Cortex: Consolidating Context into Parametric Memory

Reading and discussion of arXiv:2610.20399v1

Paper: Jintian Zhang, Feng He, Xinnian Liang, Ziwen Xu, Haoming Xu, Haojing Huang, Shuangzhi Wu, Ningyu Zhang

Research group presentation | 25 October 2026

https://arxiv.org/abs/2610.20399v1

The problem: context helps, then the gain disappears

1. Give model M useful context C.
2. Observe better responses to target queries Q_C^+.
3. Remove C: the benefit often vanishes.

Takeaway

Can a weight update preserve the *context-induced gain* while retaining behavior outside the intended change?

Paper, Secs. 1–2, PDF pp. 1–4.

Crucial scope

A large context gap identifies a *memory failure*: the model can already exploit C. A small gap with poor performance may instead be a capability failure.

Define the recoverable gain and what must be preserved

$$\Delta_{\mathrm{ctx}} = \mathrm{Perf}(M, Q_C^+, C) - \mathrm{Perf}(M, Q_C^+)$$

$$\mathrm{CR} = \frac{\mathrm{Perf}(M', Q_C^+) - \mathrm{Perf}(M, Q_C^+)}{\Delta_{\mathrm{ctx}}}, \qquad \mathrm{RR} = \frac{\mathrm{Perf}(M', Q_C^-)}{\mathrm{Perf}(M, Q_C^-)}.$$

- $M' = \mathcal{A}(M, C)$ answers *without* the source context.
- CR = 100% recovers the measured context gain; RR = 100% retains the measured baseline.
- Target and retain sets may overlap when scoring different response attributes.

Paper, Sec. 2, PDF pp. 3–4. Ratios use unrounded aggregate scores and may exceed 100%.

The testbed spans five different kinds of memory

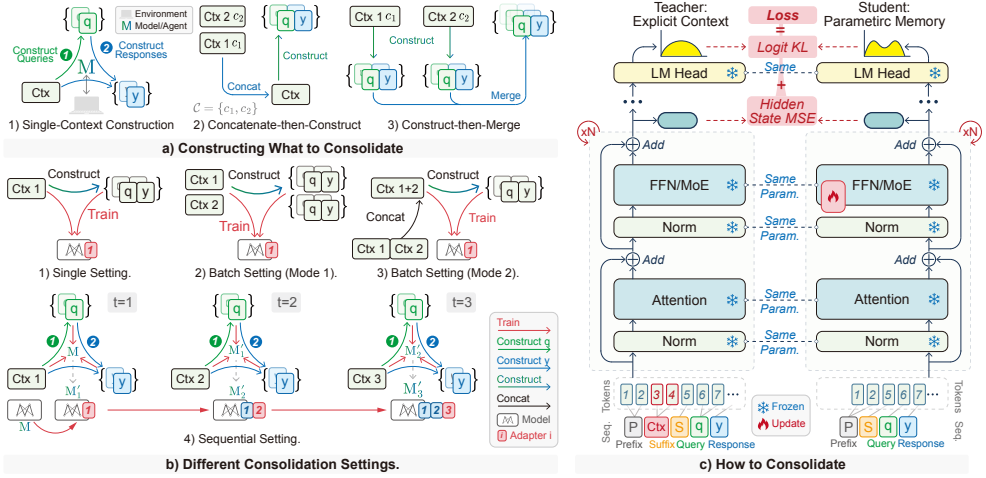
Task	Target change	Retention probe
Agent experience	ALFWorld / τ^2 success	MMLU, SimpleQA, math
New APIs	Tests for 15 Python packages	MMLU, SimpleQA, LiveCodeBench
Persona steering	Power-seeking / Psychopathy tendency	Fluency for Power-seeking only
Knowledge suppression	Less overlap with RWKU target answers	Neighbor facts and BBH
Factual editing	Efficacy, paraphrase, portability	Locality and general benchmarks

Three protocols

Single context; compatible batch of contexts; sequential arrivals with no access to earlier *source contexts*.

Paper, Sec. 4, Fig. 4 / Table 1, PDF pp. 8–10. Coverage differs across task–model–protocol pairs.

Cortex: rehearse uses, then distill into FFN memory



Prospective uses → FFN writes → context-conditioned teacher / context-free student.

Paper, Fig. 2 and Sec. 3, PDF pp. 5–7. Original paper figure.

What to write: prospective rehearsal

$$q_i \sim M(\cdot \mid I_q, C), \qquad y_i \sim M(\cdot \mid C, q_i).$$

- I_q asks the pre-update model to anticipate future uses of C.
- The same model answers with C; an optional filter removes unsuitable examples.
- Agent tasks instead use training-environment queries and successful offline trajectories.
- Batch contexts can be concatenated or rehearsed separately before merging examples.

Takeaway

Rehearsal coverage determines which future uses are exposed to the write procedure.

Paper, Sec. 3.1, PDF p. 5; Appendix B.1, PDF pp. 34–36. Held-out evaluation items are excluded.

Where to write: two mergeable FFN updates

$$F(x) = W_D\,[(W_U x) \odot \mathrm{SiLU}(W_G x)]$$

Synaptic: low-rank changes

$W' = W + B_W A_W$ for W_G, W_U, W_D.
Original matrices frozen; rank $\leq r$.

Structural: new nonlinear channels

$F'(x) = F(x) + D_s[(U_s x) \odot \mathrm{SiLU}(G_s x)]$.
Adds width k; old channels frozen.

At $k = r$ in a dense layer: Structural trains $3kd_h$ matrix weights; Synaptic trains $3r(d_h + d_{\mathrm{ff}})$.
Structural adds deployed FFN width.

Paper, Sec. 3.2 / Fig. 3, PDF p. 6; Appendix A, PDF pp. 30–32.

For MoEs, a shared branch sees every token

$$F'_{\mathrm{MoE}}(x) = \sum_e g_e(x) E_e(x) + E_s(x)$$

- Independent expert expansions receive direct gradient only on selected routes.
- Shared Structural adds one branch E_s across all routed tokens.
- If $\sum_e g_e(x) = 1$, then $F'_{\mathrm{MoE}} = \sum_e g_e(x)(E_e(x) + E_s(x))$.

Deployment detail

Folding the shared branch into each expert is algebraically valid under normalized routing, but copies its storage and executes the new channels on active routes.

Paper, Sec. 3.2, PDF p. 6; Appendix A, PDF p. 33.

How to write: context-to-memory distillation

Teacher M_{t-1}: [prefix, C, suffix, q, y]
Student M_t: [prefix, suffix, q, y]

$$\mathcal{L}_t = \mathbb{E}_{i \in \mathcal{J}} \left[\beta T^2 \mathrm{KL}_{K_v}(p_i^T \| p_i^S) + \frac{\alpha}{L d_h} \sum_{\ell=1}^{L} \| h_{\ell,i}^S - h_{\ell,i}^T \|_2^2 \right].$$

- KL matches next-token behavior; MSE aligns post-FFN residual states.
- Only the current FFN write ψ_t receives gradients.
- Teacher targets store top-K_v logits; student normalizes over the full vocabulary.

Paper, Sec. 3.3, PDF pp. 6–7; Appendix B.3, PDF pp. 48–50.

Sequential writes also need behavioral replay

1. Freeze M_{t-1}, including earlier write modules; generate current-edit rehearsal with C_t.
2. Store earlier *queries*; regenerate their answers from M_{t-1} without old contexts.
3. Include generic questions with targets cached once from M_0.
4. Train only the new FFN adapter and evaluate all still-valid edits without source contexts.

Why replay matters

Frozen old parameters do not freeze the function: later writes can change the hidden states reaching them.

Paper, Sec. 3.3, PDF p. 7; Appendix B.2, PDF p. 39. Up to 100 historical queries are stored.

Evaluation compares three conditions on paired outcomes

Condition	Target input	Role
Vanilla	Query only	Context-free baseline
Context	Query + source C	Measured recoverable gain
Updated model	Query only	CR and RR after consolidation

- Baselines include specialized editing, distillation, hypernetworks, and test-time training.
- Trainable baseline hyperparameters/duration are selected by held-out CR, then retrained and tested.
- Repeated generations are aggregated within each item before task scores and rates.

Paper, Secs. 4–5, PDF pp. 8–11; Appendix B.1, PDF pp. 34–38.

Single and batch: strong recovery is often possible

Setting	CR	RR	Comparison
Qwen2.5 batch API, Synaptic	66%	99%	OPCD: 41% / 100%
Qwen2.5 batch API, Structural	74%	98%	Same 15 packages
Qwen2.5 batch editing, Synaptic	91%	98%	SDCD: 63% / 94%
Qwen2.5 batch editing, Structural	89%	94%	Same 50 edits

Takeaway

At least one Cortex variant is on the empirical CR–RR frontier among context-free methods with both rates in each of Fig. 5's 15 panels.

Paper, Sec. 5.1 / Fig. 5, PDF pp. 10–12; Tables 11, 13, PDF pp. 76, 78. Rounded display values.

Global behavioral steering also persists

Power-seeking, Structural	CR	Fluency RR
Qwen2.5-14B-Instruct	96.9%	99.6%
Gemma3-12B-it	100.3%	99.3%

- The target is an LLM-judged tendency in open-ended answers.
- RR measures *fluency on those same answers*, not broad benchmark retention.
- Psychopathy uses behavior-matching accuracy and has no reported RR; Qwen3-4B Structural reaches 98.0% CR.

Paper, Secs. 4.2 and 5.1, PDF pp. 9–11; Tables 17–18, PDF pp. 82–83.

A high CR can hide serious retention loss

Setting and write	CR	RR
GPT-OSS-20B ALFWorld, Synaptic	100.6%	23.0%
GPT-OSS-20B ALFWorld, Structural	96.6%	95.5%
GPT-OSS-20B suppression, Synaptic	99.5%	75.9%
GPT-OSS-20B suppression, Structural	97.4%	83.3%
GPT-OSS-20B τ^2, Synaptic	−92.3%	98.5%
GPT-OSS-20B τ^2, Structural	38.5%	101.0%

In τ^2, target success is 60.3% Vanilla, 64.7% Context, 56.3% Synaptic, 62.0% Structural. Suppression's retention loss concentrates on nearby knowledge.

Paper, Sec. 5.1, PDF p. 12; Tables 10, 19, 20, PDF pp. 75, 83–84.

Continual editing: a controlled 50-update sequence

Training per edit	Evaluation
Qwen2.5-14B-Instruct; fixed sequence of 50 zsRE facts; 2 current QA pairs; up to 50 replay queries; 16 generic retention examples; 30 epochs; KL + MSE.	All seen facts after every edit; 7 generations per fact/facet. Fixed MMLU, SimpleQA, GSM8K subsets every five edits. Three training seeds, same edit order.

Target score: equal mean of efficacy, paraphrase, portability.
Retention score: equal mean of locality, MMLU, SimpleQA, GSM8K.

Paper, Sec. 5.2, PDF pp. 12–14; Appendix B.2, PDF pp. 39–41.

After 50 edits, Structural recovers 88.9% of the gap

Method	Final CR	Final RR	Portability	Added/edited per step
ROME	28.5%	94.4%	39.0%	70.78M
Cortex-Synaptic ($r = 1$)	80.6%	97.2%	30.7%	2.73M
Cortex-Structural ($k = 1$)	88.9%	99.8%	49.4%	0.74M

- Context-free target average: 8.76%; with context: 88.10%; Structural after 50 edits: 79.3%.
- After 50 edits, cumulative added/edited weights: Structural 36.86M; Synaptic 136.40M.

Paper, Fig. 6, PDF p. 13; Table 14, PDF p. 79. Means over three seeds; Structural CR SD 3.5 points.

Read the raw scores behind a 99.8% retention rate

	Locality	MMLU	SimpleQA	GSM8K
Vanilla	40.3%	80.9%	4.6%	94.2%
Structural, edit 50	38.3%	81.7%	4.8%	94.7%

- RR compares equal-weight means: 54.9% / 55.0% $\approx$ 99.8%.
- Locality still falls by 2.0 percentage points.
- Direct fact efficacy reaches 95.3%, but compositional portability reaches 49.4%.

Takeaway

The aggregate rate summarizes the measured suite; it does not imply every neighboring fact is unchanged.

Paper, Fig. 6a, PDF p. 13; Table 14, PDF p. 79.

Final performance combines acquisition and preservation

	CR at acquisition	CR at edit 50	Portability then	Portability at 50
Structural	91.3%	88.9%	52.4%	49.4%
Synaptic	85.8%	80.6%	36.1%	30.7%
WISE	52.0%	28.0%	–	–

- Compare each of the same 50 facts immediately after its own edit and at checkpoint 50.
- Structural loses only 2.4 CR points; its portability lead is already visible at acquisition.
- Separately, the first ten facts retain 97.2% mean CR after 40 later edits.

Paper, Fig. 6c, PDF p. 13; Appendix B.2, PDF pp. 40–41.

Historical-query replay protects old edits

Age-aligned ablation on the first 25 facts

After each fact has undergone 25 later edits, original-question + paraphrase accuracy stays above 96% *with* replay.

At age 25	With replay	Without replay
Structural	$> 96\%$	40.5%
Synaptic	$> 96\%$	8.0%

The no-replay control still keeps current-edit examples and 16 generic M_0 examples per update. A fixed historical QA control ends within one CR point of query replay.

Paper, Sec. 5.3 / Fig. 7e, PDF pp. 14–15; Appendix B.3, PDF pp. 58–61.

The data and the FFN location both matter

- **Same data:** on Qwen2.5 batch API, Structural exceeds LoRA-KL by 23.0 CR points.
- **Attention updates:** in continual editing at rank 4 with KL, FFN-only beats FFN + attention by 3.8 CR and 4.0 RR points.
- **Hidden-state MSE:** at Synaptic rank 4, adding MSE raises RR by 2.7 points while lowering CR by 4.0; Structural width 16 improves both.

Takeaway

The rehearsal data, supervision signal, and update parameterization jointly shape the trade-off.

Paper, Sec. 5.3 / Fig. 7a–g, PDF pp. 14–15; Appendix B.3, PDF pp. 50–63.

Shared MoE capacity and compact writes are effective here

- **Equal total expansion weights:** shared MoE branch gains 4.6 CR points on ALFWorld and 10.3 on API code versus independent branches, at about one point lower RR.
- **Equal active expansion weights:** shared remains within two CR points while training one eighth as many expansion weights.
- **Exposure:** every token trains the shared branch; 32 of 768 layer–expert branches receive under 1% of mean exposure in every ALFWorld seed.
- **Capacity sweep:** the smallest tested $k = 1$ and $r = 1$ have the highest mean CR under the common training recipe.

Paper, Sec. 5.3 / Fig. 7h–n, PDF pp. 15–16; Appendix B.3, PDF pp. 64–69.

The appendix explains capacity on finite states

FFN-local correction

At a fixed layer, align student state with the teacher by learning $R(u_i) = \delta_i = h_i^T - \bar{h}_i$ at the observed states.

$$R(u) = D_s \phi_s(u), \qquad \min_{D_s} \|D_s \Phi - \Delta\|_F^2 = \|\Delta(I - \Phi^\dagger \Phi)\|_F^2.$$

- Exact fit with fixed features iff $\mathrm{row}(\Delta) \subseteq \mathrm{row}(\Phi)$.
- With trainable Structural features and $k \geq n$, the appendix gives sufficient conditions to interpolate any n finite states.
- The experiment's $k = 1$ relies on learned approximation and rehearsal coverage beyond this guarantee.

Paper, Appendix A, PDF pp. 30–32. Finite-set result does not establish unseen-query generalization.

Interpret the evidence within its measured scope

- The 50-edit sequence uses one fact order and three training seeds; cross-task continual consolidation remains future work.
- Sequential consolidation stores historical *queries* and uses M_0 generic targets; it does not revisit old source contexts.
- Near-perfect aggregate RR coexists with a locality decrease; suppression has a larger neighbor-knowledge cost.
- Structural writes add capacity at every update, and distillation still computes through a large backbone for many optimization steps.

Paper, Secs. 5–8, PDF pp. 12–20; Appendix B.2, PDF p. 39.

Take-home messages

1. **A precise target:** recover the measurable context gap and retain specified non-target behavior.
2. **A practical write path:** prospective rehearsal + context-conditioned self-teacher + mergeable FFN updates.
3. **Strongest evidence:** 50 sequential factual edits, 88.9% CR and 99.8% RR for Structural with 0.74M parameters per edit.
4. **The remaining challenge:** protect local knowledge and improve compositional portability as memories accumulate.

Paper: Zhang et al., *Cortex: Consolidating Context into Parametric Memory*, arXiv:2610.20399v1.