Don't Read Everything

A Curvature-Conditioned Query for Linear Attention

Paper: Dong Le, Thong Nguyen, Cong-Duy Nguyen, Anh-Tuan Luu

arXiv:2606.01294v2 (30 August 2026)
Reading presentation: 16 September 2026

Memory-aware query contraction improves several retrieval results,
but the training-time mechanism remains unresolved.

The read step is a separate design choice

$$S_t = \mathrm{Update}(S_{t-1}; k_t, v_t, \ldots), \qquad o_t = S_t q_t$$

Write-side changes

Gating controls retention.

Delta updates control overwriting.

CCQ changes the read

Use the distribution of past keys to modify the current query.

Question: Can a fixed-size statistic reduce background interference?

Paper, Secs. 1–3.1, PDF pp. 1–3.

Softmax contributes a useful local geometry

$$A_t(q) = \log \sum_{j \le t} e^{k_j^\top q}, \qquad \nabla^2 A_t(q) = \mathrm{Cov}_{p(q)}[k]$$

$$\nabla^2 A_t(0) = \underbrace{\frac{1}{t} \sum_j k_j k_j^\top - \mu_t \mu_t^\top}_{\Sigma_t}, \qquad \mu_t = \frac{1}{t} \sum_j k_j$$

- At $q = 0$, all keys have probability $1/t$.
- High covariance eigenvalues identify directions of key variation.
- The shared softmax denominator does not change key rankings.

Paper, Eqs. (5), (8), (12); Secs. 3.1–3.3, PDF pp. 2–4; Appendix A, p. 11.

CCQ chooses one gradient step on a quadratic penalty

$$R_t(q) = \tfrac{1}{2} q^\top \Sigma_t q, \qquad \lambda_t = \sigma(W_\lambda q_t + b_\lambda)$$

Clean the query, then read

$$q_t^{\text{clean}} = (I - \lambda_t \Sigma_t) q_t, \qquad o_t = S_t q_t^{\text{clean}}$$

- One covariance–query product; the write recurrence is unchanged.
- This operator is a design choice, not a unique Taylor consequence.
- The identity at $q = 0$ does not bound the remainder at trained queries.

Paper, Eqs. (7)–(10), PDF p. 3; Appendix A.3, pp. 11–12. Normalization made explicit on slide 7.

Retrieval improves only under a geometric condition

$$s_{tj}^{\mathrm{CCQ}} = k_j^\top q_t - \lambda_t k_j^\top \Sigma_t q_t$$

For $\Sigma_t \approx \rho u u^\top$, set $\alpha = k_\star^\top u$ and $\beta = k_d^\top u$.

Change in target-minus-distractor margin

$$\Delta \approx \lambda \rho (\beta - \alpha)(u^\top q)$$

- The sign of $(\beta - \alpha)(u^\top q)$ decides help versus harm.
- A zero margin change need not mean the query is unchanged.
- High variance is not universally equivalent to irrelevant information.

Paper, Eq. (11), p. 3; Appendix H, Eq. (26), p. 16. Boundary qualification: reading analysis.

Unit keys and a bounded gate make the step contractive

$$\|k_j\|_2 = 1 \quad \Longrightarrow \quad 0 \preceq \Sigma_t \preceq \bar{C}_t \preceq I$$

$$\sigma_i(\Sigma_t) \in [0,1], \quad \lambda_t \in (0,1) \quad \Longrightarrow \quad 1 - \lambda_t \sigma_i \in (0,1]$$

- Query norm cannot increase; no eigendirection is annihilated.
- Dividing the accumulated second moment by t is essential.
- This is a bound on the read correction, not a retrieval guarantee.

Teaching example (not an experiment)

$\Sigma = \mathrm{diag}(0.8, 0.1)$, $\lambda = 0.5$: $(1,1)/\sqrt{2} \mapsto (0.6, 0.95)/\sqrt{2}$.

Paper, Sec. 3.3, p. 5; Appendix G, Eq. (23), pp. 15–16.

Inference adds two running statistics per head

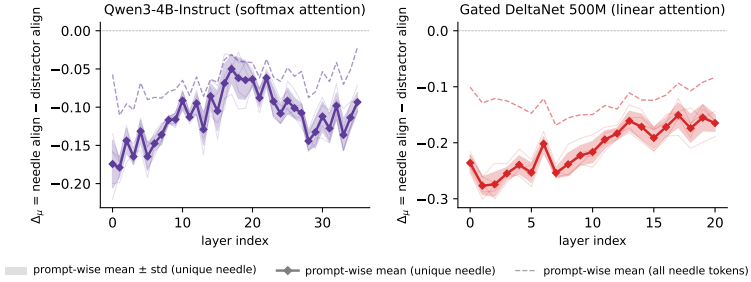
Normalize q_t, k_t to $\bar{q}_t, \bar{k}_t$, then include the current key:

$$\bar{C}_t = \bar{C}_{t-1} + t^{-1}(\bar{k}_t \bar{k}_t^\top - \bar{C}_{t-1}),$$
$$\mu_t = \mu_{t-1} + t^{-1}(\bar{k}_t - \mu_{t-1}),$$
$$q_t^{\text{clean}} = \bar{q}_t - \lambda_t[\bar{C}_t \bar{q}_t - \mu_t(\mu_t^\top \bar{q}_t)].$$

- Gate: linear projection of the pre-split, unnormalized query.
- Cache: $d_v d_k + d_k^2 + d_k$ numbers per head.
- Here $d_k = 256$, $d_v = 512$: about 50% more recurrent state.

Paper, Algorithm 2, p. 13; Appendices D–E, p. 14. Gate starts near 0.01.

The probed needles avoid high-variance subspaces



- Δ_μ = unique-needle alignment minus distractor alignment.
- Negative head averages at 36/36 Qwen layers and 21/21 GDN layers.
- Rare-token filtering and unmatched model pairs limit causal claims.

Paper, Fig. 1, p. 5; Appendix I, pp. 16–20. Solid: unique tokens; dashed: all needle-span tokens.

Training is matched within each model scale

	500M	1.3B
Training tokens	15B	40B
Hidden size	1024	2048
GLA / GDN family layers	21	21
Global batch (tokens)	262,144	524,288

- FineWeb-Edu subset; 32K BPE; 4,096-token training window.
- From scratch, AdamW, bfloat16, one 8-H200 node.
- Controls: Transformer, Mamba2, GLA, GLA-Hedgehog, Gated DeltaNet.

Paper, Sec. 4.1 and Table 1, pp. 5–6; Appendix D and Table 6, p. 14. GDN = Gated DeltaNet.

Four evaluations measure different capabilities

LM + reasoning WikiText / LAMBADA PPL and seven-task zero-shot accuracy.

S-NIAH Pass-key, number, UUID retrieval; 1K–8K. UUID only to 4K.

Length extrapolation Token PPL on six corpora; 4K–20K in 2K steps.

LongBench Fourteen English tasks at 1.3B; task-specific official metrics.

LongBench Avg mixes F1, ROUGE-L, accuracy and code similarity.

Main quality tables do not report multi-seed uncertainty.

Paper, Sec. 4; Appendix F, pp. 14–15. Seven-task Avg excludes LAMBADA.

Language modelling gains depend on the backbone

Scale	Backbone	Wiki PPL ↓	7-task Avg ↑
500M	GLA	27.27 → 26.17	48.54 → 49.73
500M	GDN	26.33 → 25.59	48.98 → 49.87
1.3B	GLA	18.80 → 16.92	55.75 → 56.10
1.3B	GDN	17.98 → 18.01	55.64 → 56.24

- Arrows compare the backbone with its CCQ variant.
- 1.3B GLA: Wiki PPL falls by 1.88 (10% relative).
- 1.3B GDN: average accuracy rises, but Wiki PPL is slightly worse.

Paper, Table 2, p. 7. Avg values use a percentage scale; no significance claim.

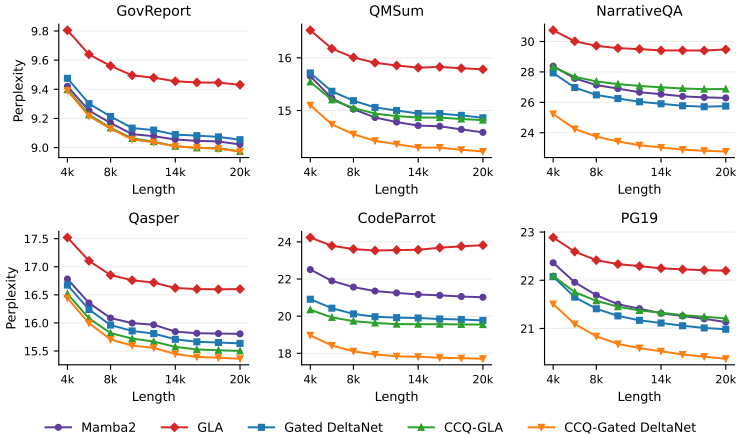
Several retrieval gains are large; others reverse

Scale / backbone	Task / length	Base	CCQ	Δ pp
500M / GLA	pass-key / 2K	26.2	97.6	+71.4
500M / GDN	number / 8K	24.8	50.8	+26.0
1.3B / GLA	pass-key / 4K	47.8	96.8	+49.0
1.3B / GLA	number / 8K	31.2	60.6	+29.4
1.3B / GDN	UUID / 4K	6.0	46.6	+40.6
500M / GLA	UUID / 1K	39.8	11.0	−28.8
1.3B / GDN	pass-key / 8K	99.8	98.2	−1.6

All models fail on UUID beyond 4K; those 8K results are not tabulated.

Paper, Table 3, p. 8. Accuracy in percent; pp = percentage points. Selected gains and counterexamples.

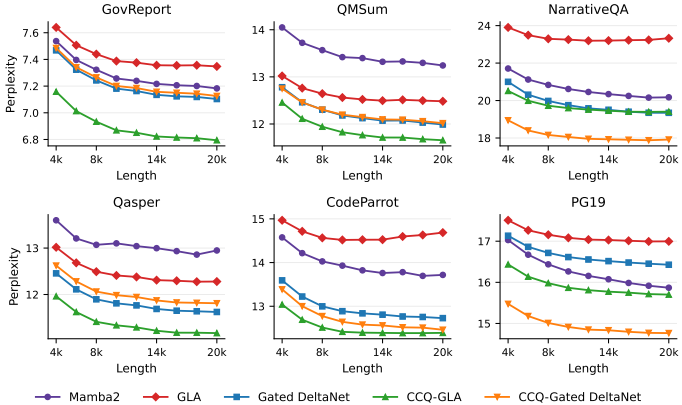
500M: CCQ improves many extrapolation curves



Axes are scaled separately. Transformer and GLA-Hedgehog are omitted as off-range.

Paper, Fig. 3, p. 6. Token-level PPL; lower is better. No numerical values digitized from curves.

1.3B: the figure does not show universal improvement



CCQ-GDN (orange) is slightly worse than GDN (blue) on GovReport, QMSum and Qasper.

Paper, Fig. 4, p. 7. This conflicts with the all-panels improvement claim in Sec. 4.4, p. 6.

LongBench averages improve, with task-level regressions

1.3B model	Avg	Change from backbone
GLA	14.07	—
CCQ-GLA	15.77	+1.70
GDN	13.80	—
CCQ-GDN	14.78	+0.98

- CCQ-GLA: TREC 24.75 → 39.50; GovReport 7.46 → 9.11.
- CCQ-GDN: TriviaQA 21.01 → 28.28, but HotpotQA 5.77 → 3.39.
- RepoBench-P: GLA 26.81 exceeds CCQ-GLA 26.37, despite bolding.

Paper, Table 4, p. 8. Mixed-metric average; best-value markings and prose require checking.

Centring is not necessary for the observed gains

1.3B GDN variant	Wiki PPL ↓	Avg ↑	UUID 4K ↑
No CCQ	17.98	55.64	6.0
Centred, learned gate	18.01	56.24	46.6
Uncentred, learned gate	17.30	56.20	55.6
Centred, fixed $\lambda = 0.1$	17.99	55.85	20.0

- Uncentred second moments win on WikiText and UUID 4K.
- A learned gate beats fixed 0.1 on UUID, but strength is not matched.
- Exact Hessian correspondence is motivation, not established necessity.

Paper, Appendix E, p. 14; Table 7, p. 15. Final bullet: reading interpretation.

Turning CCQ off at inference preserves most retrieval gains

Intervention on the trained CCQ-Gated DeltaNet checkpoint

Disabling the correction changes accuracy by at most:

- 2.2 percentage points: pass-key, 4K–100K;
- 1.6 percentage points: number / UUID, 4K–16K.

- The advantage over a separately trained backbone largely remains.
- This weakens a purely immediate read-time explanation.
- It does not identify the training mechanism or prove all tasks are safe to disable.

Paper, Appendix E, "Inference-time intervention, p. 14. Interpretation distinguished from observation.

Decode throughput is a material implementation cost

1.3B, base → CCQ	GLA	GDN
Train, k tokens/s/GPU	41.6 → 34.8	45.1 → 37.8
Peak training memory, GiB	87.5 → 99.9	81.5 → 93.6
4K prefill, ms	198 → 217	189 → 204
Decode, tokens/s	454 → 298	457 → 297

- About 16% less training throughput and 35% less decode throughput.
- Inference measurements: one H200, batch size 8.
- Recurrent state: 64 to 96 MiB/sequence in the fp32 implementation.

Paper, Appendix C, pp. 13–14; Appendix D, p. 14. Current unfused decode path, not projected performance.

The covariance can disagree with the effective memory

- **Query mismatch:** Σ_t is uniform over past keys; the exact Hessian uses $p(q_t)$ weights.
- **Retention mismatch:** CCQ keeps undecayed statistics while the backbone forgets or overwrites.
- **Task mismatch:** relevant frequent patterns may lie in high-variance directions.
- **Scale limit:** trained only at 500M/1.3B, on GLA and GDN, up to 40B tokens.

The Qwen 4B probe is a diagnostic, not a trained 4B CCQ model.

Paper, Limitations, p. 9; Appendix J, p. 18; Appendix I, pp. 16–18.

What is established, and what should be tested next?

Established within the reported scope

A recurrent query correction with a spectral bound, and several large matched-backbone retrieval improvements after training.

Still unresolved

How much comes from the online correction versus representations learned during CCQ training?

- Next: matched train-on/off and test-on/off experiments.
- Add normalization-only and matched-strength gate controls.
- Measure quality and speed when disabling CCQ at deployment.

Synthesis of Secs. 3–4 and Appendix E. Proposed controls are reading analysis, not completed experiments.

Source and coverage

Dong Le, Thong Nguyen, Cong-Duy Nguyen, Anh-Tuan Luu.

Don't Read Everything: A Curvature-Conditioned Query for Linear Attention.

arXiv:2606.01294v2, revised 30 August 2026.
`https://arxiv.org/abs/2606.01294v2`

- Read the version-matched author TeX and 20-page PDF, including Appendices A–J.
- Figures reproduced from the author's source package; all locators use PDF pages.
- Results are reported evidence, not an independent training reproduction.

Reading date: 16 September 2026. Companion Chinese notes include detailed evidence checks.

Backup: a shape-consistent chunkwise formulation

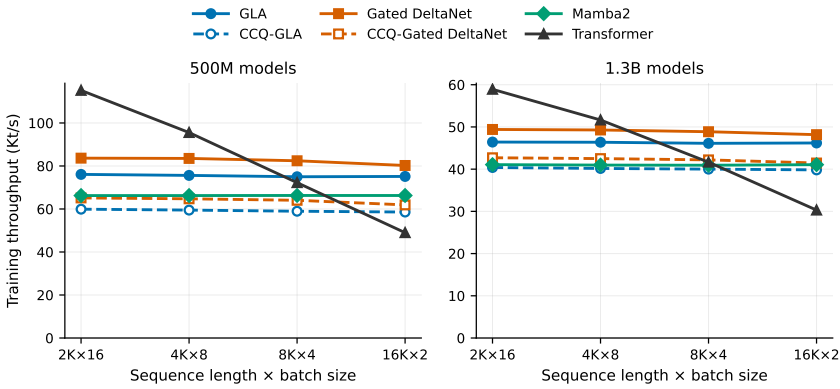
Store queries and keys as rows. Let L be the causal mask and m_i the prefix key sum at position t_i.

$$
\begin{aligned}
P_{\text{prev}} &= \bar{Q} S^{K}_{\text{prev}}, \\
P_{\text{intra}} &= (\bar{Q}\bar{K}^{\top} \odot L)\bar{K}, \\
(\bar{C}q)_i &= (P_{\text{prev}} + P_{\text{intra}})_i / t_i, \\
(\text{mean correction})_i &= m_i (m_i^{\top} \bar{q}_i)/t_i^2.
\end{aligned}
$$

- Update chunk-prefix $\sum kk^{\top}$ and $\sum k$ after the chunk.
- Mean correction is rowwise; avoid cross-token mixing from shorthand.

Paper, Appendix B and Algorithm 1, pp. 12–13. Row-oriented restatement for clarity.

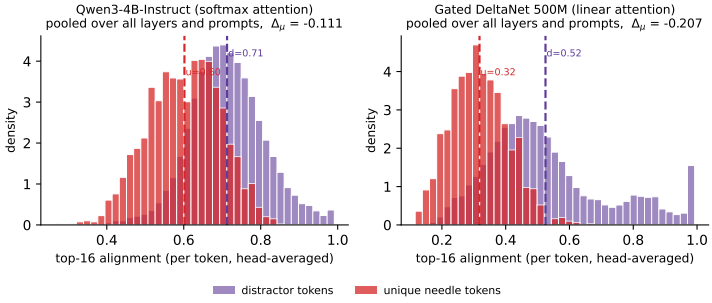
Backup: training throughput preserves linear scaling



32K tokens per step; length and batch size vary together.
Controlled sweep values differ from full-run median throughput.

Paper, Fig. 5, p. 15; full-run medians in Appendix C, p. 13.

Backup: pooled geometry is supportive, not causal



- Red: unique needles; purple: distractors. Lower means less top-16 energy.
- GDN gap is -0.207 when pooled, versus -0.209 in the depth summary.
- Rare-token selection and missing query-sign conditions limit inference.

Paper, Fig. 2, p. 5; Appendix I, pp. 16–18. Last bullet: reading interpretation.

Backup: qualify several claims using the displayed evidence

- **Length PPL:** Fig. 4 contradicts the all-panels 1–2.5 PPL improvement claim for both variants.
- **RepoBench-P:** GLA 26.81 exceeds CCQ-GLA 26.37 in Table 4.
- **Mamba2 comparison:** CCQ-GDN gains $14.78 - 13.13 = 1.65$, not 2.64; 2.64 is CCQ-GLA's gain.
- **Boundary case:** equal target/distractor suppression leaves the margin unchanged, not necessarily the query.

Use numerical entries and explicit equations instead of best-value bolding or broad prose.

Paper, Secs. 3.2, 4.4–4.5, pp. 4, 6–7; Table 4, p. 8; Appendix H, p. 16.