OSDN: Online Preconditioning inside the Delta Rule

A paper reading: mechanism, guarantees, and evidence

Paper: Chenyu Zhou, Hongpei Li, Yuerou Liu,
Jianghao Lin, Dongdong Ge, Yinyu Ye

arXiv:2605.13473v1 (13 May 2026)
Reading: 22 September 2026

A small write-side change has a specific payoff

- OSDN learns a diagonal write multiplier online from keys and scalar gates.
- At 340M: JRT recall rises from 0.150 to 0.198 (+32% relative).
- At 1.3B: local residual ratio falls by 39%; repeated recall does not improve.
- Central question: how can a preconditioner remain compatible with chunkwise DeltaNet?

Paper, Secs.4–6; Tables 1 and 8, PDF pp.7 and 35. Main talk: approximately 25 minutes.

DeltaNet writes an error, using one scalar step

$$S_{t-1} \in \mathbb{R}^{V \times K}, \quad k_t, q_t \in \mathbb{R}^K, \quad v_t \in \mathbb{R}^V$$

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

An inner regression step

$f_t(S) = \frac{1}{2}\|Sk_t - v_t\|^2$, so $\nabla f_t(S_{t-1}) = -u_t k_t^\top$.

- The learned gate $\beta_t \in (0, 1)$ is shared across key coordinates.
- Repeated, rare, and distracting directions may need different write scales.

Paper, Secs.1, 3 and 4.1, PDF pp.1–5.

Diagonal right-preconditioning scales only the write key

$$D_t = \mathrm{Diag}(d_t), \qquad \widetilde{k}_t = d_t \odot k_t$$

$$\begin{aligned} S_t &= S_{t-1} - \beta_t \nabla f_t(S_{t-1}) D_t \\ &= S_{t-1} + \beta_t \underbrace{(v_t - S_{t-1} k_t)}_{\text{original read key}} \underbrace{\widetilde{k}_t^\top}_{\text{scaled write key}} . \end{aligned}$$

- Readout remains $o_t = S_t q_t$; the memory keeps its $V \times K$ shape.
- Extra persistent state per head: $d_t \in \mathbb{R}^K$.
- Scaling both keys would define a different recurrence.

Paper, Sec.4.1, Eq.(2), PDF pp.4–5.

The residual cancels out of the hypergradient

Let $s_t = k_t^{\odot 2}$, $n_t = \|k_t\|^2$, and $a_t = \beta_t \langle d_t, s_t \rangle$.

$$S_t k_t - v_t = -u_t(1 - a_t)$$

$$h_t(d_t) = \frac{f_t(S_t) - f_t(S_{t-1})}{\|\nabla f_t(S_{t-1})\|_F^2} = \frac{(1 - a_t)^2 - 1}{2n_t}$$

$$\nabla_d h_t(d_t) = -\frac{\beta_t}{n_t}(1 - a_t)s_t$$

Why this matters

The feedback needs keys and gates, but no value, memory matrix, or residual.

Paper, Lemma 4.1 / Eq.(3), PDF p.5; App.A.3, pp.15–16. Derivation assumes nonzero key and residual.

Emit the current write key before updating the scale

1. Emit $\widetilde{k}_t = d_t \odot k_t$.
2. Compute $s_t = k_t^{\odot 2}$ and $\bar{n}_t = \max(\|k_t\|^2, \epsilon)$.
3. Update

$$d_{t+1} = \Pi_{\mathcal{D}}\left[d_t + \eta\beta_t \frac{1 - \beta_t\langle d_t, s_t\rangle}{\bar{n}_t} s_t\right].$$

Reported 340M final configuration

$d_1 = \mathbf{1}$, $\eta = 0.003$, $\mathcal{D} = [0.5, 2]^K$.

With unit keys, $0 < a_t < 2$, hence the local loss ratio $(1 - a_t)^2 < 1$.

Paper, Algorithm 1 / Eq.(4), PDF p.5; Cor.A.2, p.16; App.F, p.32.

Decoupling preserves the expensive chunk computation

Phase 1: a lightweight recurrent sweep

$(k_t, \beta_t, d_t) \longrightarrow \widetilde{k}_t, d_{t+1}$. Work/token: $O(K)$; persistent state/head: $O(K)$.

Phase 2: the asymmetric DeltaNet write

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

Chunk Gram: $BKK^\top \to BK\widetilde{K}^\top$. Local score: $QK^\top \to Q\widetilde{K}^\top$.

Matrix shapes and triangular solves are retained. Phase 1 still has sequence-length sequential depth in Table 2; clipping makes it piecewise affine.

Paper, Secs.4.3–4.5, PDF pp.5–6; Table 2, p.17; App.C, pp.20–23.

APF forgets calibration in the preconditioner

$$r_{t,h} = \sigma(w_{r,h}^{\top} x_t + b_{r,h})$$

$$d_{t+1} = \Pi_{\mathcal{D}}\left[r_{t,h} d_t + \eta \beta_t \frac{1 - \beta_t \langle d_t, s_t \rangle}{\bar{n}_t} s_t \right]$$

- A token-wise, head-wise retention gate responds to changing contexts.
- It acts on d_t; the memory S_t is not directly decayed by APF.
- Adds $H(d_m + 1)$ parameters/layer; reported as at most 0.3% at 340M.
- The current super-geometric analysis does not directly cover APF.

Paper, Sec.4.4 / Eq.(7), PDF p.6; App.A.5, p.16; Discussion, p.9.

Two guarantees control different mathematical objects

Population objective: ideal full-gradient dynamics

$f = \mathbb{E}[f_t]$, $\Sigma_k = \mathbb{E}[kk^\top]$, $L = \lambda_{\max}(\Sigma_k)$.

$$f(S_{T+1}) - f^\star \le [f(S_1) - f^\star] \left(\frac{2L\mathcal{R}_T}{T}\right)^T$$

Monotone updates; cumulative regret $\mathcal{R}_T$ against $D_\star = \Sigma_k^\dagger$.

Token-local objective: diagonal surrogate

$$\prod_{t=1}^{T} \frac{f_t(S_t)}{f_t(S_{t-1})} \le \left(2\varepsilon_T^{\mathrm{diag}} + \frac{2R_T}{T}\right)^T$$

Unit keys; diagonal regret R_T; fixed-comparator gap $\varepsilon_T^{\mathrm{diag}}$.

Paper, Theorems 5.1–5.2, PDF pp.6–7; detailed proofs in App.D.

Super-geometric contraction is conditional

$$\varepsilon_T^{\mathrm{diag}} = \min_{d \in \mathcal{D}} \frac{1}{2T} \sum_t (1 - \beta_t \langle d, k_t^{\odot 2} \rangle)^2$$

- Zero comparator gap requires one feasible $d^\star$ satisfying $\beta_t \langle d^\star, k_t^{\odot 2} \rangle = 1$ for every token.
- Together with $R_T = O(\sqrt{T})$, the bound becomes $(O(1)/\sqrt{T})^T$.
- The practical fixed step $\eta = 0.003$ has no dedicated regret derivation here.
- Reader inference: if $\beta_t < 0.5$ and $d_j \leq 2$, exact matching is impossible at that token.

Paper, Thm.5.2, p.7; App.D.4–D.5, pp.28–31; Discussion, p.9.

Perfect local writes can still overwrite useful memory

The paper's counterexample

$K = V = 1$, $k_t = 1$, $v_t = (-1)^t$, and an exact local write.

$$f_t(S_t) = 0, \qquad S_t = (-1)^t, \qquad S^{\star}_{\text{population}} = 0.$$

- Consecutive losses f_t differ; their ratios do not telescope into a global loss.
- A repeated-key corollary needs fixed targets, disjoint key supports, and compatible blocks.
- Natural-language repeated prompts are not shown to satisfy those assumptions.

Paper, App.D.4, counterexample and Cor.D.8, PDF pp.29–30.

Evaluation separates recall, mechanism, and broad capability

Question	Measurement
Recall	JRT contains accuracy at 2K; single/repeated FDA, SWDE, SQuAD
Write mechanism	Residual-ratio geometric mean on twice prompts
Broad capability	LM PPL; 8-task commonsense; 14-task LongBench
Length extrapolation	PG-19 20K blocks, 2K position buckets
Cost	H100 prefill + decode throughput; state count

340M: 10.74B tokens, fixed seed 42, matched within each backbone. Packed 64K tensors contain segments capped at 4K with state resets.

Paper, Sec.6 and Apps.E–F, PDF pp.7–8 and 31–34. The 2K training claim elsewhere conflicts with App.F.

340M: recall gains are strongest on repeated contexts

Model	All ↑	Single ↑	Repeated ↑	LM PPL ↓
DeltaNet	0.150	0.155	0.145	32.00
OSDN	**0.198**	**0.179**	**0.218**	31.67
OSDN-APF	0.176	0.152	0.199	**30.99**

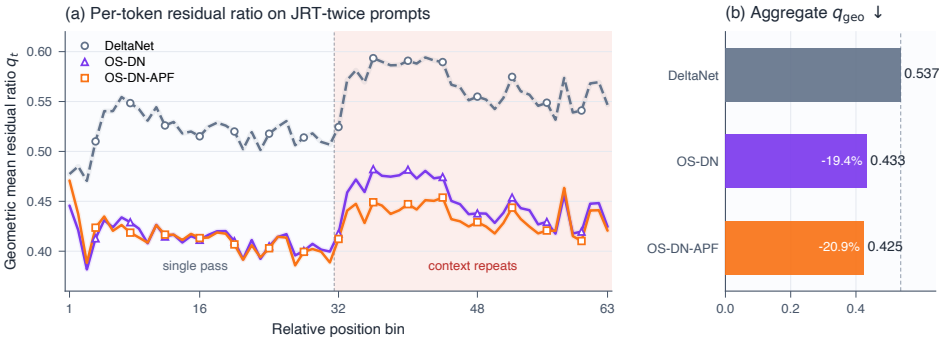
Interpret the headline correctly

Overall: +4.8 percentage points, or +32% relative.
Repeated: +7.3 points, or about +50% relative.

Gains vary by task: OSDN SQuAD-twice is 0.231→0.383; FDA-twice is 0.045→0.044.

Paper, Table 1, PDF p.7; Table 7, p.34. LM PPL = WikiText/LAMBADA geometric mean.

The residual diagnostic supports the intended mechanism



q_{geo}: 0.537→0.433 (OSDN), 0.425 (APF).
48 prompts; 24 layers; 8 heads; about 7.85 million correlated measurements. Lower curves show relative improvement, not uniformly lower values after repetition.

Paper, original Fig.2, PDF p.9; Table 8, p.35. Scalar residual ratio q is distinct from the query vector.

The same intervention has different backbone tradeoffs

Model	Repeated recall ↑	LM PPL ↓	LongBench ↑
GDN	0.139	30.01	0.073
OSGDN	0.195	29.78	0.073
OSGDN-APF	0.221	29.50	0.080
KDA	0.150	**26.75**	0.088
OSKDA	0.133	27.93	0.090
OSKDA-APF	0.179	28.02	0.098

GDN benefits on repeated recall. KDA variants trade language-model perplexity against selected recall and LongBench gains.

Paper, Table 1, PDF p.7; Tables 6–7, p.34. Comparisons are within each host family.

At 1.3B, stronger local contraction does not lift repeated recall

Metric	DeltaNet	OSDN-APF
Residual $q_{\mathrm{geo}} \downarrow$	0.432	**0.265**
Single recall $\uparrow$	0.293	0.315
Repeated recall $\uparrow$	0.227	0.217
LM PPL $\downarrow$	14.28	14.22
Common. $\uparrow$	0.560	0.566
LongBench $\uparrow$	0.115	0.116

Recall averages FDA/SWDE only; SQuAD accuracy is excluded after harness failures. No matched PG-19 result. No multi-seed interval establishes equivalence.

Paper, Table 8, PDF p.35; Tables 15–18, p.40. 1.3B / 100B tokens.

The 1.3B update differs from the main theoretical algorithm

340M beta-aware phase 1, before projection

$$d_{t+1} = r_t d_t + \eta \beta_t (1 - \beta_t \langle d_t, k_t^2 \rangle) k_t^2$$

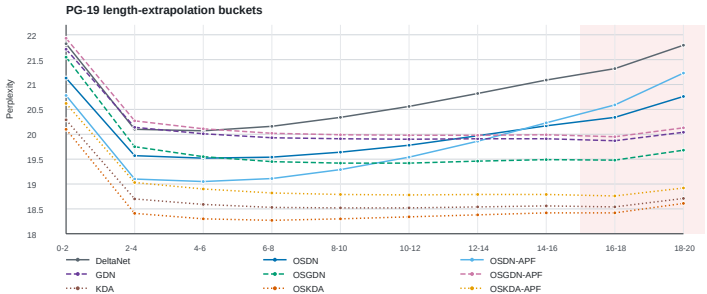
Reported 1.3B non-beta-aware phase 1

$$d_{t+1} = \rho_t^d d_t + \eta (1 - \langle d_t, k_t^2 \rangle) k_t^2$$

Unit-key form shown. The memory write still uses β_t.
The 1.3B configuration also uses a learned initial scale and retention. Its empirical result is not a direct verification of the beta-aware, unforgotten theorem.

Paper, App.B, PDF p.17; App.C, p.20; theoretical scope in Discussion, p.9.

PG-19: lowest final PPL is not lowest late-context drift



DeltaNet / OSDN / APF: final PPL = 20.78 / 20.02 / 19.85.
Late drift = +1.69 / +1.19 / +2.13. APF has the lowest final PPL, while vanilla OSDN has the smallest drift within this family.

Paper, original Fig.4 and Table 13, PDF p.38. Drift = PPL at 18–20K minus PPL at 2–4K.

Ablations and broad checks narrow the claim

Variant	Recall ↑	PG-19 PPL ↓	Late drift ↓
OSDN	0.198	20.02	+1.19
OSDN + learned D_0	0.172	21.57	+5.34
OSDN + $D_0^\star$	0.164	21.42	+4.25
OSDN-APF	0.176	19.85	+2.13

- Learning the initial scale does not recover the best recall or length profile.
- DROP/NQ/TriviaQA averages are all 0.016 for the DeltaNet family.
- Training CE spans only 0.058 nats across completed runs; it is a different loss from the theorem.

Paper, Table 14, p.38; Table 12, p.37; Fig.5, p.39. $D_0^\star$: relaxed initial-scale projection.

Inference overhead is modest, with a state-percentage error

Model	tokens/s ↑	Change	State MiB
DeltaNet 340M	784.5	—	6.000
OSDN 340M	782.7	−0.2%	6.047
OSDN-APF 340M	767.2	−2.2%	6.047
DeltaNet 1.3B	794.0	—	12.000
OSDN-APF 1.3B	739.9	−6.8%	12.094

Single H100; batch 1; 2048-token prefill + 128-token decode.

Reader calculation: $0.047/6.000 \approx 0.78\%$, inconsistent with the paper's "at most 0.05%" claim. Counts use bf16 and exclude activations and short-conv cache.

Paper, Tables 19–20 and state-count formula, PDF pp.41–42. Throughput includes prefill; not decode-only.

What the evidence establishes, and what remains open

- **Established algebra:** diagonal right-preconditioning is an asymmetric write-key substitution with decoupled feedback.
- **Positive evidence:** 340M controlled recall gains and smaller local residual ratios at both scales.
- **Limits:** conditional theory, APF and scale-specific update gaps, one seed, missing 1.3B evaluations, mixed broad-task results.
- **Next test:** match beta-aware/non-beta-aware and APF/no-APF runs at 1.3B; measure retained associations as well as local write errors.

Paper, Discussion, PDF pp.9–10; Apps.B–L. Proposed next test is reader analysis.

Reference and reading materials

Chenyu Zhou, Hongpei Li, Yuerou Liu, Jianghao Lin, Dongdong Ge, and Yinyu Ye.

OSDN: Improving Delta Rule with Provable Online Preconditioning in Linear Attention.

arXiv:2605.13473v1, 13 May 2026.

This presentation is a reading of the paper, prepared on 22 September 2026. It is not an author presentation or an independent experimental reproduction.

Companion: Chinese reading notes, original figure assets, and editable Beamer source in this reading directory.

Source coverage: main text and Appendices A–M; PDF and matching author TeX source.

Backup: the token-local proof is regret plus AM–GM

For nonzero pre-write residuals and unit keys,

$$q_t = (1 - \beta_t \langle d_t, s_t \rangle)^2 = 1 + 2h_t(d_t).$$

For any fixed comparator $d \in \mathcal{D}$,

$$\begin{aligned} \frac{1}{T} \sum_t q_t &= 1 + \frac{2}{T} \sum_t h_t(d_t) \\ &\le 1 + \frac{2}{T} \left(\sum_t h_t(d) + R_T(d) \right) \\ &= 2\varepsilon_T(d) + \frac{2R_T(d)}{T}. \end{aligned}$$

Apply $\prod_t q_t \le (\frac{1}{T} \sum_t q_t)^T$.

Paper, Thm.D.7 proof, PDF p.29. Zero-residual conventions need separate treatment.

Backup: the chunk matrices keep the read/write asymmetry

For a chunk of C tokens, $K, \widetilde{K} \in \mathbb{R}^{C\times K}$, $V \in \mathbb{R}^{C\times V}$, and $B = \mathrm{Diag}(\beta)$:

$$T = \left[I + \mathrm{tril}(BK\widetilde{K}^{\top}, -1)\right]^{-1}$$

$$W = TBK, \qquad U = TBV$$

$$S_{\mathrm{next}} = S + (U - WS^{\top})^{\top}\widetilde{K}.$$

The physical key remains in W. The scaled key appears in storage interactions and the final write. Materializing $\widetilde{K}$ also consumes temporary sequence storage.

Paper, App.A.2, PDF p.15; Listing 1, pp.20–22.

Backup: GDN and KDA scale the same storage factor

GDN: apply its state gate first,

$$\bar{S} = \alpha_t S_{t-1}, \quad u_t = v_t - \bar{S} k_t, \quad S_t = \bar{S} + \beta_t u_t \widetilde{k}_t^\top.$$

KDA: use its native transposed state, $S_t \in \mathbb{R}^{K \times V}$,

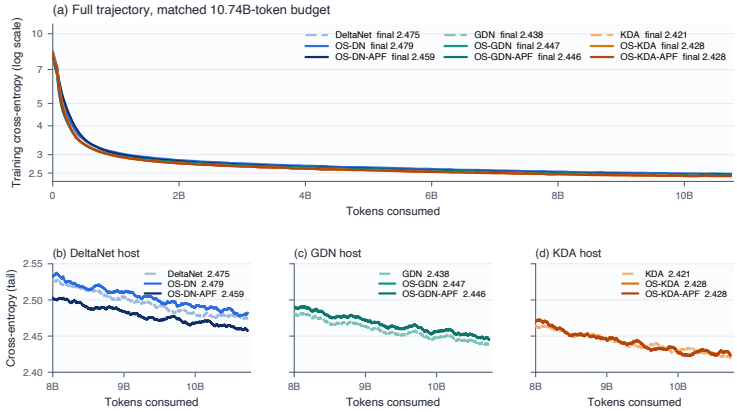
$$\bar{S} = \mathrm{Diag}(\alpha_t) S_{t-1}, \quad u_t = v_t - \bar{S}^\top k_t,$$

$$S_t = \bar{S} + \beta_t \widetilde{k}_t u_t^\top.$$

The state-retention gate and preconditioner-retention gate have distinct roles. Post-gate residuals preserve the local feedback derivation.

Paper, App.B and Table 3, PDF pp.17–19.

Backup: training cross-entropy is a separate diagnostic



257-step centered smoothing; final values average 256 steps. These curves do not establish a super-geometric language-model loss rate.

Paper, original Fig.5, PDF p.39. Full panel has log y-axis; late panels have linear y-axes.

Backup: reporting details matter for reproduction

- App.E says 2K training context; App.F and Fig.4 specify segments capped at 4K.
- App.C mentions OSKDA retention 0.999; App.F labels refreshed final OSKDA as no-retention.
- Table 9 selects the best completed no-APF cell per task before averaging commonsense results.
- APF commonsense is 0.454 in the ablation but 0.456 in the main results.
- Fig.1 uses d_{t+1} for the write key; Algorithm 1 emits $d_t \odot k_t$ before updating.

Paper, PDF pp.2, 5, 23, 32, 35 and 38. Reading follows explicit final configurations and algorithm ordering.