History Replays Itself

Unleashing Long-Context Capability in Linear Attention

NoteBank: preserve text, retrieve, replay

Paper by Li Nanbo, Firas Laakom, et al.
King Abdullah University of Science and Technology

Reading presentation · 2 November 2026
arXiv:2610.15715v1 · submitted 11 October 2026

A small replay budget recovers distant evidence

Central result

Frozen GDN at 1M tokens: **1.7%** $\rightarrow$ **70.3%**
Mean over six RULER retrieval tasks; replay at most 2×64 tokens.

- Preserve selected text before the question is known.
- Use the backbone's features to address it, and its forward pass to read it.
- Distinguish retrieval gains, read adaptation, and state reset.

Publication not found as of 2026-11-02; author lists a preprint.

Code forthcoming: official project page.

Paper, Sec. 4.1 and Fig. 3, PDF p. 6. Lookup date: 2026-11-02.

A recurrent state must compute and retain evidence

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

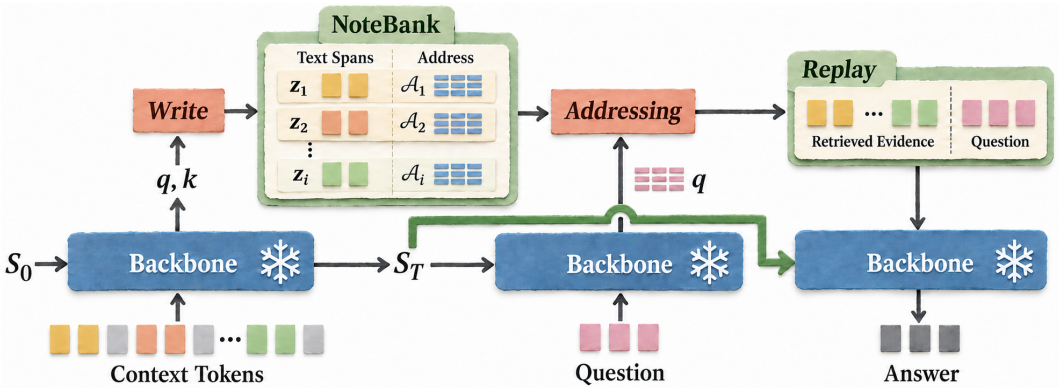
- $\mathbf{q}_t, \mathbf{k}_t \in \mathbb{R}^{d_k}$; $\mathbf{v}_t \in \mathbb{R}^{d_v}$; $\mathbf{S}_t \in \mathbb{R}^{d_v \times d_k}$.
- $\mathbf{A}_t$ controls retention and transformation; β_t controls writing.
- Context arrives first; the question arrives later.

The access bottleneck

Ordinary inference can use history only through the retained state $\mathcal{S}_T$. Evidence can be lost before its relevance is known.

Paper, Sec. 2, Eq. (1), PDF p. 3. Surrounding gates and normalization are omitted.

NoteBank separates computation from text storage



Write without the question → **address** with the question → **replay** through the same backbone.

Paper, Fig. 2, PDF p. 4. Reused author figure. The answer branch starts from a separate state copy.

Writing selects locally distinctive text

$$\overline{\boldsymbol{\mu}}_{\mathcal{C}}^{\ell,h} = \widehat{\frac{1}{|\mathcal{C}|}\sum_{t\in\mathcal{C}} \widehat{\mathbf{k}}_t^{\ell,h}}, \qquad d_t^{\ell,h} = 1 - \langle \widehat{\mathbf{k}}_t^{\ell,h}, \overline{\boldsymbol{\mu}}_{\mathcal{C}(t)}^{\ell,h} \rangle.$$

- Hats denote ℓ_2 normalization; average scores across J channels.
- Score 48-token windows; start with the center token's sentence.
- Expand to adjacent sentences within the note budget B.
- Exclude overlapping sentences before selecting a second note.

Default short-note configuration

$W = 2$ notes/chunk, $B = 64$ tokens/note; $C = 512$ for GDN/GDN-2, $C = 256$ for Mamba-2. No future question is used.

Paper, Eqs. (3), (8), PDF pp. 4, 18; Apps. B.2 and C.3, pp. 17–21.

Keys choose positions; queries become addresses

$$t_i^{*,\ell,h} = \arg\max_{t \in I_i} d_t^{\ell,h}, \qquad \boldsymbol{\rho}_i^{\ell,h} = P^{\ell,h}\mathbf{q}_{t_i^{*,\ell,h}}^{\ell,h}.$$

- I_i: positions in note i; each channel chooses its own token.
- $P^{\ell,h}$ selects 16 fixed random coordinates, without replacement.
- The same coordinate map is used for notes and questions.
- Store original token IDs plus addresses, not historical full KV tensors.

A useful distinction

The routing comparison is **query-to-query** after compaction. Key distinctiveness determines the representative position.

Paper, Eq. (4), PDF p. 4; App. B.3, Eq. (9), PDF p. 19.

Each channel votes for its best matching note

$$\mathbf{u}_\tau^{\ell,h} = P^{\ell,h}\mathbf{q}_\tau^{\ell,h}, \qquad s_i^{\ell,h} = \max_\tau \langle \widehat{\mathbf{u}}_\tau^{\ell,h}, \widehat{\boldsymbol{\rho}}_i^{\ell,h} \rangle$$

$$V_i = \sum_{\ell,h} \mathbf{1}\left[i = \arg\max_j s_j^{\ell,h} \right].$$

- Probe the question from a temporary copy of $\mathcal{S}_T$.
- Each channel finds one winning note; retrieve top-R by votes.
- Ties favor earlier insertion. Replay follows document order.

Paper, Eq. (5), PDF pp. 4–5. Matching casts BF16 features to FP32 and normalizes them.

Replay uses a fresh branch of the retained state

$$\mathcal{S}_{\text{read}} \leftarrow \text{copy}(\mathcal{S}_T), \quad \mathcal{S}_{\mathbf{R}} = F_\phi(\mathcal{S}_{\text{read}}, \mathbf{R}), \quad \mathbf{Y} \sim p_\phi(\cdot \mid \mathcal{S}_{\mathbf{R}}, \mathbf{Q}).$$

1. Discard the question-probing state after retrieval.
2. Replay at most RB original tokens through the full backbone.
3. Process the question again, then generate the answer.

Positions continue from the context boundary; notes are re-encoded at their new positions.

Architecture-specific intervention

Mamba-2 experiments reset recurrent states before replay, retaining convolution caches and positions. We separate this effect later.

Paper, Eq. (6), p. 5; Algorithm 1, p. 18; Apps. B.4/C.3, pp. 19, 21.

Read adaptation leaves writing and routing frozen

NB-TF

No weight updates.

Native writing, routing, replay and generation.

NB-LoRA

Adapters enabled during replay, question processing and generation.

Disabled during context ingestion and retrieval.

$$\Delta W = (\alpha/r)BA, \qquad \mathcal{L} = -\frac{1}{|\widetilde{\mathbf{Y}}|}\sum_{t} \log p_{\phi,\theta}(\widetilde{y}_t \mid \mathcal{S}_{\mathrm{read}}, \mathbf{R}, \mathbf{Q}, \widetilde{\mathbf{Y}}_{<t}).$$

Answer-only loss; 2,400 adaptation examples; fewer than 9.3M tokens.

Paper, App. B.4, Eq. (10), PDF p. 19; App. C.4, p. 21.

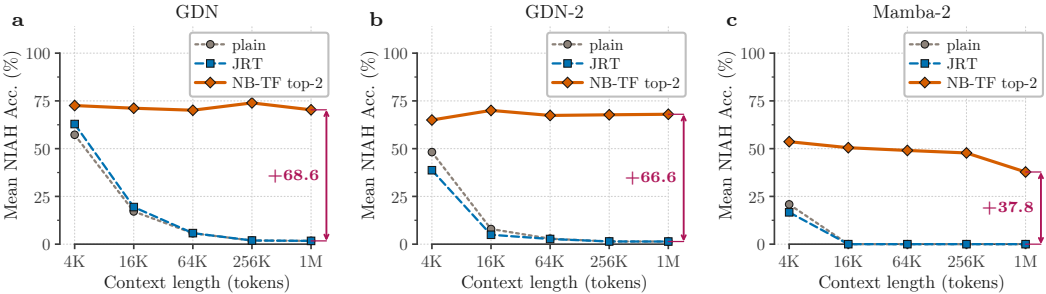
The benchmarks test different kinds of access

Benchmark	Evaluation scope	Replay / metric
RULER	Six tasks; 100 examples each per length; 4K–1M	Top-2; task mean
LongBench v1	12 subsets; 30 examples each; 360 total	GDN top-1; Mamba top-4; macro score
MemoryAgent-Bench	Selected QA tasks; reported 2,100 questions	Top-4; task macro accuracy
LongBench v2	375 full-input QA/dialogue examples	Qwen top-8; accuracy

Within each backbone: same checkpoint and examples. Qwen uses YaRN for both Plain and NoteBank.

Paper, Sec. 4, PDF pp. 5–7; App. C.2, p. 20. MemoryAgentBench count inconsistency: backup slide.

Frozen GDN sustains retrieval from 4K to 1M



At 1M, gains over Plain: **+68.6, +66.6, +37.8 pp**.

Top-2 replay; six-task means. Mamba-2 also uses a readout state reset.

Paper, Fig. 3 and Sec. 4.1, PDF p. 6. Original figure; JRT repeats the full prompt.

Natural-language gains vary by task and reader

LongBench v1, 12-task macro	Plain	NB-TF	NB-LoRA
GDN, top-1	15.49	16.07	23.03
Mamba-2, top-4	6.79	13.43	17.10

- GDN training-free gain: **+0.58 points**.
- LoRA improves QA, but hurts summarization relative to NB-TF.
- Summarization: GDN 13.52 → 11.37; Mamba-2 16.51 → 7.46.

Published Transformer score 16.05 is a contextual reference with a different training protocol.

Paper, Table 1, PDF p. 6; Sec. 4.2, p. 7; Table 5, p. 25. Scores combine task-specific metrics.

Most GDN LongBench improvement comes from adaptation

	RULER @ 1M		LongBench v1	
GDN	Frozen	LoRA	Frozen	LoRA
Without NoteBank	1.7	1.7	15.49	22.72
With NoteBank, top-1	59.8	58.4	16.07	23.03

Two different bottlenecks

RULER: memory access recovers distant evidence.

LongBench: adaptation drives most of the gain; adapted models differ by only **0.31 points**.

Paper, Table 4 and Sec. 6.3, PDF p. 10. Adapters trained separately with matched data and budgets.

A strong hybrid also benefits without training

Qwen3.5-9B	Plain	NB-TF	Gain
MemoryAgentBench, top-4	55.7	56.8	+1.1 pp
LongBench v2, top-8	46.1	49.3	+3.2 pp

- 24 GDN blocks + 8 attention blocks.
- Write and address with GDN features; replay through every block.
- Both methods use YaRN; LongBench v2 uses 375 selected examples.
- On MemoryAgentBench, top-1 matches Plain; top-4 improves it.

Paper, Tables 2/6, PDF pp. 6/25; Apps. C.1/C.2, p. 20.

State reset explains much of Mamba-2-8B recovery

MemoryAgentBench	Overall	Single-hop FactCon.
Plain	0.0	0.0
Clean start, no replay	13.0	14.0
NB-TF, top-1	14.5	19.0
NB-TF, top-4	14.7	22.0

Use the matched control

Top-4 adds **1.7 pp** over clean start, versus 14.7 pp over Plain. The accumulated long-context state can impair readout.

More replay is task-dependent: EventQA drops 17.9 $\rightarrow$ 16.4 from top-1 to top-4.

Paper, Table 6, PDF p. 25; state-reset discussion, pp. 21, 26.

Extra arithmetic is tiny relative to full ingestion

$$\Delta F = F_{\text{write}} + F_{\text{route}} + (|\mathbf{Q}| + |\mathbf{R}|) f_\phi, \qquad |\mathbf{R}| \le RB.$$

$$F_{\text{route}} \approx 2|\mathbf{Q}| N_T J d_\rho + 3(|\mathbf{Q}| + N_T) J d_\rho.$$

Here N_T counts notes, $d_\rho = 16$, and f_ϕ is backbone FLOPs per token.

- GDN at 1M: Plain $\approx$ 2891.4 TFLOPs; extra $\approx$ 0.722 TFLOPs.
- Added ratio: **0.025%**; three backbones span 0.0171%–0.0250%.
- Fixed probing/replay is amortized; writing and exhaustive routing grow with history.

These estimates exclude many non-arithmetic system costs. Hybrid attention needs additional accounting.

Paper, Sec. 5.1, pp. 7–8; App. C.6, Eqs. (12)–(16), PDF pp. 22–23.

Persistent payload is compact, but grows with history

GDN + NoteBank at 1M

- Recurrent state: 48 MiB.
- Conv. cache: 1.125 MiB.
- Archive: 24.959 MiB.
- Helpers: ≈ 0.402 MiB.

Total: ≈ 74.49 **MiB**.

The comparison

$$M_{\mathrm{MHA}} = 4|\mathcal{L}|DP$$

P: input tokens. BF16 full MHA; 24 layers; width 2048; no KV-head sharing.

$$\frac{M_{\mathrm{NB}}}{M_{\mathrm{MHA}}} \approx \frac{1}{2638}.$$

Reference: ≈ 192 GiB.

Excludes weights, workspace, branch copies and Python-object overhead. This is a persistent-payload ratio for the specified GDN configuration.

Paper, Sec. 5.2, PDF p. 8; App. C.7, Eqs. (17)–(21), p. 23.

Measured latency overhead is about 12.4%

GDN at 1M	Plain	NB-TF top-2
End-to-end time	125.11 s	140.64 s
Peak GPU memory	2.93 GiB	3.27 GiB
RULER score (separate quality run)	1.71	70.33

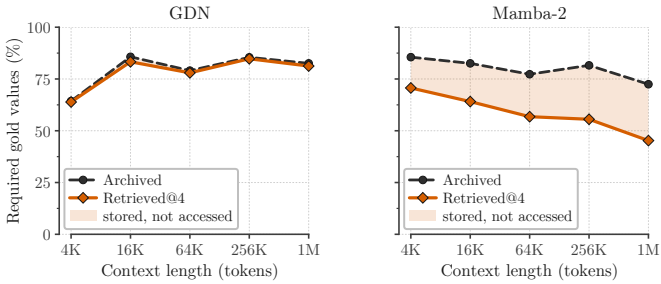
- H100 80GB, BF16, batch 1, exactly 32 generated tokens.
- Archive rebuilt each run; probing, routing and replay included.
- Peak memory includes weights and transient allocations.

Practical interpretation

Very small arithmetic overhead coexists with measurable runtime overhead.

Paper, Sec. 5.3, PDF p. 8; App. C.5, pp. 21–22; Table 7, p. 26.

Stored evidence can remain inaccessible



Oracle routing at 1M

GDN
75.0 → 75.7

Mamba-2
40.4 → 61.6

Archive, reader and replay budget held fixed.

GDN retrieves most archived gold values. Mamba-2 leaves a larger gap between preservation and access.

Paper, Fig. 5a/b and Sec. 6.1, PDF pp. 8–9. Original panel; diagnostics differ from top-2 headline results.

Alternative routers help Mamba-2 on LongBench

LongBench, fixed LoRA reader, top-1	GDN	Mamba-2
Random	21.96	13.83
Mean input embedding	22.40	16.70
Native feature voting	23.03	14.25
BM25	23.50	18.66
BGE	23.77	17.24

Native routing avoids an auxiliary encoder. Its quality is backbone-dependent. BM25 realizes some of the observed routing headroom. Reported timing scopes differ: native 26 ms includes probing; BGE timing includes full-bank encoding.

Paper, Tables 3/8, PDF pp. 10/27. Timing values do not form a matched per-query latency comparison.

The best evidence budget depends on the backbone

- **Write more:** $W = 2$ improves coverage over $W = 1$; storage rises, arithmetic changes little.
- **Write more often:** small chunks help Mamba-2; very coarse $C = 2048$ hurts both backbones.
- **Replay more:** at 1M, $R = 1 \rightarrow 4$ gives GDN 59.5 $\rightarrow$ 75.2, Mamba-2 37.1 $\rightarrow$ 40.8.
- **Replay longer:** longer spans are not uniformly better. The 1M grid peaks at $(B, R) = (128, 4)$.

Natural-language interaction

With LoRA, LongBench peaks at $(256, 2)$ for GDN (24.13) and $(128, 4)$ for Mamba-2 (17.81).

Paper, App. D.2, Figs. 8–12, PDF pp. 27–30. Ablation scores belong to their own configurations.

The contribution is a stronger memory interface

1. **Preserve independently.** A fixed recurrent state need not be the only route to historical evidence.
2. **Reuse native computation.** Frozen features write and route; ordinary forward passes read selected text.
3. **Track the actual bottleneck.** Coverage, routing, read adaptation and state stability contribute differently.

Main trade-off

Sparse query-independent storage can miss evidence. Archive size and exhaustive matching grow with history; the replay budget stays bounded.

Paper, Secs. 3, 6, 7, PDF pp. 3–5, 8–10. Synthesis of reported mechanisms and controls.

Paper and source links

History Replays Itself: Unleashing Long-Context Capability in Linear Attention

Li Nanbo, Firas Laakom, Yilan Zhang, Piotr Piękos, Yimeng Chen, Liangyu Wang, Yifan Xu, Mateusz Ostaszewski, Dandan Guo, Wenyi Wang, Yanning Dai, Yuhui Wang, Jürgen Schmidhuber.

- arXiv:2610.15715v1: the version read in this presentation.
- Official project: code marked *Coming soon*.
- Coauthor publication list: listed as a 2026 preprint.

Publication not found as of 2026-11-02.
All results are from the paper; teaching explanations and calculations are identified where used.

Backup: the writer has explicit boundary rules

- Candidate region includes the preceding chunk's last 32 tokens. Their original features and scores are retained.
- A long initial sentence is clipped to a valid contiguous B-token segment.
- Selecting part of a sentence excludes that entire sentence from the second selection.
- Notes from one write operation do not overlap; notes across chunk boundaries may overlap.
- A final incomplete chunk is processed but writes no note.

Teaching calculation: $W = 2, B = 64, C = 512$ permits up to 25% as many stored token occurrences as input tokens. At 1M, the observed ratio is about 22.2%; replay is at most 128 tokens.

Paper, App. B.2, PDF pp. 17–19; Eq. (18), p. 23. Percentages calculated from stated budgets/counts.

Backup: adaptation is shared across downstream tasks

- Natural Questions-derived short-answer corpus: 2,400 examples, including 678 without supporting evidence.
- Original gold answers are retained for these no-support examples.
- Rank $r = 8$, scale $\alpha = 16$; adapters on four selected layers.
- 600 AdamW updates; four examples accumulated per update; learning rate 5×10^{-4}.
- Answer-only loss including EOS; frozen context state and discrete routing.

Each backbone/budget adapter is evaluated across all 12 LongBench tasks without task-specific fine-tuning.

Paper, Apps. B.4/C.4, PDF pp. 19/21. Weight decay 0.01; gradient clipping 1.0.

Backup: more span tokens can reduce retrieval quality

RULER at 1M: six-task mean, NB-TF

GDN

	$B = 64$	128	256
$R = 1$	59.5	57.2	52.8
$R = 2$	70.3	72.4	71.2
$R = 4$	75.2	**76.8**	74.5

Mamba-2

	$B = 64$	128	256
$R = 1$	37.1	33.5	33.0
$R = 2$	37.8	39.9	35.8
$R = 4$	40.8	**41.3**	37.3

Longer notes and more notes change different aspects of evidence access. The best configuration uses more than the default replay budget.

Redrawn from Paper, Fig. 11, PDF p. 29 (1M column only). Do not substitute these values for other runs.

Backup: LoRA does not remove budget sensitivity

LongBench macro score: NB-LoRA

GDN

	$R = 1$	2	4
$B = 64$	23.26	21.30	20.04
$B = 128$	23.20	22.12	21.19
$B = 256$	22.59	**24.13**	22.86

Mamba-2

	$R = 1$	2	4
$B = 64$	14.26	15.77	16.88
$B = 128$	16.55	17.43	**17.81**
$B = 256$	14.25	16.20	17.18

GDN's score averaged over span budgets falls from 23.02 to 21.36 as R increases from 1 to 4. The full grid reveals an interaction hidden by that average.

Redrawn from Paper, Fig. 12, PDF p. 30. These ablation runs differ slightly from the main table.

Backup: exact payload accounting uses recorded counts

P: input tokens; N_T: notes; U: stored token occurrences.

$$\bar{P} \approx 1{,}048{,}019, \qquad \overline{N_T} \approx 4092.3, \qquad \bar{U} \approx 232{,}514.$$

$$M_{\text{archive}} = 2N_T J d_\rho + 4U + 24N_T \quad \text{bytes.}$$

$$M_{\text{NB}} = M_{\text{archive}} + 4J d_k d_v + 2c_{\text{conv}} J(2d_k + d_v) + M_{\text{helper}}.$$

- BF16 addresses, int32 token IDs, three uint64 fields per note.
- GDN: $J = 192$, $d_k = d_v = 256$, $c_{\text{conv}} = 4$.
- Payload measured at answer start; exclude transient branch copies.

Paper, App. C.7, Eqs. (18)–(21), PDF p. 23.

Backup: preserve the paper's evaluation qualifications

- **MemoryAgentBench count:** the paper reports 2,100 questions, but its 11-configuration enumeration sums to 2,300. Keep reported scores without reconstructing the sample set.
- **Router reader label:** Table 3 calls readers LoRA; its GDN RULER value 59.8 is labeled Frozen in Table 4 (LoRA: 58.4). Avoid cross-table configuration equivalence.
- **Transformer comparison:** the runtime baseline uses GQA and unscaled RoPE beyond a configured 4K range. The 1/2600 payload reference instead assumes full MHA.

Paper, App. C.2, p. 20; Tables 3/4, p. 10; Apps. C.5/C.7, pp. 22–23.

Backup: retrieval quality and evidence use interact

Four LongBench QA tasks	Reader	Native top-1	Oracle-1
GDN	Frozen	7.05	7.59
GDN	LoRA	19.11	22.58
Mamba-2	Frozen	4.22	7.23
Mamba-2	LoRA	9.09	26.43

- Fixed archive, reader and one-note replay budget.
- Oracle chooses an archived note containing a reference answer.
- Answer containment is a relevance proxy, not guaranteed support; these scores are not a strict natural-language upper bound.

Paper, Fig. 5c, PDF p. 9. Tasks: HotpotQA, 2WikiMQA, MuSiQue, Qasper.