# SpectralShift

## Context extension through Gated DeltaNet spectral dynamics

Independent research paper discussion

Paper: Zian Liu et al.

arXiv:2609.14320v1 | September 18, 2026

# Longer inputs require longer-lived state directions

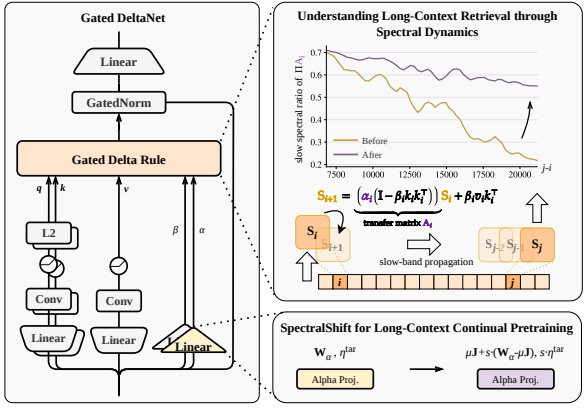
## Research question

How should Gated DeltaNet recurrence change when the training window grows?

- Longer sequences and positional scaling do not directly fix recurrent forgetting.
- A few long-lived directions may be insufficient: capacity and write alignment both matter.
- Fast decay remains useful for state clearing and context switching (mechanistic interpretation).

Paper, Secs. 1–3, PDF pp. 1–4.

# SpectralShift changes the alpha branch at training time




Paper, Fig. 1, PDF p. 2. Original author figure; quantitative evidence follows.

# Alpha controls shared decay; keys control direction

$$\mathbf{S}_t = \alpha_t(\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top)\mathbf{S}_{t-1} + \beta_t \mathbf{k}_t \mathbf{v}_t^\top$$

$$\mathbf{S}_t \in \mathbb{R}^{d_k \times d_v}, \quad \|\mathbf{k}_t\|_2 = 1, \quad \mathbf{o}_t = \mathbf{q}_t^\top \mathbf{S}_t$$

- $\alpha_t$: retention shared across a head; $\beta_t$: directional update strength.
- One-step eigenvalues: $\alpha_t(1 - \beta_t)$ along the key, and $\alpha_t$ orthogonally.
- Multi-step transitions generally do not commute.

App. A, pp. 12–13; App. E.4, pp. 18–19. We use the E.4 write orientation; Sec. 2 reverses it.

# Finite-time singular values measure survival

$$\mathbf{\Phi}_{i\to j} = \mathbf{A}_j \cdots \mathbf{A}_{i+1} = \left( \prod_{t=i+1}^{j} \alpha_t \right) \mathbf{R}_{i\to j}, \qquad \ell = j - i$$

$$\gamma_r = -\frac{\log \sigma_r(\mathbf{\Phi})}{\ell} = \underbrace{-\frac{1}{\ell} \sum_t \log \alpha_t}_{\text{shared decay } \gamma_\alpha} + \underbrace{-\frac{1}{\ell} \log \sigma_r(\mathbf{R})}_{\text{directional decay}}$$

- The relevant spectrum belongs to the full interval, conditioned on the input.
- Alpha scales all singular values within one head on a fixed directional trajectory.

Paper, Sec. 2, PDF p. 3; App. A.2, pp. 12–13.

# A slow spectral band counts usable propagation directions

$$\mathcal{S}_\ell = \{r : \log \sigma_r(\mathbf{\Phi}) \geq -1\} = \{r : \gamma_r \leq 1/\ell\}$$

- $H_\ell^{\text{slow}}$: layer–heads with at least one qualifying mode.
- $M_\ell^{\text{slow}} = \sum_h |\mathcal{S}_{\ell,h}|$: total qualifying directions.
- The threshold means at least $e^{-1} \approx 0.368$ amplitude survives the interval.
- This is a capacity proxy, not a guarantee of successful retrieval.

**Teaching example (not experimental data)**

Retention 0.9999 per step leaves about 0.449 after 8,000 steps, but 0.041 after 32,000.

Paper, Sec. 2, PDF p. 3. Illustrative scalar calculation added for this presentation.

# Writing into slow directions matters as well

$$\mathbf{\Phi} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \qquad E_{\text{write}} = \frac{\|\mathbf{V}_{\mathcal{S}_\ell}^\top \mathbf{k}_i\|_2^2}{\|\mathbf{k}_i\|_2^2}$$

| Window | $H^{slow}$ | $M^{slow}$ | $E_{write}$ | NIAH-3 avg. |
|---|---|---|---|---|
| 8K, Low | 18 | 1036 | 0.1006 | 8.60 |
| 8K, High | 22 | 1308 | 0.1503 | 32.93 |
| 32K, Low | 15 | 984 | 0.0964 | 10.07 |
| 32K, High | 19 | 1109 | 0.1819 | 30.47 |

Capacity and alignment improve together; their causal contributions are not isolated.

Tables 1–2, pp. 3–4. Capacity: longest distance; alignment: mean over heads and four positions.

# Broader slow capacity coexists with very fast heads

$$\tau_1 = -1/\log \sigma_1(\mathbf{\Phi}), \qquad \text{fast head: } 0 < \tau_1 < 0.1$$

| Window | Low-Retrieval | High-Retrieval |
|---|---|---|
| 8K | 4 | 5 |
| 32K | 5 | 5 |

- Even the slowest direction in a qualifying head decays quickly.
- Counts support preservation of an extreme fast component.
- They do not establish identical head identities or unchanged switching behavior.

Paper, App. B.3–B.4, Table 8, PDF p. 14.

# Initialization contracts weights around their global mean

$$\rho = \frac{L_{\mathrm{ref}}}{L_{\mathrm{tar}}}, \qquad s_1 = \sqrt{\rho}$$

$$\boxed{\mathbf{W}_\alpha^* = \mu_\alpha \mathbf{J} + s_1(\mathbf{W}_\alpha^{\mathrm{ref}} - \mu_\alpha \mathbf{J})}$$

$$\mathbf{a}_t(s_1) = \mathbf{c}_t + s_1 \boldsymbol{\xi}_t, \qquad \mathbf{c}_t = \mu_\alpha \mathbf{J} \mathbf{h}_t$$

- Use the global element mean of each alpha projection matrix; preserve that mean.
- Apply once when loading the checkpoint for the new extension stage.
- No additional inference-time spectral computation is required.

Paper, Algorithm 1 and Sec. 4.1, PDF pp. 4–5.

# Smaller scaling does not make every token forget less

$$\alpha_t(s_1) = \exp\left[-e^{A_{\log}}\,\mathrm{softplus}(c_t + s_1\xi_t + b_{dt})\right]$$

| When $s_1$ decreases | Retention | Local forgetting |
|---|---|---|
| $\xi_t > 0$ | $\alpha_t$ increases | Slower |
| $\xi_t < 0$ | $\alpha_t$ decreases | Faster |

- The accumulated head-wise effect depends on the realized sequence.
- Exact shared displacement assumes fixed hidden states, keys, and beta gates.
- Upstream changes can alter later-layer trajectories in the full network.

Paper, Lemma 1, p. 5; App. D.2–D.4, pp. 16–17.

# The square-root rule matches a conditional variance scale

$$\mathcal{D}_L(s) = \mathrm{Var}[\Delta_L(s)] = \Theta(s^2 L)$$

$$\frac{\mathcal{D}_{L_{\mathrm{tar}}}(\rho^k)}{\mathcal{D}_{L_{\mathrm{ref}}}(1)} = \Theta(\rho^{2k-1}) \quad \Longrightarrow \quad k = \tfrac{1}{2}$$

- Assumption: centered token-response accumulation has variance of order length.
- Objective: preserve input-dependent log-retention dispersion across lengths.
- Optimal exponent within this power-law family and scale-matching criterion.
- This is not a global optimum for language-model loss or retrieval accuracy.

Paper, App. C, PDF pp. 14–15.

# A smaller alpha learning rate limits shared spectral drift

$$\eta_\alpha^{\mathrm{tar}} = s_2 \eta^{\mathrm{tar}}, \qquad s_2 = s_1 = \sqrt{\rho}$$
$$|\gamma_\alpha^{(u+1)} - \gamma_\alpha^{(u)}| \leq C_\alpha s_2 \eta + O((s_2\eta)^2)$$

- Local bound: fixed layer inputs and bounded sensitivities / update directions.
- Other trainable parameters retain their base learning rate.
- Both the shared-mean and centered alpha components continue to learn.
- Actual implementation splits fused input projections into optimizer groups.

Theorem 1, p. 5; App. E, pp. 17–20; implementation: App. G.2, p. 25.

# Strict slow-band expansion needs margins and update budgets

- Positive initialization gain requires activation balance, not just weight centering.
- Bound scalar and directional changes along the realized full-network CPT trajectory.
- Then $D_r \geq g_\star/(2\ell) > 0$ for certified reference indices.
- Existing slow indices remain slow; near-boundary indices enter if candidates exist.
- Fast-head preservation requires a separate upper-survival condition.

## Evidence boundary

These are sufficient conditions. The paper does not certify the budgets along its actual 10B/20B-token training runs.

Theorem 2, p. 5; App. F, PDF pp. 20–25. Details in backup slides.

# Main experiments compare matched context-extension training

- 1.5B total / 0.6B active GDN-MoE; 16 layers; 3 GDN layers per full-attention layer.
- Base: 8K context, 500B Dolma3 tokens, Muon optimizer.
- 32K extension: 10B tokens. 64K / 128K: direct 20B or staged 10B + 10B via 32K.
- Default full-attention positional extension: ABF. GDN layers: NoPE.
- General benchmarks, DROP / RACE, and RULER at 8K–128K. Higher is better.

Paper, Sec. 5.1, PDF p. 6; App. G, p. 25. Spectral probes use a separate 4K reference.

# Target-length RULER improves across the five main settings

| Target | CPT tokens | Baseline | SpectralShift | Gain |
|---|---|---|---|---|
| 32K | 10B | 40.6 | 45.8 | +5.2 |
| 64K | 10B+10B | 41.6 | 45.4 | +3.8 |
| 64K | 20B | 46.4 | 48.4 | +2.0 |
| 128K | 10B+10B | 43.1 | 44.6 | +1.5 |
| 128K | 20B | 41.9 | 42.2 | +0.3 |

- Each row uses its own target evaluation length and matched baseline.
- Benefits depend on training path; staged extension is not always best.
- No significance claim is made for small differences.

Paper, Table 3, PDF p. 6. Gains are score-point differences computed from the table.

# The reported 5.35 percent gain averages five evaluation lengths

| 128K staged CPT | 8K | 16K | 32K | 64K | 128K |
|---|---|---|---|---|---|
| Baseline | 64.3 | 57.2 | 51.9 | 45.4 | 43.1 |
| SpectralShift | 64.6 | 60.1 | 56.7 | 49.9 | 44.6 |

Mean: $52.38 \rightarrow 55.18$, $\quad \dfrac{55.18 - 52.38}{52.38} \approx 5.35\%$

- Absolute mean gain: 2.80 points. Target-point gain at 128K: 1.5 points.
- An exception elsewhere: 64K direct CPT, evaluated at 16K, gives 59.7 vs 60.5.
- General average decreases at 32K: 53.1 vs 54.3.

Paper, Table 3 and Sec. 5.2, PDF pp. 6–7.

# Initialization and learning-rate scaling work better together

| 8K → 128K ablation | General | R-32K | R-64K | R-128K |
|---|---|---|---|---|
| No changes | 54.0 | 47.4 | 43.6 | 42.8 |
| Init only, $\rho^{0.5}$ | 54.1 | 50.6 | 43.5 | 39.8 |
| Init + LR, $\rho^{0.5}$ | 54.3 | 52.5 | 48.5 | 42.9 |
| Init + LR, $\rho$ | 52.2 | 51.4 | 43.5 | 40.1 |

- Adding LR scaling to square-root initialization improves R-64K by 5.0 points.
- The final 128K gain over no changes is only 0.1 point in this ablation.
- No LR-only row: independent necessity of initialization is not isolated.

Paper, Table 4, PDF p. 7. Selected rows; this baseline differs from Table 3. $\rho = L_{ref}/L_{tar}$.

# Compatibility extends to several positional strategies

| Target-point gain | DroPE | YaRN | ABF |
|---|---|---|---|
| 8K $\rightarrow$ 32K | +1.0 | +0.2 | +5.2 |
| 32K $\rightarrow$ 64K | +1.1 | +1.0 | +6.3 |

- Table 5 shows positive target-point differences with all three strategies.
- Some shorter evaluation lengths decline, including YaRN at 8K / 16K.
- ABF entries disagree with Table 3 in several places; retain table-specific comparisons.

Paper, Table 5, PDF p. 7. Gains computed within Table 5; discrepancies documented in notes.

# Pure GDN gains remain far from reliable retrieval

| RULER Single-1 | 4K $\rightarrow$ 8K | 4K $\rightarrow$ 16K | 4K $\rightarrow$ 32K |
|---|---|---|---|
| Baseline | 11.11 | 8.00 | 6.43 |
| SpectralShift | 11.96 | 11.25 | 10.63 |
| Gain | +0.85 | +3.25 | +4.20 |

- The absolute improvement grows with target length in this experiment.
- This is a single retrieval subtask, not the full RULER aggregate.
- The 32K score is still 10.63; pure-model configuration details are incomplete.

Paper, Table 6 and Sec. 5.3, PDF p. 8; App. G mainly specifies the hybrid model.

# Aggregate gains can hide substantial subtask losses

| 64K direct CPT, evaluated at 64K | Baseline | SpectralShift |
|---|---|---|
| NIAH multi-key | 31.9 | 48.1 |
| Variable tracking | 11.3 | 24.6 |
| NIAH multi-query | 55.4 | 35.4 |
| NIAH multi-value | 61.2 | 37.5 |

- 128K direct CPT: variable tracking is 0.0 at 32K / 64K / 128K.
- Fast-head counts and general averages do not imply all tasks are preserved.
- Reported evaluation variance lacks enough protocol detail for significance tests.

Paper, Tables 10–11, PDF pp. 27–28; variance description on p. 26.

# Takeaways: useful intervention, conditional explanation

- Finite-time slow capacity and write alignment are useful diagnostics for GDN extension.
- Centered alpha initialization plus a smaller alpha learning rate improves several matched target-length results.
- The square-root derivation and preservation guarantees have explicit assumptions.
- Scale, subtask tradeoffs, and reporting inconsistencies limit broader claims.

**Discussion**

Would readout alignment and write interference explain the multi-query / multi-value regressions?

Synthesis of Secs. 3–5 and Apps. B–I. Discussion question is reader analysis.

# Reference and reading scope

**SpectralShift: Effective Context Window Extension of Gated DeltaNet via Spectral Reparameterization**

Zian Liu, Yiwen Hu, Zican Dong, Tian Xie, Wayne Xin Zhao, Yucheng Ding, Ran Tao, Bryan Dai.

arXiv:2609.14320v1, submitted September 13, 2026.

Reading date: September 18, 2026. Source: version-matched author TeX and 28-page PDF, including substantive appendices A–I.

Independent paper discussion; no model training or experimental reproduction was performed.

Local companion: reading-notes.zh.md. Main presentation: approximately 25 minutes; backup follows.

# Backup: spectral probes have a distinct protocol

- Appendix B: a 4K reference, paired 8K / 32K extension; 96 analyzed layer–heads.
- Four needle positions: 0.2, 0.4, 0.6, 0.8. Query fixed at the end.
- Forward pass in BF16; transition products and SVD in FP64.
- Write alignment averages all heads and positions, counting heads without slow modes as zero.
- Table 1 calls the model pure GDN; Appendix B calls it hybrid. This inconsistency is unresolved.

Paper, App. B, PDF pp. 13–14; Table 1 caption, p. 3.

# Backup: a positive margin survives bounded perturbations

$$M = \sum_t \sigma(b_t)\xi_t, \quad V = \sum_t \kappa_t \xi_t^2, \quad M \geq -\frac{1+s}{4} V, \quad V > 0$$

$$g \geq G = \frac{e^{A_{\log}}(1-s^2)}{4} V \geq g_\star > 0$$

$$|P^{(u)} - \widetilde{P}| \leq g_\star/4, \qquad \|\mathbf{R}^{(u)} - \mathbf{R}^{ref}\|_2 \leq a(1 - e^{-g_\star/4})$$

$$\sigma_r(\mathbf{R}^{ref}) \geq a \quad \Longrightarrow \quad D_r^{(u)} \geq \frac{g_\star}{2\ell}$$

Here $b_t = c_t + b_{dt}$; $\kappa_t$ is a lower bound on softplus curvature along the offset segment.

Paper, App. F.2–F.3, PDF pp. 21–22. These conditions concern a fixed input and interval.

# Backup: finite-run guarantees constrain all parameter updates

$$\sum_{u=0}^{K-1} \|\boldsymbol{\theta}^{(u+1)} - \boldsymbol{\theta}^{(u)}\|_2 \le r_\star$$

$$\eta_{base} \le \frac{r_\star}{K(s_2 B_\alpha + B_{oth})}$$

- $r_\star$: radius allowed by initial slack and local sensitivity.
- $B_\alpha, B_{oth}$: bounds on effective optimizer update directions, including decay.
- Smooth local region; hard MoE routing must stay in a fixed routing region.
- Fast-head sufficient condition: $\widetilde{P} - b_\alpha > 1/\tau_0$; for $\tau_0 = 0.1$, the threshold is 10.

Paper, App. F.5–F.6, PDF pp. 23–24.

# Backup: reporting issues to resolve before reproduction

- Write orientation: Sec. 2 uses value–key; App. E.4 uses key–value.
- Probe architecture: Table 1 says pure GDN; Appendix B specifies hybrid.
- 32K ABF DROP: Table 3 gives 21.5 / 22.3; Table 5 gives 25.1 / 25.9.
- 64K ABF RULER entries also differ across Tables 3 and 5.
- 128K RoPE base, pure-GDN training details, and variance protocol need clarification.

Paper, pp. 2–3, 6–7, 13, 18–19, 25–28. Values ordered SpectralShift / baseline.