log}}, \quad \Delta_t = \mathrm{softplus}(a_t + b_\Delta)$$

$$g_t = \ell + (-\ell)\exp\left(-\frac{A}{|\ell|} \odot \Delta_t\right), \qquad \alpha_t = e^{g_t}$$

- With $\ell = -5$: $e^{-5} < \alpha_{t,i} \leq 1$; near zero, $g \approx -A\Delta$.
- Large decay inputs saturate instead of sending log-retention to $-\infty$.
- Important attribution limit: reported EDA/KDA comparisons also differ in safe gate and parameter count.

Paper, Eqs. (14)–(15), p. 5; Appendix A, pp. 13–14.

Two virtual steps recover a standard delta recurrence

Sub-step	q'	k'	v'	β'	α'
Erase $(2t-1)$	0	e_t	0	γ_t	α_t
Delta $(2t)$	q_t	k_t	v_t	β_t	1

- Odd steps decay and erase without writing a value.
- Even steps correct/write and produce the real output.
- WY / UT chunk formulations aim to turn recurrence into matrix operations.
- Twice as many sub-steps increases prefill work, while state size is unchanged.

Paper, Sec. 3.4, Eqs. (17)–(22), pp. 6–7.

Efficiency is argued, not quantitatively benchmarked

- New kernel inputs: erase address e and scalar gate γ.
- Authors argue fusion keeps extra HBM traffic modest and decoding remains state-I/O dominated.
- No latency, throughput, hardware sweep, or measured overhead table is provided.
- Fixed-size recurrent memory does not imply fixed memory for the whole 3:1 hybrid model.

Paper, Sec. 3.5 and Sec. 4.1, p. 7; last bullet: architectural implication.

Evaluation uses matched training within two model families

	Dense 2.5B	MoE 25B-A2.8B
Layers (linear/full)	24 (18/6)	28 (21/7)
Model width	2048	2048
Pretraining tokens	400B	400B
Sequence / global batch	4096 / 1024	4096 / 1024
Extra midtraining	–	80B at 32k

- All recurrent variants: three linear layers, then one full-attention layer.
- Dense includes Transformer, GDN, GDN-2, KDA, EDA; MoE includes GDN, KDA, EDA.
- BF16, AdamW; parameter counts are close but not identical.

Paper, Sec. 4.1, p. 7; Tables 4–5, pp. 13–14.

Metrics mix knowledge, reasoning, and code

- MMLU / MMLU-Pro: broad knowledge and harder multiple-choice questions.
- GSM8K / MATH: mathematical reasoning; BBH: difficult reasoning tasks.
- EvalPlus: code generation under stricter functional tests.
- Tables 1–2: percentages, two evaluations of the same checkpoint; Avg. is the unweighted six-task mean.
- No independent-training-seed variance or confidence intervals are reported.

Paper, Sec. 4.1, p. 7; Appendix B, p. 14.

Dense EDA leads the mean, not every benchmark

Model	MMLU	GSM8K	EvalPlus	Six-task Avg.
Transformer	50.11	20.26	30.82	27.75
GDN	49.99	20.83	31.08	27.58
GDN-2	49.90	21.40	32.95	28.14
KDA	50.03	21.30	30.73	27.81
EDA	49.27	23.90	32.50	**28.44**

- EDA vs. KDA: +0.63 pp overall; +2.60 pp GSM8K, but −0.76 pp MMLU.
- Only selected task columns shown; Avg. still uses all six tasks.

Paper, Table 1, p. 8. Percentages; higher is better; 400B tokens.

MoE EDA leads five of six pretraining tasks

Model	GSM8K	MATH	EvalPlus	Six-task Avg.
GDN	58.89	31.87	50.09	48.16
KDA	58.61	30.78	51.60	48.31
EDA	57.71	33.57	53.37	**49.38**

- EDA vs. KDA: +1.07 pp overall; MATH +2.79 pp, EvalPlus +1.77 pp.
- GSM8K is the exception: EDA trails both baselines.
- Evidence across two settings, not a fitted scaling law.

Paper, Table 1, p. 8. Percentages; higher is better; 400B tokens.

Midtraining retains only a small mean advantage

Model	GSM8K	MATH	EvalPlus	Six-task Avg.
GDN	75.93	45.94	50.09	57.37
KDA	76.21	46.87	53.82	58.31
EDA	75.99	49.28	51.78	**58.45**

- EDA vs. KDA: +0.14 pp overall, down from +1.07 pp before midtraining.
- MATH +2.41 pp is partly offset by EvalPlus −2.04 pp.
- Reported means do not establish statistical significance.

Paper, Table 2, p. 8. MoE: 400B pretraining + 80B tokens at 32k.

RULER gains are strongest at 4k–16k

Model	4k	8k	16k	32k	64k	128k	Avg.
GDN	92.40	90.09	87.28	82.15	67.62	45.16	77.45
KDA	93.33	89.29	85.13	79.66	72.00	42.70	77.02
EDA	93.84	90.88	87.55	81.52	71.90	44.22	78.32

- EDA has the highest six-length average, not the best score at every length.
- 64k / 128k extrapolate beyond 32k midtraining; all models degrade substantially.
- Four evaluation runs; averages across subtasks and displayed lengths.

Paper, Table 3, p. 9. Percentages; higher is better.

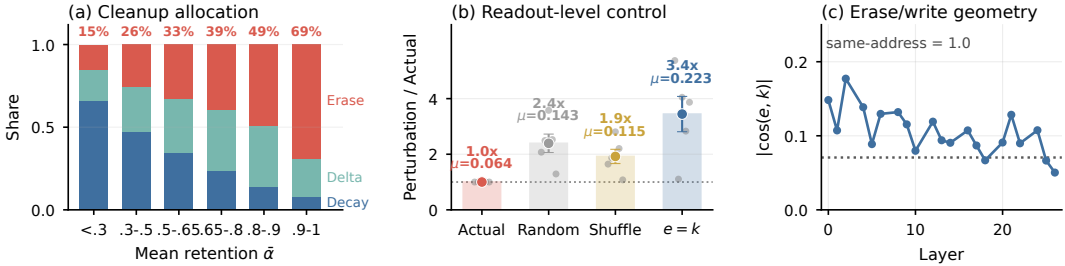
The erase path receives more budget when decay is weak

$$b_D = 1 - \bar{\alpha}, \quad b_\Delta = \bar{\alpha}\beta, \quad b_E = \bar{\alpha}\gamma, \quad q_m = \frac{b_m}{b_D + b_\Delta + b_E}$$

- Global gate shares: decay 35.4%, delta 31.8%, erase 32.8%.
- For $\bar{\alpha} \geq 0.9$, erase share is 69.1%, about $3\times$ delta.
- These are gate-strength shares, not fractions of removed state energy.
- State projections, anisotropy, and address overlap are not accounted for.

Paper, Sec. 4.5, Eqs. (23)–(24), pp. 9–10; Fig. 2(a).

Learned addresses perturb current reads less



Actual / random / shuffled / same-key raw perturbation means:
0.064 / 0.143 / 0.115 / 0.223.

Layer-normalized ratios in panel (b) are not ratios of raw means.

Paper, Fig. 2, p. 10; same erase gates, counterfactual direction replay.

The probes support decoupling, not causal task benefit

- Mean $|\cos(e, k)| \approx 0.105$: far from same-address collapse at 1.
- Independent random directions in 128D already have mean $|\cos| \approx 0.0707$.
- Low perturbation is not a direct measurement of deleting stale facts or improving predictions.
- KDA performs better on raw write-key recall; the paper gives no numerical recall curve.

Paper, Sec. 4.5, pp. 9–10; Sec. 6, p. 11. Random-reference value: reader calculation.

A useful update rule with clear evidence boundaries

- Contribution: independently addressed cleanup while retaining corrective delta writing.
- Evidence: better reported task averages in the tested hybrid settings.
- Limits: different safe gate, small parameter increases, no seed uncertainty, no timed kernels.
- Next tests: matched-gate ablations, controlled invalidation tasks, causal interventions, and recurrence/kernel equivalence.

Paper, Sec. 6, p. 11; limitations and next tests include reader analysis.

Primary source and discussion route

Xiao Li et al.
Erase-then-Delta Attention: Decoupling Erase and Write Addresses in Delta-Rule Linear Attention.
arXiv:2606.26560v1, 25 June 2026.

- Read Eq. (8) for the update, Table 3 for long-context boundaries, and Sec. 4.5 for probe interpretation.
- This presentation uses the same-version author TeX, PDF, and original figures.
- Discussion: when should finite memory sacrifice raw historical recall for current prediction?

Independent reading presentation; no pretraining or kernel performance reproduction.

Backup: printed chunk formulas need implementation checks

Paper Eq. (21) prints

$$M = I + \mathrm{StrictTril}(\cdots), \quad W = M(A \odot K'), \quad U = MV'.$$

- Under ordinary ungated K', V' notation, a single sub-step gives $M = I$ and $U = V'$; Eq. (20) instead requires the β' factor.
- A triangular-solve / inverse convention and gate placement need clarification; matrix layout also needs checking.
- Do not silently repair the paper. Validate an implementation against unambiguous Eqs. (8), (17)–(18).

Reader consistency check against Paper, Eqs. (20)–(22), pp. 6–7; not an author correction.

Backup: parameter differences are small but nonzero

Scale	Model	Total (B)	Active (B)
Dense	KDA	2.5218	2.5218
Dense	EDA	2.5295	2.5295
MoE	KDA	24.6324	2.7885
MoE	EDA	24.6558	2.8119

- Dense increase: 7.7M, approximately 0.31% total parameters.
- MoE increase: 23.4M, approximately 0.84% active parameters.
- MoE erase projection: rank 16 per head; safe log-decay bound -5.

Paper, Table 5, p. 14; Appendix A, p. 13. Percent changes: reader calculations.