How Linear Attention Remembers

Causal evidence for writing, retention, and readout

Paper authors: Kichang Lee, JaeYeon Park,
Songkuk Kim, and JeongGil Ko

arXiv:2609.33093v1 | 27 September 2026

Reading presentation | 8 October 2026

Selective recall can emerge from shared, causally coupled state.

20–30 minute main discussion; slides 21–28 provide additional evidence and details.

Fixed memory changes the organization of history

Full attention	Recurrent linear attention
Past tokens retain explicit key–value entries.	Many past tokens update one matrix per head.
The cache grows with context length.	State size stays fixed as context grows.

- Which writes carry the remembered content?
- How does a query access one fact among many?
- What fails as writes compete, or when KV memory is available?

Paper, Sec. 1, PDF pp. 1–2. This is a mechanism study of pretrained models.

A head transforms old state and adds a new write

$$S_t = \mathcal{T}_t(S_{t-1}) + W_t, \qquad o_t = q_t^\top S_t$$

- $S_t, W_t \in \mathbb{R}^{d_k \times d_v}$; $q_t \in \mathbb{R}^{d_k}$; $o_t \in \mathbb{R}^{d_v}$.
- A common write is $W_t = u_t v_t^\top$: direction u_t, content v_t.
- Conditioned on token representations, $\mathcal{T}_t$ is linear in state.

The update rules differ

GLA uses data-dependent multiplicative gating. GDN also corrects the content associated with the current key direction.

Paper, Sec. 2.1, Eqs. (1)–(2), PDF p. 2. Core read omits output gating and normalization.

Retention and query access are separate operations

$$C_i^{(t)} = (\mathcal{T}_t \circ \cdots \circ \mathcal{T}_{i+1})(W_i)$$

$$\underbrace{S_t = \sum_{i=1}^{t} C_i^{(t)}}_{\text{contributions retained in state}} \qquad \underbrace{o_t = \sum_{i=1}^{t} q_t^\top C_i^{(t)}}_{\text{contributions accessed by the query}}$$

- Exact for $S_0 = 0$ along a fixed forward trajectory.
- A retained contribution need not influence the current output.
- The aggregate state does not uniquely identify its source terms.

Paper, Sec. 2.2, Eqs. (3)–(5), PDF pp. 2–3. End-to-end interventions can change the later trajectory.

Shared storage can support selective access

Teaching example: two writes, no gating

$$S = u_A v_A^\top + u_B v_B^\top, \qquad q = u_A, \qquad \|u_A\| = 1$$

$$q^\top S = v_A^\top + \underbrace{(u_A^\top u_B)}_{\rho} v_B^\top$$

- Orthogonal directions give a selective read of A.
- Overlap introduces a contribution from B.
- Learned gates and downstream computation determine the real effect.

Presenter's illustration of the paper's framework, not a reported experiment.

Paired counterfactuals connect writing to recall

Recipient: Alice lives in **Paris**. Donor: Alice lives in London.

Intervention	Causal question
Block / replace write	Is the fact-time update necessary and content-specific?
Swap state	Is donor information already present before the query?
Patch query output	Can this read pathway transmit the donor effect?
Restore clean output	Does it block the effect of a donor-state swap?

Only a development-selected head is changed in the main single-fact tests.

Paper, Fig. 1, PDF p. 4; Secs. 3.1–3.2. Example follows the intervention schematic.

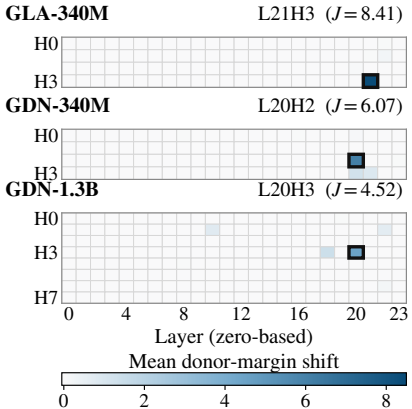
Frozen models answer controlled recall questions

Model label	Recurrent layers	Heads/layer	State/head
GLA-340M	24	4	128×256
GDN-340M	24	4	256×256
GDN-1.3B	24	8	256×256

- WikiText-103 context, controlled associations, four balanced answers.
- Donor/recipient pairs differ in the target value; base weights stay frozen.
- Development and confirmation use separate source passages.
- Additional hybrids: GDN-340M and Qwen3.5-4B/9B.

Paper, Sec. 2.3, PDF p. 3; Appendix A, Table 1, PDF pp. 12–14. Model labels follow released checkpoints.

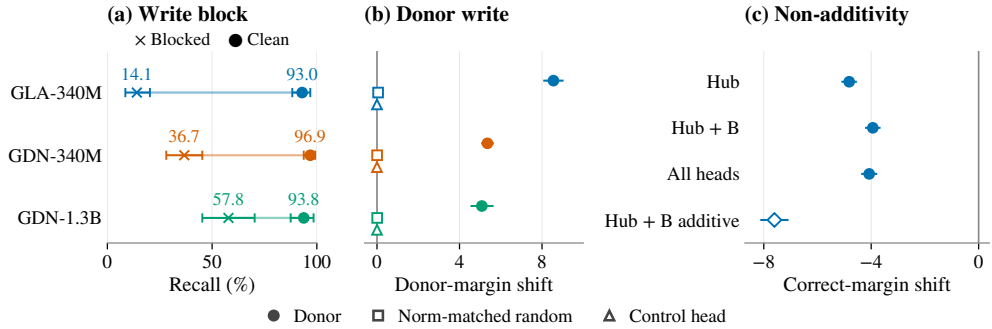
A small set of heads has strong causal effects



- Screen all 96 / 96 / 192 heads on development data.
- $\Delta(z_d - z_r) > 0$ means a shift toward the donor answer.
- Freeze GLA **L21H3**, GDN-340M **L20H2**, and GDN-1.3B **L20H3**.
- Keep those heads fixed for confirmation.

Paper, Fig. 2, PDF p. 4; Appendix B, pp. 14–15. Indices are zero-based; z_d, z_r are donor/recipient logits. See slide 21 for the v1 score notation issue.

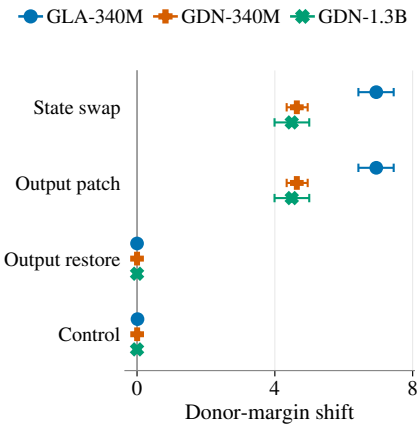
Fact-time writes carry persistent, specific content



- Blocking the selected write lowers recall by **36.0–78.9 pp**.
- Donor writes transfer the donor answer in **68–90%** of cases.
- GLA joint effects are non-additive: the selected head interacts with its circuit.

Paper, Fig. 3, Sec. 3.1, PDF p. 5. Error bars: source-bootstrap 95% CIs; pp = percentage points.

Query-time output mediates the state-swap effect

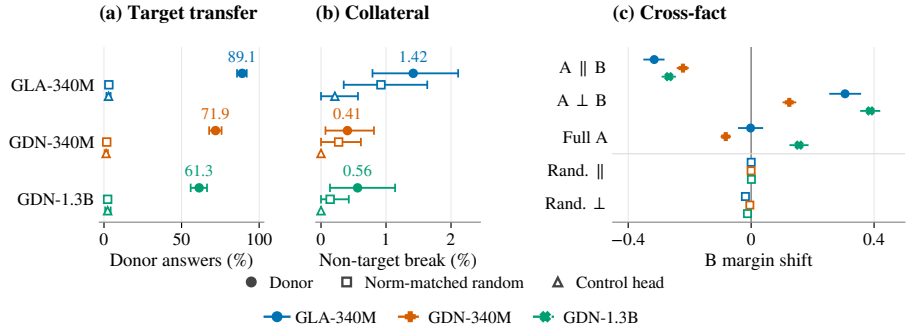


1. Donor state before the query shifts the answer.
2. Donor output alone reproduces a similar shift.
3. Restoring recipient output after the state swap removes the shift.

Stored information affects behavior through the selected read pathway.

Paper, Fig. 4, Sec. 3.2, PDF p. 5. Horizontal axis: donor-versus-recipient margin change.

Four facts remain behaviorally distinguishable



- Donor-answer rates: **89.1** / **71.9** / **61.3%** across the three models.
- Non-target break rates: **1.42** / **0.41** / **0.56%**.
- Break rate: previously correct non-target answers turned wrong.

Paper, Fig. 5, Sec. 3.3, PDF p. 6. Model order: GLA-340M, GDN-340M, GDN-1.3B.

A selective edit contains opposing cross-fact effects

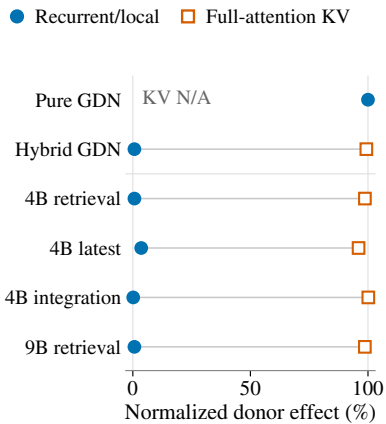
Let Δ_A, Δ_B be donor–recipient state differences for facts A, B.

$$\Delta_{A\|B} = \frac{\langle \Delta_A, \Delta_B \rangle_F}{\|\Delta_B\|_F^2} \Delta_B, \qquad \Delta_{A\perp B} = \Delta_A - \Delta_{A\|B}$$

- Add each component separately, then query fact B.
- Parallel and orthogonal components have opposing effects on B.
- The full edit has a weaker signed effect, consistent with cancellation.
- Norm-matched random directions have much smaller mean effects.

Paper, Sec. 3.3, Fig. 5(c), PDF p. 6. Projection written out from the text; $\Delta_B \neq 0$. Geometric orthogonality does not guarantee behavioral independence.

Hybrid recall is carried mainly by the KV cache

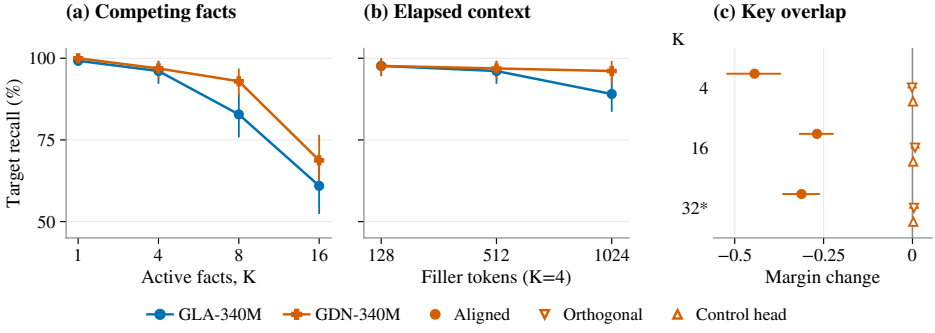


- Near-matched hybrid GDN: recurrent/local $< 1\%$, KV $> 99\%$.
- Qwen3.5-4B: recurrent **0.2–3.6%**, KV **96–100%**.
- Denominator: effect of replacing *both* cache substrates.

This identifies the direct pre-query carrier. Recurrent layers may still form KV contents and queries.

Paper, Sec. 4.1, Fig. 6, PDF p. 7. Separately trained checkpoints; two chance-level clean conditions excluded.

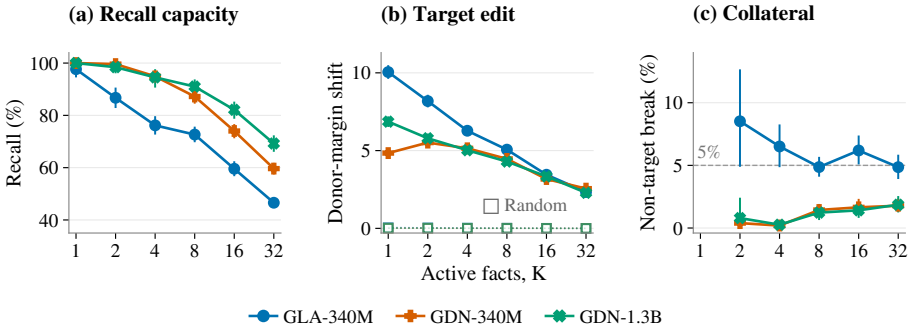
Competing and overlapping writes weaken recall



- Fixed length, $K = 1 \rightarrow 16$: recall falls by about **30–40 pp**.
- Fixed $K = 4$, distance $128 \rightarrow 1024$: a much smaller decline.
- Aligning a later write key lowers target margin by **0.27–0.44 logits**.

Paper, Sec. 4.2, Fig. 7, PDF pp. 7–8. Key norm is preserved. $K = 32$ uses a separate high-load layout.

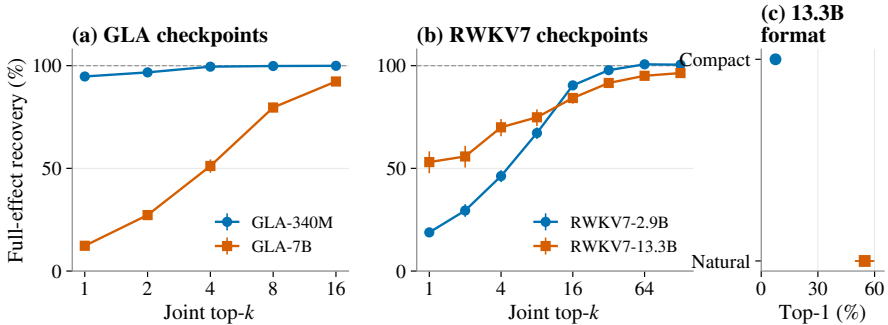
Retention, edit strength, and collateral fail differently



- At $K = 32$, recall is about **47%, 60%, 69%**; edit strength also falls.
- GDN collateral stays below **2%**; GLA is higher, near or above **5%**.
- Low collateral alone is insufficient if the intended edit is already weak.

Paper, Fig. 8, PDF p. 8. GLA points at $K = 8, 32$ are slightly below 5%, qualifying the prose. Missing GLA edit points were not evaluated. Fig. 5 uses a separate comparison.

Single-component dominance depends on the checkpoint



- GLA-340M top-1: **94.7%**; GLA-7B top-1: **12.3%**, top-16: **92.3%**.
- RWKV7-2.9B / 13.3B top-32: **97.8%** / **91.6%**.
- Joint top-k effects are measured directly; there is no monotonic scale trend.

Paper, Appendix E.2, Fig. 14, PDF pp. 19–20. 32 development / 64 test sources; formats differ across model families.

Native errors can retain useful state information

Frozen backbone / condition	Native accuracy	With correction
GLA, colors	24.2%	**52.3%**
GLA, shapes + changed template	30.5%	**49.2%**
GDN, colors	46.9%	48.4%
Qwen3.5-4B recurrent feature	78.9%	46.9%

- A separate correction model combines native logits and state features.
- Trained on source-disjoint data, then frozen for evaluation.
- GLA supports a storage/access distinction; benefits do not transfer uniformly.

Paper, Appendix F.1, Fig. 15, PDF pp. 20–21. This is a four-choice diagnostic, not a general decoding method.

The causal account has concrete experimental boundaries

- **Task competence:** the prose pilot misses an 80% top-1 gate, so its lifecycle interventions are not run.
- **Architecture attribution:** pure and hybrid checkpoints were trained separately; GLA and GDN also differ in state size.
- **Localization:** a high-impact head is an interacting pathway; dominance varies with checkpoint and input format.
- **Practical editing:** donor and neutral counterfactuals require information unavailable to a general memory manager.

Paper, Sec. 6; Appendices A, D–F, PDF pp. 9, 13, 17–21.

Measure how memories compete, survive, and are read

Mechanistic picture

Content-specific writes persist in a shared state and influence answers through learned read pathways. Selectivity coexists with causal coupling.

1. Separate memory load from elapsed context in evaluation.
2. Report retention, intended edit strength, and collateral together.
3. In hybrids, distinguish representation formation from the direct runtime carrier of recalled identity.

Synthesis of Secs. 3–6, PDF pp. 4–9. Write partitioning, refresh, and dynamic recurrent/KV allocation remain design directions.

Paper reference and discussion prompts

Kichang Lee, JaeYeon Park, Songkuk Kim, JeongGil Ko.

How Linear Attention Remembers.
arXiv:2609.33093v1, 27 September 2026.

`https://arxiv.org/abs/2609.33093v1`

- Would the GLA/GDN gap persist with matched state capacity and training?
- Can useful edits be found without an oracle donor execution?
- Which facts should a hybrid preserve in growing KV memory?

All experimental figures in this presentation are reproduced from the version above. The two-write example is a teaching illustration.

Backup: head scoring needs a v1 notation clarification

The prose defines a mean *change* in the donor–recipient margin:

$$m_i = z_{i,d_i} - z_{i,r_i}, \qquad J_{\ell,h}^{\text{prose}} = \frac{1}{16} \sum_i \left(m_i^{(\ell,h)} - m_i^{\text{clean}} \right).$$

- Printed Eq. (6) has adjacent margin factors, with no minus sign.
- The text and Fig. 9 describe a signed margin *shift*; implementation needs clarification.
- The 340M screen replicates earlier selection; the 1.3B screen is prospective.
- Source-wise winners: 16/16, 16/16, 15/16; leave-one-source-out winners are stable.

Paper, Appendix B, Eq. (6), Fig. 9, PDF pp. 14–15. Equation above follows the prose, not the printed formula.

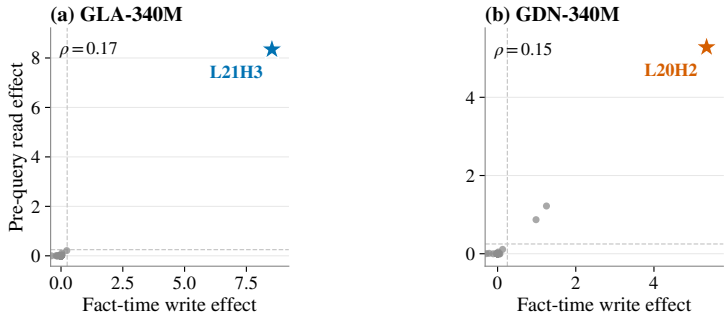
Backup: model names do not imply matched budgets

Model	Actual parameters	Recurrent	Full attention
GLA-340M	341.7M	24	0
GDN-340M	399.5M	24	0
Hybrid GDN-340M	382.6M	16	8
GDN-1.3B	1.466B	24	0
Qwen3.5-4B / 9B	4.206B / 8.954B text	24	8

- Hybrid GDN attention indices: $2, 5, 8, 11, 14, 17, 20, 23$.
- These counts imply 2:1 recurrent/attention; Appendix A calls it 3:1.
- Qwen's 24/8 pattern is 3:1. GLA's state/head has half as many elements as GDN's.

Paper, Appendix A, Table 1, PDF p. 13. Counts and explicit layer list are used here to resolve the ratio inconsistency.

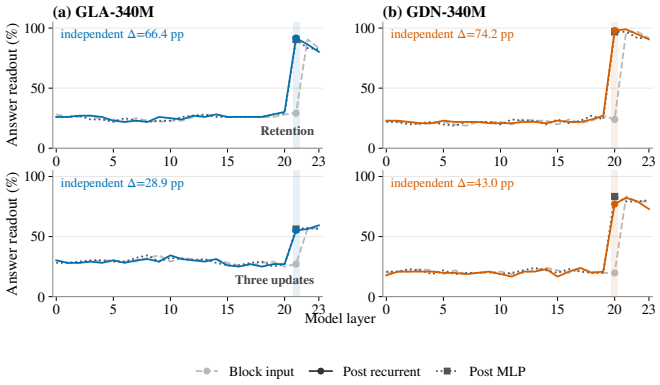
Backup: held-out hubs agree, global rankings do not



- Selected heads rank first at both write and pre-query stages on 32 sources.
- Bootstrap rank intervals: $[1, 1]$ in all four model/stage comparisons.
- Across all 96 heads, write/read rank correlations are only 0.173 and 0.149.

Paper, Appendix C.1, Fig. 10, PDF pp. 15–16. Correlation 95% CIs cross zero for both models.

Backup: answer readout rises sharply at the hub layer



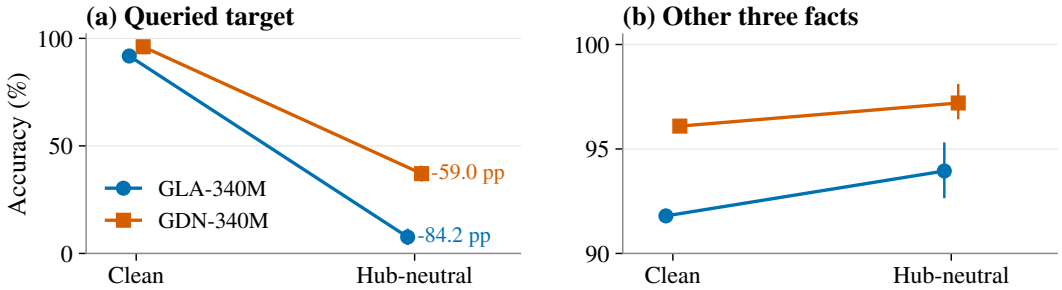
Independent confirmation:

- GLA retention: 28.9% → 95.3%.
- GDN retention: 25.8% → 100%.
- Three updates: +28.9 / +43.0 pp.

Logit lens localizes accessible readout, not the first existence of information.

Paper, Appendix C.2, Fig. 11, PDF pp. 16–17. Curves: 96-source localization; annotations: disjoint 128-source confirmation.

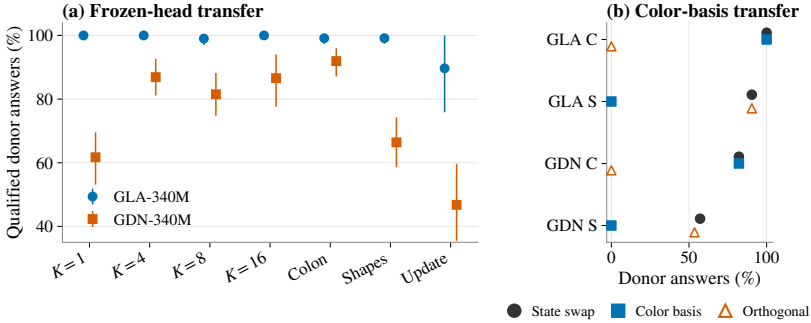
Backup: a neutral state selectively suppresses a fact



- Replace the chosen head with a state from a target-absent execution.
- Target accuracy: GLA **91.8%** → **7.6%**; GDN **96.1%** → **37.1%**.
- Other-fact mean accuracy changes by +2.15 / +1.11 pp, despite some individual disruptions.

Paper, Appendix D.1, Fig. 12, PDF p. 17. 128 sources; oracle counterfactual, not a guaranteed deletion procedure.

Backup: the same head carries task-dependent directions



- Frozen-head effects persist across bindings, templates, values, and updates.
- A rank-4 color basis transfers colors but yields no shape donor answers.
- Most shape transfer lies in the displayed rank-8 orthogonal component.

Paper, Appendix D.2, Fig. 13, PDF p. 18. Panel (a): 128 sources, qualified rates. Panel (b): exploratory 32-source analysis.

Backup: KV dominance survives more natural tasks

Qwen3.5-4B task	n	Clean	Rec. donor	KV donor	KV/full
Entity update, WikiText	128	93.0	0.0	85.2	97.2
Entity update, PG19	128	89.1	0.0	89.8	96.2
RULER single-NIAH, $\sim$4K	64	100.0	0.0	100.0	97.2

- All performance/effect entries are percentages.
- Donor columns measure answer rates; KV/full measures margin-effect recovery.
- Recurrence can still influence how KV contents were formed.

Paper, Appendix E.1, Table 2, PDF p. 19. Source-bootstrap / paired intervention framework follows Sec. 4.1.

Backup: the prose pilot fails a prerequisite for attribution

Filler tokens	Recipient	Donor	Evidence removed
512	56.3%	53.1%	0.0%
128	71.9%	68.8%	3.1%

- Localized evidence sentence; 31 city identities; two fixed query paraphrases.
- 32 source families; joint full-vocabulary top-1 accuracy across the paraphrases.
- Both recipient and donor must reach 80%; neither setting qualifies.
- Lifecycle interventions are therefore not applied in this pilot.

Paper, Appendix F.2, Table 3, PDF p. 21. An unmet competence gate leaves this extension open.