On State Reduction in Linear Attention

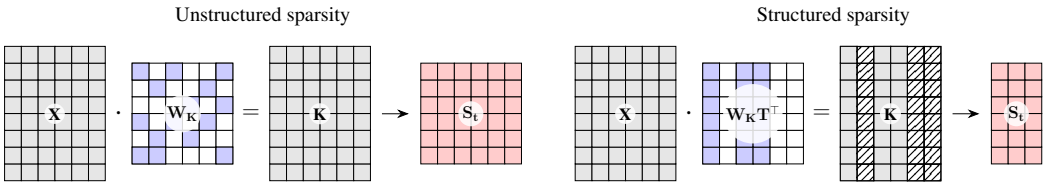
Research group reading presentation

Paper by Philipp Nazari and T. Konstantin Rusch

arXiv:2602.04852v3 (29 September 2026)
Reading date: 13 October 2026

Joint key/query channel selection can halve recurrent state,
with task-dependent quality costs and workload-dependent speedups.

Deleting channels makes the recurrent state smaller



$$\mathbf{S}_t \in \mathbb{R}^{d_v \times d_k} \quad \longrightarrow \quad \mathbf{S}'_t \in \mathbb{R}^{d_v \times d'_k}, \qquad d'_k < d_k$$

- Sparse projection weights can leave the state shape unchanged.
- Select entire key/query channels and their convolution filters.
- 50% compression means half the recurrent matrix entries in the pruned layers; total model memory has other components.

Paper, Fig. 1, PDF p. 2; Secs. 3.2–3.3, pp. 5–6.

DeltaNet writes a prediction residual into memory

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \beta_t(\mathbf{v}_t - \mathbf{S}_{t-1}\mathbf{k}_t)\mathbf{k}_t^\top, \qquad \mathbf{o}_t = \mathbf{S}_t\mathbf{q}_t$$

- $\mathbf{k}_t, \mathbf{q}_t \in \mathbb{R}^{d_k}$: write and read addresses.
- $\mathbf{v}_t, \mathbf{o}_t \in \mathbb{R}^{d_v}$; β_t: scalar update strength.
- The key controls both erasure and writing; the query controls readout.

The state inherits its available directions

$$\operatorname{rank}(\mathbf{S}_t) \le \min\{\operatorname{rank}(\mathbf{K}_t), \operatorname{rank}(\mathbf{V}_t)\} \le t$$

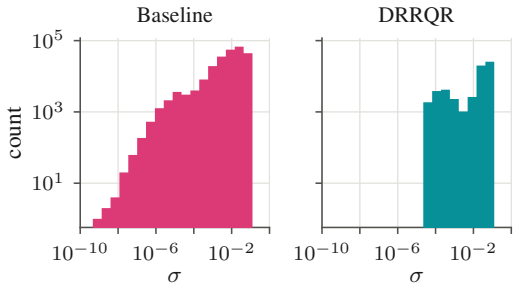
Paper, Eqs. (1)–(3), PDF pp. 2–3; proof in App. F.2, pp. 25–26. Here $\mathbf{K}_t, \mathbf{V}_t$ stack historical vectors as columns.

Stable rank measures concentration of spectral energy

$$\mathrm{er}(\mathbf{S}) = \frac{\|\mathbf{S}\|_F^2}{\|\mathbf{S}\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2}$$

$$u(\mathbf{S}) = \frac{\mathrm{er}(\mathbf{S})}{d}, \quad d = \min(d_v, d_k)$$

- $u = 1$: all d singular values are equal.
- Low u: energy is concentrated in a few directions.



DeltaNet 370M, one head.
75% reduction, before recovery fine-tuning (RFT).

Paper, Sec. 2.2, Eq. (5), Fig. 2, PDF pp. 3–4. Both histogram axes are logarithmic; FineWeb, $T = 2048$, first 128 tokens skipped.

Key conditioning enters a lower bound on state rank

For additive memory, $\mathbf{S} = \mathbf{V}\mathbf{K}^\top$:

$$\boxed{\mathrm{er}(\mathbf{S}) \geq \frac{\mathrm{er}(\mathbf{V}_\parallel)}{\kappa(\mathbf{K})^2}}, \qquad \kappa(\mathbf{K}) = \frac{\sigma_{\max}(\mathbf{K})}{\sigma_{\min}(\mathbf{K})}$$

- $\mathbf{K} \in \mathbb{R}^{d_k \times T}$ has full row rank.
- $\mathbf{V}_\parallel \neq 0$: value rows projected onto $\mathrm{row}(\mathbf{K})$.
- DeltaNet uses pseudo-values $\mathbf{u}_t = \beta_t(\mathbf{v}_t - \mathbf{S}_{t-1}\mathbf{k}_t)$.

Interpretation

Avoid nearly redundant key directions, while preserving useful signals.
The projected values also change when channels change.

Paper, Eq. (4), PDF p. 3; assumptions and proof in App. F.1, p. 25. The paper's $\nu(\mathbf{V}) = \mathrm{er}(\mathbf{V}_\parallel)$ also depends on $\mathbf{K}$.

An uneven spectrum permits large relative readout error

With $\widetilde{\mathbf{q}} = \mathbf{q}^* + \varepsilon$, define

$$R = \frac{\|\mathbf{S}\varepsilon\|_2 \,/\, \|\mathbf{S}\mathbf{q}^*\|_2}{\|\varepsilon\|_2 \,/\, \|\mathbf{q}^*\|_2}.$$

For a full-column-rank state:

$$\max_{\|\mathbf{q}^*\|_2=1,\ \varepsilon\neq 0} R = \kappa(\mathbf{S}) \geq \frac{1}{\sqrt{u(\mathbf{S})}}$$

- Worst case: query along the weakest direction, noise along the strongest.
- At exact isotropy, $u = 1$ and $R = 1$.
- This describes worst-case sensitivity; actual alignment matters.

Paper, Proposition 2.1, Eqs. (6)–(7), PDF p. 4; App. F.3, pp. 26–27.

Joint orthogonal changes preserve the sequence mixer

For a square orthogonal $\mathbf{T}$:

$$\mathbf{k}'_t = \mathbf{T}\mathbf{k}_t, \qquad \mathbf{q}'_t = \mathbf{T}\mathbf{q}_t, \qquad \mathbf{S}'_t = \mathbf{S}_t\mathbf{T}^\top$$

$$\mathbf{S}'_t\mathbf{q}'_t = \mathbf{S}_t\mathbf{T}^\top\mathbf{T}\mathbf{q}_t = \mathbf{S}_t\mathbf{q}_t.$$

- The Delta recurrence keeps the same form; scalar decay also commutes.
- Both keys and queries must use the same transformation.

Dimension reduction changes the problem

For $\mathbf{T} \in \mathbb{R}^{d'_k \times d_k}$, $\mathbf{T}^\top\mathbf{T} \neq \mathbf{I}$.
Compression seeks an acceptable approximation to the original dynamics.

Paper, Proposition 3.1 and Secs. 3.1–3.2, PDF p. 5.

Channel selection preserves the efficient layer structure

$$\underbrace{\mathbf{T}\,\mathrm{diag}(\mathbf{w}^{(j)})\mathbf{T}^{\top}}_{\text{generally dense}} \quad \text{vs.} \quad \underbrace{\mathbf{P}_{\mathcal{I}} = [\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_{d'_k}}]^{\top}}_{\text{select original coordinates}}$$

1. Pick one retained set $\mathcal{I}$ per head.
2. Slice matching columns of $\mathbf{W}_K$ and $\mathbf{W}_Q$.
3. Slice the matching depthwise convolution filters.
4. Execute the same recurrence with a smaller d'_k.

Dense PCA rotations conflict with depthwise convolution.
Shared filters address convolution, but SiLU still breaks general rotation equivariance.

Paper, Secs. 3.2–3.3, PDF pp. 5–6; App. B.3–B.5, pp. 17–18. The retained channels are normalized in the reduced model.

Three baselines score individual channels

L1: projection-weight magnitude

$$s_j = \|\mathbf{W}_{Q,:j}\|_1 + \|\mathbf{W}_{K,:j}\|_1$$

S-Wanda: magnitude weighted by input activation

$$s_j = \sum_i (|W_{Q,ij}| + |W_{K,ij}|)\, \|\mathbf{X}_{:i}\|_2$$

Grad: first-order task-loss sensitivity

$$s_j = \sum_i (|W_{Q,ij} \nabla_{W_{Q,ij}} \mathcal{L}| + |W_{K,ij} \nabla_{W_{K,ij}} \mathcal{L}|)$$

Select the top d'_k scores within each head.

Grad alone uses full-model loss gradients.

Paper, Sec. 3.4, PDF pp. 6–7. Rand selects channels randomly.

DRRQR selects original columns with a stable basis

$$\mathbf{M} = \begin{bmatrix} \mathbf{K}_{\text{cal}} \\ \mathbf{Q}_{\text{cal}} \end{bmatrix} \in \mathbb{R}^{2N \times d_k}, \qquad \mathbf{M}\Pi = \mathbf{Q}_{\text{QR}} \begin{pmatrix} \mathbf{A}_k & \mathbf{B}_k \\ 0 & \mathbf{C}_k \end{pmatrix}$$

1. Start with column-pivoted QR; $k = d'_k$.
2. Evaluate swaps between selected and unselected columns.
3. Accept swaps that improve the selected volume beyond tolerance f.
4. Return the first k original channel indices in the permutation.

What is guaranteed

Strong RRQR bounds the selected joint activation spectrum.
The model uses the selected channels, with no QR rotation at inference.

Paper, Sec. 3.4, PDF pp. 6–7; Algorithm 2, p. 15; App. B.2, pp. 15–16. The full bound and swap rule are in backup slide 23.

Calibration and recovery answer different questions

Calibration and models

- Default: 128 FineWeb samples.
- DRRQR: subsample 2.5k keys and 2.5k queries.
- DeltaNet / Gated DeltaNet: 370M and 1.3B.
- Mamba-2 370M, Codestral 7B; hybrid Qwen3.5 9B.

Recovery fine-tuning

- LoRA rank 32, $\alpha = 64$.
- 32k FineWeb samples; length 2048; batch 16.
- LR $10^{-4} \rightarrow 10^{-5}$; 5% warmup.
- Train convolutions and norms; distill from original model.

Pre-RFT isolates immediate pruning quality. Post-RFT includes recovery.

Paper, App. A, PDF p. 14. 370M DN/GDN: 10B pretraining tokens. Mamba results are pre-RFT. Qwen RFT updates recurrent components only.

Half-sized states often recover language-model quality

DRRQR at 50% state reduction

	WikiText-2 PPL ↓			Zero-shot avg. % ↑	
Model	Base	Pre	Post	Base	Post
DN 370M	29.8	31.4	29.4	44.2	44.5
DN 1.3B	16.7	20.5	17.5	50.0	49.7
GDN 370M	28.8	31.6	28.7	44.4	43.8
GDN 1.3B	16.8	17.3	16.3	59.4	58.0

- Recovery matters: DN 1.3B LAMBADA PPL is $11.9 \rightarrow 47.2 \rightarrow 11.8$.
- Small post-RFT gains include added training and distillation.

Paper, Table 20, PDF p. 29. DN = DeltaNet; GDN = Gated DeltaNet. Six-task average: ARC-E/C, HellaSwag, WinoGrande, PIQA, LAMBADA.

More compression exposes a gradual quality trade-off

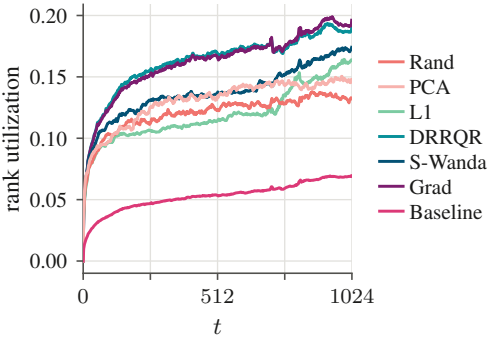
Gated DeltaNet 1.3B, after RFT

Reduction	Selector	Wiki PPL ↓	LMB PPL ↓	Avg. % ↑
0%	Baseline	16.8	9.7	59.4
30%	DRRQR	15.9	10.7	58.8
30%	Grad	15.9	9.9	59.0
50%	DRRQR	16.3	12.0	58.0
50%	Grad	16.4	10.5	58.3
75%	DRRQR	17.9	14.7	56.2
75%	Grad	17.8	12.5	56.8

At 50%, DRRQR has lower WikiText PPL; Grad has better LAMBADA and average task accuracy. The selector ranking depends on the metric.

Paper, Table 1, PDF p. 8; full task breakdown in Table 24, p. 33.

Higher utilization can coexist with lower absolute rank



DeltaNet 370M, 75% reduction

	d_k	Mean u
Baseline	128	0.053
DRRQR	32	0.170
L1	32	0.124

Reading-derived calculation:

mean stable rank $\approx 6.78 \rightarrow 5.44$.

Most effective dimensions survive in one quarter of the capacity.

Paper, Fig. 3, PDF p. 7; Table 5, p. 18. All compressed curves have the same dimension; baseline has a different denominator. Pre-RFT, FineWeb.

A low condition number can hide discarded signal

Selected joint key/query columns: DeltaNet 370M, 75% reduction

Selector	$\sigma_{\min}$	$\sigma_{\max}$	κ
Baseline (128)	0.10	49.4	529
DRRQR (32)	0.96	49.4	65
Grad (32)	0.62	48.5	89
S-Wanda (32)	0.65	22.5	42
L1 (32)	0.64	15.3	32

- DRRQR retains the dominant spectral scale and lifts the smallest singular value.
- L1 has a smaller κ, but removes much of the dominant signal.
- DRRQR and Grad overlap on a median 26/32 channels; random overlap is 8.

Paper, App. C.2, Table 6, PDF p. 18. Each column reports a per-head median; median ratios need not equal ratios of medians.

DRRQR transfers more reliably across calibration corpora

DeltaNet 370M, 75% reduction, before RFT

Calibration corpus	DRRQR PPL ↓	Grad PPL ↓
FineWeb	41.6	44.9
PG19	41.5	53.8
Pile	41.7	51.0

- Fixed evaluation: WikiText-2; 128 calibration sequences of length 2048.
- DRRQR: 41.6–41.7 PPL across 8–512 FineWeb samples.
- Grad is stable in-domain, but its scores depend on calibration loss.

Paper, Table 2, PDF p. 8; full calibration study, Table 7, p. 19. These dedicated ablation values differ from the standard 75% run in Table 21.

The framework extends to Mamba, with visible quality cost

Mamba-Codestral 7B, 50% reduction, before RFT

Selector	Wiki PPL ↓	LMB PPL ↓	Avg. % ↑
Baseline	14.5	6.7	64.4
Grad	16.7	12.7	59.9
DRRQR	17.6	28.7	57.1
L1	25.7	73.7	48.8

- Mamba-2's $\mathbf{B}, \mathbf{C}$ play the roles of keys and queries.
- Both selectors outperform magnitude pruning; Grad is stronger here.
- DRRQR loses 7.3 accuracy points; 7B results are reported before recovery.

Paper, Tables 9–10, PDF pp. 19–20; main-text Table 3, p. 8.

Retrieval separates pure recurrence from the tested hybrid

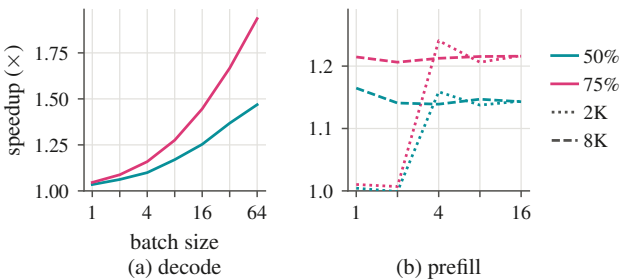
50% recurrent-state reduction, after RFT; scores ↑

Model	Configuration	RULER 4K	NIAH 4K
GDN 1.3B	Baseline	38.8	81.3
	DRRQR	27.5	61.9
Qwen3.5 9B	Baseline	93.9	100.0
	DRRQR	93.6	100.0

- GDN loses 19.4 NIAH points despite competitive language-model metrics.
- Qwen prunes and fine-tunes recurrent components; attention stays frozen.
- Supports a memory-path hypothesis; models also differ in scale and training.

Paper, Table 4, PDF p. 9; Tables 11–12, p. 20. The reported context lengths are 2K and 4K.

End-to-end speedup grows with decode batch size



GDN 1.3B, H100

50% state reduction

- Decode, $B = 1$: 1.03×.
- Decode, $B = 64$: 1.47×.
- Prefill: usually 1.14–1.16×.

At $B = 64$, $11.61/1.47 \approx 7.90$ ms/token:
47% more speed, about 32% less time.

Paper, Fig. 4, PDF p. 9; Tables 17–18, p. 28. CUDA graphs, unchanged FLA kernels, 10 warmups, 100 measured steps, median of 3 trials. Decode prompt: 512 tokens. Prefill at 2K, $B = 1$–2: 1.00×.

Coupled scoring helps DRRQR, but magnitude has a caveat

DeltaNet 370M, 50% reduction; WikiText-2 PPL ↓

Selector	Score from K+Q	Score from K	Score from Q
L1	3843.5	425218.3	33.0
S-Wanda	915.9	376074.0	33.0
DRRQR	31.4	286035.7	32.1
Grad	31.7	17709.6	31.9

- Every column still applies the same selected indices to both K and Q.
- Joint scoring works best for DRRQR and Grad.
- Query-only magnitude is a much stronger baseline than joint magnitude.

Paper, Table 16 and App. C.7, PDF p. 23. Uncompressed baseline: 29.8.

Capacity should be chosen for the capability you need

1. **A practical compression primitive.**
   Coupled channel selection shrinks state while preserving kernel compatibility.

2. **A useful numerical criterion.**
   DRRQR preserves useful activation directions without task gradients,
   and transfers well across the tested calibration corpora.

3. **A capability-dependent budget.**
   Language modeling tolerates more compression than pure recurrent retrieval;
   batch size determines how much speed is recovered.

Discussion: could per-head budgets and retrieval-aware calibration
preserve the rare directions that long-context recall needs?

Synthesis of Secs. 2–6 and Apps. C/G. The discussion question is a reading-derived research direction.

Paper and source materials

Philipp Nazari and T. Konstantin Rusch
On State Reduction in Linear Attention
arXiv:2602.04852v3, revised 29 September 2026.

arxiv.org/abs/2602.04852v3
github.com/camail-official/LinearAttentionPruning

- All experimental numbers are from this paper version.
- Figures are cropped from the paper; tables select reported entries.
- Page locators refer to the original 35-page PDF.
- The following six slides provide optional technical backup.

Backup: strong RRQR controls the selected spectrum

Let $k = d'_k$, $n = d_k$, $\operatorname{rank}(\mathbf{M}) \geq k$, and $f \geq 1$.

$$\rho_{ij} = \sqrt{|(\mathbf{A}_k^{-1}\mathbf{B}_k)_{ij}|^2 + \left(\frac{\|(\mathbf{C}_k)_{:,j}\|_2}{1/\left\|(\mathbf{A}_k^{-1})_{i,:}\right\|_2} \right)^2}.$$

Swap columns while $\max_{ij} \rho_{ij} > f$. At termination:

$$c_f = \sqrt{1 + f^2 k(n-k)}, \qquad \sigma_i(\mathbf{A}_k) \geq \frac{\sigma_i(\mathbf{M})}{c_f}$$

$$\|\mathbf{C}_k\|_2 \leq c_f \sigma_{k+1}(\mathbf{M}), \qquad \boxed{\kappa(\mathbf{A}_k) \leq c_f \frac{\sigma_1(\mathbf{M})}{\sigma_k(\mathbf{M})}}.$$

Selected original columns share the singular values of $\mathbf{A}_k$.

The guarantee concerns the calibration matrix, not a bound on task loss.

Paper, Algorithm 2 and App. B.2, PDF pp. 15–16; based on Gu & Eisenstat (1996).

Backup: the rank bound and local gradient conditioning

Projected-value proof sketch

$$\mathbf{S} = \mathbf{V}_{\parallel}\mathbf{K}^{\top}, \quad \|\mathbf{S}\|_F \geq \left\|\mathbf{V}_{\parallel}\right\|_F \sigma_{\min}(\mathbf{K}), \quad \|\mathbf{S}\|_2 \leq \left\|\mathbf{V}_{\parallel}\right\|_2 \sigma_{\max}(\mathbf{K}).$$

Divide the squared inequalities to obtain the stable-rank lower bound.

Local query-weight gradient

$$\mathbf{o} = \mathbf{S}\mathbf{W}_Q^{\top}\mathbf{x}, \qquad \mathbf{J} = \frac{\partial \mathbf{o}}{\partial \operatorname{vec}(\mathbf{W}_Q)} = \mathbf{S} \otimes \mathbf{x}^{\top}$$

$$\sigma_i^{+}(\mathbf{J}^{\top}) = \|\mathbf{x}\|_2 \, \sigma_i(\mathbf{S}).$$

For fixed $\mathbf{S}$ of full column rank and $\mathbf{x} \neq 0$, the backward map inherits $\kappa(\mathbf{S})$ on its nonzero singular subspaces.

Paper, App. F.1, PDF p. 25; App. F.4, p. 27. Projected values must be nonzero and keys full row rank for the first bound.

Backup: fixing convolution alone does not rescue PCA

Closest diagonal filters after a square orthogonal rotation:

$$\mathbf{W}' = (\mathbf{T} \odot \mathbf{T})\mathbf{W}.$$

Shared filters commute with rotations; elementwise SiLU generally does not.

DN 370M, 50% reduction	Standard conv	Shared conv
Proper PCA, Wiki PPL ↓	3.2×10^5	155.76
Grad, Wiki PPL ↓	31.75	33.73
Uncompressed, Wiki PPL ↓	29.83	29.55

Keeping high-variance components is insufficient for preserving the full layer.

Paper, Proposition B.1, PDF p. 17; SiLU qualification, p. 18; Table 15, p. 22. Compression results are pre-RFT.

Backup: hybrid robustness still depends on the selector

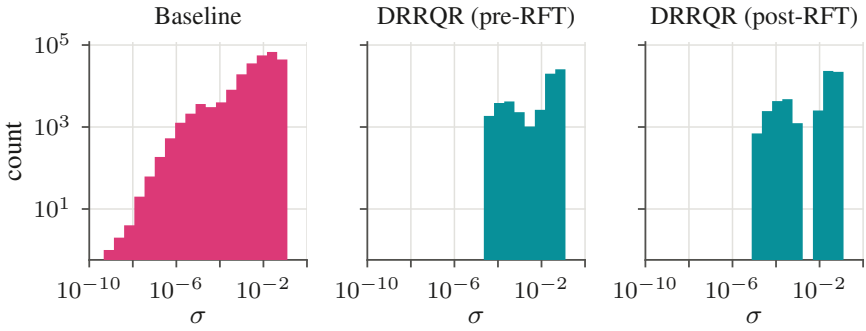
Qwen3.5 9B, before RFT

Configuration	Wiki PPL ↓	LMB PPL ↓	Avg. % ↑
Baseline	9.0	3.8	73.3
DRRQR 25%	9.7	4.3	73.0
DRRQR 50%	11.6	5.7	71.4
L1 25%	18.8	7.1	61.3
L1 50%	38.0	33.4	50.9

After recurrent-only RFT at 50%, DRRQR also retains completion scores:
SQuAD 65.3 $\rightarrow$ 66.0; SWDE 92.7 $\rightarrow$ 92.3.
Softmax attention layers remain frozen throughout recovery.

Paper, Table 13, PDF p. 21; Table 12, p. 20.

Backup: recovery can further reshape the state spectrum



- One DeltaNet 370M head, 75% reduction, FineWeb sequence length 2048.
- Histograms aggregate singular values after the first 128 tokens.
- Compression shortens the tail; RFT can introduce a spectral gap.

Paper, Fig. 6 and App. C.5, PDF pp. 20–22. Both axes are logarithmic; this is a selected-head illustration.

Backup: state size, mixer memory, and model speed differ

GDN 1.3B, H100, 2K prefill, 50% state reduction

Batch	Mixer speedup	Mixer VRAM ratio	Full-model speedup
8	1.26×	0.73×	1.14×
16	1.28×	0.73×	1.14×
32	1.28×	0.72×	1.14×

- Mixer measurements exclude projections, norms, and feed-forward layers.
- A smaller recurrent matrix is only part of peak inference memory.
- Larger decode batches expose more benefit from reducing per-sequence state.

Paper, Tables 18–19 and App. G.1, PDF pp. 27–28. Mixer setup: 16 heads, hidden size 2048, $d_v = 128$.