BEYOND ENDPOINT SCORES: WHEN CONTINUAL KNOWLEDGE-UPDATING RANKINGS REVERSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual knowledge-updating methods are usually compared at a single final checkpoint. On a retrospective 24-update factual stream we show that this reporting choice, the initialization of the replay baseline, and the query form can each reverse the outcome of one and the same comparison. The testbed pairs a fast/slow LoRA hierarchy—a slow adapter re-fitted every K updates plus a small monthly adapter—against cumulative LoRA replay on Llama-3.2-1B and Qwen2.5-1.5B. *Baseline:* carrying replay's adapter weights across updates raises held-out paraphrase retention by 12.8 points at rank 8 and turns the hierarchy's $+7.3$-point full-history lead into a -5.4-point deficit, while lowering unchanged-fact accuracy by 12–19 points. *Window:* because the final update is a consolidation state for every tested K, the endpoint favors the hierarchy by 9–11 points even where warm replay leads by 5–18 points on average; the reversal appears in 16 of 18 seed-level paraphrase comparisons. *Query form:* training-form queries sit near ceiling under warm replay and cannot discriminate. An exact decomposition of reporting error into within-cycle phase and between-cycle drift explains why averaging the last cycle still selects the wrong winner in four of 18 cases. A retrospective drift check passes in seven of 12 aggregates and withholds the one wrong last-cycle sign, and spreading the same number of checkpoints over the stream removes every sign error on the warm panel. We distill these findings into a reporting protocol for continual-updater comparisons.

1 INTRODUCTION

Consider an assistant whose factual store is refreshed monthly with records such as an athlete's current team. One design learns each month's records in a small adapter and periodically consolidates the accumulated history into a larger one; another replays the whole history after every update. Users query the assistant throughout the interval, not only right after a consolidation, and they phrase questions in their own words rather than in the training wording. A benchmark that scores only the final consolidated checkpoint answers a narrower question—*which model is better at that state*—and, as we show, can select the opposite updater from the one that answered more queries correctly over the stream.

This paper is a controlled audit of that comparison rather than a proposal of a new updater. We fix a two-level LoRA hierarchy as the consolidating method and pair it with cumulative LoRA replay on a 24-update Wikidata stream. We then vary three things a benchmark designer chooses: the initialization of the replay baseline (fresh refit versus weight carry-over), the reporting window (endpoint, last cycle, or the full stream), and the query form (training wording versus a held-out paraphrase). Each choice changes the answer.

Our findings are concrete (Figure 1). On Llama-3.2-1B, the hierarchy's average paraphrase retention exceeds fresh rank-8 replay by 7.33 points. Initializing replay from its previous adapter—same data, same per-update recipe, optimizer reset—lifts replay's retention by 12.76 points and puts it ahead by 5.42 points on average, at a cost of 12–19 points of unchanged-fact accuracy. Yet at the final checkpoint the hierarchy leads that stronger baseline by 9.44 points. The reason is structural: the final update is a consolidation state for every tested period, so the endpoint samples the hierarchy at its

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

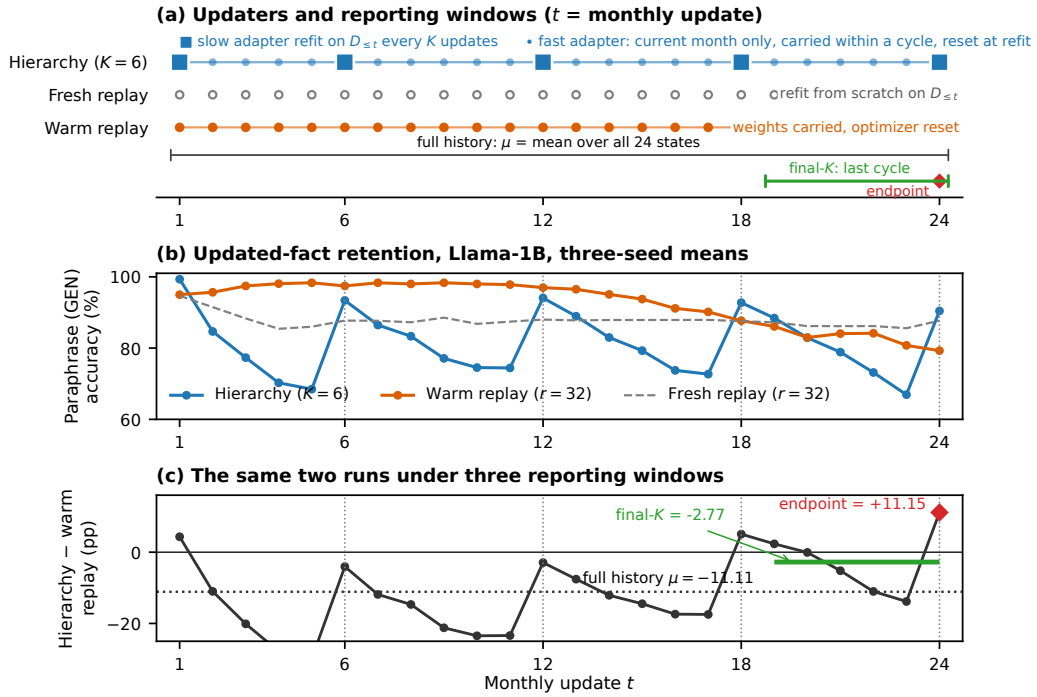

Figure 1: **One comparison, three answers.** (a) The three updaters and the three reporting windows. (b) Held-out paraphrase accuracy on updated facts for the hierarchy (K=6) and rank-32 warm replay, Llama-1B, three-seed means; dotted lines mark slow-adapter refits. The hierarchy peaks at every refit and loses 20–25 points within each cycle; warm replay peaks near 98% and declines to 79% by update 24. (c) The paired margin of the same two runs: the full-history mean favors warm replay by 11.11 pp, the last-cycle mean by 2.77 pp, and the endpoint favors the hierarchy by 11.15 pp.

within-cycle peak, 20–25 points above its trough, while the warm baseline has drifted downward late in the stream. Averaging the last cycle removes the first effect but not the second: it still selects the wrong winner in four of 18 paraphrase comparisons. On training-form queries, where warm replay scores 98–100% at almost every update, no reversal is possible.

Contributions. **(1) A stronger-baseline audit.** Six warm-replay trajectories (two ranks, three seeds) extend a fresh-replay rank sweep on two model families. Weight carry-over improves updated-fact retention in every measurement, costs unchanged-fact accuracy, and leaves paraphrase endpoint reversals in 16 of 18 seed-level comparisons; the training-form ceiling that prevents reversals there is made explicit. **(2) A structural account of endpoint bias.** The last update is consolidation-aligned for every K; an endpoint sweep over all 24 candidate reporting months separates this alignment from late-stream level changes. **(3) Phase, drift, and what to do about them.** An exact phase–drift decomposition with a recency-weighted drift identity and a retrospective drift check on all 12 warm aggregates. A comparison of sparse reporting rules at equal checkpoint budget then shows that spreading checkpoints across the stream leaves no sign error in the 36 warm-panel cells, whereas sampling only consolidation states reproduces most of the endpoint's 16.

A preliminary report by Choi (2026) presented the fresh-replay rank sweep on this stream (Table 2); the warm panel, the preservation accounting, the endpoint sweep, the decomposition, and the reporting-rule comparison are new here. Temporary forgetting between task boundaries (De Lange et al., 2023) and evaluation-dependent editor rankings (Pohl et al., 2025) have precedents; the contribution is their controlled conjunction in one updater decision, with the mechanism behind the reversal identified.

2 SETUP: STREAM, UPDATERS, AND REPORTING TARGETS

2.1 STREAM AND QUERY FORMS

The stream contains 24 event-time buckets, January 2024 through December 2025, built retrospectively from Wikidata statements with temporal qualifiers; team membership (property P54) dominates. Up to 200 records are selected per bucket (21 buckets contain 200; the others 198, 194, and 147), for 4,739 training records. These are event-time records extracted later, not 24 archived snapshots; pretraining overlap has not been measured, and a no-adapter baseline is the first item of future work (Section 6).

Each fact has a canonical training sentence, a training-form completion prompt (MEM), and one held-out paraphrase prompt (GEN) that asks for the *same fact* in different wording. GEN therefore measures access to a trained fact through another phrasing, not generalization to unseen facts. The score ranks the gold answer against three same-relation distractors by mean answer-token log-likelihood (chance 25%), on up to 100 facts per monthly cohort. A never-trained set of 300 unchanged facts, scored the same way, measures preservation. An auxiliary free-generation probe with a different grader is confined to Appendix D.

2.2 THREE UPDATERS

The backbone stays frozen. Let $D_{\leq t}$ be all selected records through update t. **Fresh replay** fits a newly initialized LoRA adapter (Hu et al., 2021) of rank r to $D_{\leq t}$ at every update. **Warm replay** fits the same cumulative data but initializes the adapter from update $t-1$; AdamW state and the learning-rate schedule are recreated each month in both variants, so warm start carries *weights*, not optimizer moments. The **hierarchy** H keeps a rank-64 slow adapter and a rank-8 fast adapter. The slow adapter is freshly fitted to $D_{\leq t}$ at the first update and every K updates thereafter; the fast adapter learns the current month's records on top of the fixed slow adapter, is carried within a cycle, and is reset at each refit. Inference composes the two. This is a fast/slow testbed in the spirit of Progress & Compress (Schwarz et al., 2018), not a proposed method.

All variants use one fixed recipe (six epochs, learning rate 3×10^{-4}, $\alpha/r = 2$; Appendix B); nothing is validation-tuned. The main sweep varies replay rank $r \in \{8, 16, 32, 64, 72\}$ and hierarchy period $K \in \{3, 6, 12\}$; the warm extension fixes $r \in \{8, 32\}$ with seeds $0, 1, 2$ on Llama-1B. Because the hierarchy replays the full history only every K updates, its idealized training exposure at $K = 6$ is about 28% of cumulative replay's in record-epochs (Appendix B); we make no wall-clock claim.

The endpoint is consolidation-aligned by design. $T = 24$ is divisible by every tested K, so the final update always follows a complete slow-adapter refit on $D_{\leq 24}$ with a freshly reset fast adapter. For a fixed seed that state is identical across K, which is why the hierarchy's endpoint numbers repeat across periods in every table below. At the endpoint the hierarchy is, in effect, fresh rank-64 replay plus one month of fast adaptation; its endpoint margin against fresh rank-64 replay is $+0.14$ pp (GEN) and -0.47 pp (MEM).

2.3 REPORTING TARGETS AND WINDOWS

Let $A_M(t, j, s)$ be method M's accuracy after update t on facts from cohort $j \leq t$ under query form s. The benchmark score at state t, the paired margin, and the full-history target are

$$R_M(t, s) = \frac{1}{t} \sum_{j=1}^{t} A_M(t, j, s), \qquad m_t = R_{\mathrm{H}}(t, s) - R_{\mathrm{R}}(t, s), \qquad \mu = \frac{1}{T} \sum_{t=1}^{T} m_t, \tag{1}$$

with the dependence on rank, initialization, period, and seed suppressed; positive margins favor the hierarchy. Equation 1 weights update states equally and, within a state, observed cohorts equally. It approximates a service queried throughout the stream under that cohort mixture; a final-only deployment legitimately targets $R_M(T, s)$ instead, and other traffic distributions define other targets.

Three sparse reports of μ recur below: the *endpoint* m_T, the *final-K* mean of the last K states, and (Section 4) evenly spread states. An *endpoint reversal* means μ and m_T have opposite signs; with $\delta = m_T - \mu$ this is exactly $\mu(\mu + \delta) < 0$. Reference margins with $|\mu| < 0.005$ are ties and excluded

Table 1: **Each evaluation choice reverses the same comparison.** Hierarchy minus replay on held-out paraphrases (GEN), Llama-1B, three-seed means, in percentage points. Positive favors the hierarchy. Details in the sections indicated.

Evaluation choice (other settings fixed)	A vs. B	A	B	Where
Replay initialization (r=8, K=6, full history)	fresh / warm	$+7.33$	-5.42	§3.2
Reporting window (r=32, K=6, warm)	full history / endpoint	-11.11	$+11.15$	§3.3
Query form (r=32, K=6, warm, endpoint)	GEN / MEM	$+11.15$	-1.57	§3.3

Table 2: **Rank sweep against fresh replay.** Hierarchy minus fresh replay on MEM, in percentage points, $K = 6$. Average is the full-history margin; Endpoint is m_{24}. Seeds are Llama/Qwen. Qwen rank 64 uses its completed three-seed extension.

	Llama-1B		Qwen-1.5B		
r	Average	Endpoint	Average	Endpoint	Seeds
8	+3.96	+13.13	+5.02	+17.36	3/3
16	-2.53	+6.67	-1.92	+10.22	3/3
32	-6.75	+1.92	-7.98	+3.48	3/3
64	-9.02	-0.47	-11.31	+0.04	3/3
72	-9.32	-0.40	-11.62	-0.25	3/3

from counts; a non-tie estimate is classified by its raw sign. This is a reporting convention, not a significance test, and estimate-side band sensitivity is reported alongside. Cells across K, seeds, and query forms share trained runs and are correlated, so counts are descriptive coverage, not independent trials.

3 Empirical audit: rank, initialization, and window

3.1 Replay rank moves the crossover

The fresh-replay panel has 44 aggregate cells (30 Llama-1B, 10 Qwen-1.5B, four Llama-3B; only the Llama-3B rank-72 cells are one-seed). Table 2 shows the $K = 6$ training-form slice. On both families the full-history winner changes between ranks 8 and 16, whereas the endpoint winner changes only between ranks 32 and 64 (Llama) or 64 and 72 (Qwen). In the bracket between, the two windows disagree: at Llama rank 16 replay leads by 2.53 points on average while the endpoint favors the hierarchy by 6.67; Qwen rank-64 paraphrases favor replay by 12.20 points on average and the hierarchy by 3.19 at the endpoint. Across the panel, endpoint and full-history winners disagree in 18 of 43 non-tie cells (14 of 43 if endpoint margins below $0.5\,\mathrm{pp}$ are also treated as ties), and the last-cycle mean disagrees in none. Rank changes parameterization and optimization, so the sweep does not identify a storage threshold (Back et al., 2026; Jiang et al., 2024); it does raise the question the next experiment answers: is restarting replay's adapter every month what makes the low-rank baseline weak?

3.2 Weight carry-over strengthens retention and costs preservation

Six warm trajectories (ranks 8 and 32, seeds 0–2) reuse the fresh runs' data and per-update recipe. Warm initialization improves updated-fact retention in all 12 rank–seed–query-form measurements: by $14.35 \pm 0.25\,\mathrm{pp}$ (MEM) and $12.76 \pm 3.25\,\mathrm{pp}$ (GEN) at rank 8, and by $3.93 \pm 0.13\,\mathrm{pp}$ and $4.83 \pm 1.96\,\mathrm{pp}$ at rank 32 (Table 3). The rank-8 gain alone shows that rank does not fix this baseline's achievable retention. Against the hierarchy, warm replay wins 11 of 12 three-seed full-history aggregates; four non-tie winners switch from the hierarchy to replay, and the only remaining hierarchy advantage is rank 8, $K = 3$, GEN ($+1.09\,\mathrm{pp}$; Appendix C).

The gain is not free. Unchanged-fact accuracy falls by 11.7–18.8 points; at rank 32 the GEN figure drops from 69.23% (fresh) to 50.43% (warm), below the hierarchy's 60.36%. The trajectories show where the loss accrues (Figure 1b). Rank-32 warm replay's paraphrase retention peaks near 98.3% in updates 5–9 and ends at 79.3%, while fresh replay stays near 88%; over the same span warm

Table 3: **Warm replay is a stronger retention baseline, not an overall winner.** Full-history accuracy (%), mean ± sample standard deviation over three seeds, Llama-1B. The four columns are separate objectives and are not combined.

	Updated facts		Unchanged facts	
Method	MEM	GEN	MEM	GEN
Fresh replay, $r = 8$	84.59 ± 0.22	74.11 ± 0.49	73.92 ± 0.23	71.29 ± 0.19
Warm replay, $r = 8$	98.95 ± 0.21	86.86 ± 3.14	62.25 ± 0.60	57.98 ± 1.89
Fresh replay, $r = 32$	95.30 ± 0.15	87.72 ± 0.78	72.61 ± 0.38	69.23 ± 0.81
Warm replay, $r = 32$	99.23 ± 0.11	92.55 ± 2.43	55.99 ± 1.40	50.43 ± 1.28
Hierarchy, $K = 6$	88.55 ± 0.79	81.44 ± 0.77	64.60 ± 0.67	60.36 ± 0.49

Table 4: **Held-out paraphrases against warm replay.** Hierarchy minus replay, in percentage points, three-seed means. "End flips" counts seeds whose endpoint sign opposes their non-tie full-history sign. Endpoint values repeat across K because the endpoint state is consolidation-aligned (Section 2).

r	K	Full history	Endpoint	Final-K	End flips
8	3	+1.09	+9.44	+2.54	1/3
8	6	-5.42	+9.44	-4.69	3/3
8	12	-12.05	+9.44	-10.53	3/3
32	3	-4.60	+11.15	+4.29	3/3
32	6	-11.11	+11.15	-2.77	3/3
32	12	-17.73	+11.15	-12.37	3/3

replay's unchanged-fact GEN accuracy falls from 54.6% to 39.1%. Under this recipe every update re-trains six epochs on the cumulative data, so a first-month record has been trained on for 144 epochs by update 24. Progressive degradation under long edit sequences is a known failure mode (Li & Chu, 2024; Gupta et al., 2024). Whether the decline here is caused by that exposure or by weight carry-over itself is not identified by these runs, so we do not attribute it to warm initialization as such; an epoch-sensitivity study is the natural next experiment.

3.3 Endpoint reversals survive the stronger baseline

If the earlier reversals were artifacts of a weak baseline, they should vanish against warm replay. They do not (Table 4). The endpoint opposes the full-history winner in 16 of 18 seed-level GEN comparisons—all nine at rank 32, and all seeds at rank 8 for $K = 6$ and $K = 12$ ($K = 3$ reverses only in seed 2). At rank 32, $K = 6$, warm replay leads by 11.11 points over the stream while the endpoint reports an 11.15-point hierarchy lead. The estimate-side tie band does not change the 16/18 count.

On training-form queries none of the 18 comparisons reverses, but this is a ceiling effect rather than a robustness result. Rank-32 warm replay scores between 97.8% and 99.9% on MEM at every update (rank 8: above 98.9% from the third update on), while the hierarchy's endpoint MEM accuracy is 96.9%, so no endpoint state of the hierarchy overtakes it. The paraphrase form retains headroom and exposes the discrepancy on the same facts. A single held-out template bounds the scope of that observation.

Figure 2 asks the question that a single endpoint cannot: which reporting months would have reversed the winner? For $K = 12$ only the consolidation-aligned months 24 and 1 do; for $K = 6$, months 18, 19, and 24; for $K = 3$, months 18, 19, 21, 22, and 24. At $t = 23$ no period reverses. Endpoint phase therefore explains the $K = 12$ case entirely, while for short periods the late rise in the hierarchy's relative level—driven by warm replay's decline—reverses several non-aligned months as well. Both effects are visible in Figure 1c, and the next section separates them exactly.

4 Phase and drift: why last-cycle averaging is not enough

A natural remedy for a consolidation-aligned endpoint is to average the last full cycle. On the fresh panel this works: final-K agrees with every full-history winner (0/43) and cuts the mean absolute

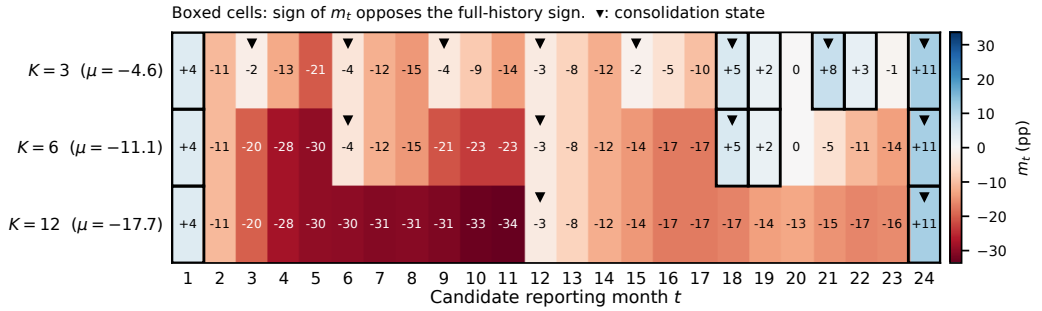

Figure 2: **Which reporting month reverses the winner?** Three-seed mean GEN margins m_t against rank-32 warm replay, for each candidate reporting month, judged against each period's full-history mean μ. Boxed cells have the opposite sign from μ; triangles mark consolidation states. All periods reverse at $t = 24$ and none at $t = 23$; short periods also reverse at several late months. The 24 candidates share the same trained runs.

margin error from 11.05 to 1.20 points. On the warm panel it still reverses four of 18 paraphrase comparisons (all three seeds at rank 32, $K = 3$, and seed 2 at rank 32, $K = 6$), including the rank-32, $K = 3$ aggregate. The decomposition below identifies the error that averaging cannot remove.

Proposition 1 (Exact phase–drift decomposition)**.** *For $T = CK$, write the paired margin in cycle c at phase ϕ as $m_{c,\phi} = u_c + v_{c,\phi}$ with $u_c = K^{-1}\sum_{\phi=1}^{K} m_{c,\phi}$ and $\sum_\phi v_{c,\phi} = 0$. Under the uniform full-history target,*

$$m_T - \mu = \underbrace{u_C - \bar{u}}_{\textit{between-cycle drift}} + \underbrace{v_{C,K}}_{\textit{endpoint phase}}, \tag{2}$$

$$\widehat{\mu}_{\text{final-}K} - \mu = u_C - \bar{u}, \qquad \bar{u} = C^{-1}\sum_{c=1}^{C} u_c. \tag{3}$$

No periodicity or stationarity is assumed.

Proposition 2 (Recency-weighted drift and a sufficient sign condition)**.** *Let $\Delta_c = u_{c+1} - u_c$. Then*

$$\widehat{\mu}_{\text{final-}K} - \mu = \frac{1}{C}\sum_{c=1}^{C-1} c\,\Delta_c, \qquad \left|\widehat{\mu}_{\text{final-}K} - \mu\right| \le \frac{1}{C}\sum_{c=1}^{C-1} c\,|\Delta_c| \le \frac{L(C-1)}{2} \tag{4}$$

whenever $|\Delta_c| \le L$ for all c. Consequently $|\widehat{\mu}_{\text{final-}K}| > L(C-1)/2$ is sufficient for the last-cycle mean to carry the full-history sign, and the last bound is attained by linear drift.

Proofs are in Appendix A. Two readings matter in practice. First, final-K estimates the last cycle's level, not the average level: phase cancellation says nothing about agreement with μ. Second, drift error is a recency-weighted sum of cycle-to-cycle changes, so a late change in relative level counts most, and no amount of within-cycle averaging offsets it.

A measured decomposition. For rank-32 warm replay, $K = 3$, GEN, the full-history margin is -4.60 pp, the last-cycle mean is $+4.29$ pp, and the endpoint is $+11.15$ pp (Figure 3). Equation 2 splits the 15.75 pp endpoint error into 8.89 pp of drift and 6.86 pp of phase. Averaging removes the phase term, but the drift alone exceeds the magnitude of the target, so final-K must reverse the sign. Table 5 applies the drift check to all 12 warm aggregates after the fact: it passes in seven, every passing cell has the right sign, and it withholds exactly the failing cell. For $K = 3$ the bound is loose (19–33 points over eight cycles), which is the price of assuming only a maximum increment; a prospective use would need L justified from a pilot stream rather than the evaluated trajectory.

Which sparse rule to use. The decomposition prescribes the remedy: phase error is removed by sampling several phases, drift error by sampling several cycles. Table 6 compares four rules that score at most K checkpoints. Sampling *only* consolidation states—the protocol implicit in a benchmark

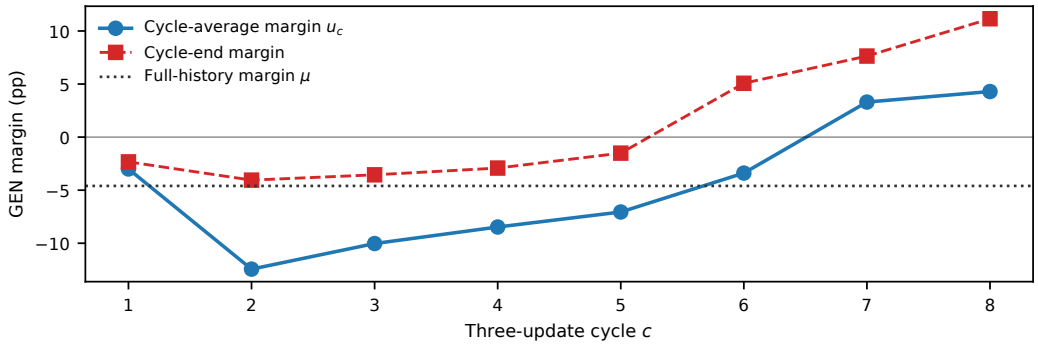

Figure 3: **Removing phase does not remove drift.** Llama-1B, rank-32 warm replay, $K = 3$, GEN, three-seed means. Cycle averages remove within-cycle fluctuation, yet the last three cycle means rise above the negative full-history margin, so the last cycle selects the opposite winner even without its peak.

Table 5: **Retrospective drift check on all warm-panel aggregates.** Percentage points. Drift is the exact recency-weighted sum of Equation 4, equal to final-K minus full; $B = L(C-1)/2$ with L the largest *observed* adjacent-cycle change. "Pass" means $|\text{final-}K| > B$, the sufficient condition of Proposition 2. It passes in 7 of 12 cells, every passing cell has the correct sign, and the one cell whose last-cycle sign is wrong (rank 32, $K = 3$, GEN) is withheld. Because L is computed from the evaluated trajectory, this is a diagnostic, not a prospective certificate.

r	K	Query	Full	Final-K	Drift	B	Pass
8	3	MEM	-4.66	-5.14	-0.48	32.75	no
8	3	GEN	+1.09	+2.54	+1.45	33.37	no
8	6	MEM	-10.40	-9.92	+0.47	1.43	yes
8	6	GEN	-5.42	-4.69	+0.74	3.46	yes
8	12	MEM	-17.66	-15.90	+1.76	1.76	yes
8	12	GEN	-12.05	-10.53	+1.52	1.52	yes
32	3	MEM	-4.94	-4.61	+0.34	18.67	no
32	3	GEN	-4.60	+4.29	+8.89	33.04	no
32	6	MEM	-10.68	-9.44	+1.23	1.55	yes
32	6	GEN	-11.11	-2.77	+8.34	11.82	no
32	12	MEM	-17.94	-15.55	+2.38	2.38	yes
32	12	GEN	-17.73	-12.37	+5.36	5.36	yes

that evaluates after every refresh—inherits most of the endpoint's failures (14/36 on the warm panel, 18/33 on the fresh panel). Final-K removes phase and fails only under drift (4/36 warm, 0/33 fresh). Spreading the same K checkpoints evenly across the stream samples both phases and cycles. It makes no sign error on the warm panel with the lowest error (1.30 points), and its two fresh-panel errors are both cells with $|\mu| < 1\,\text{pp}$, where any sparse rule is fragile. It also beats K states drawn at random, which err in about two cells per panel with roughly twice the error, so the gain comes from the design rather than from using several states. No sparse rule is safe without a bound; among cheap rules, spreading is the one that the decomposition recommends and the data support.

General weights and traffic models. For a target distribution p over states and reporting weights q, the error is $(q-p)^\top m$, and a winner is certified only when a justified bound on that error is smaller than the reported margin (Appendix A). The full-history winner is not fragile to plausible traffic models: weighting every refit state twelve times as heavily as other states (about 70% of traffic at $K = 6$) still leaves rank-32 warm replay ahead on paraphrases by $2.4\,\text{pp}$ at $K = 6$ and $7.3\,\text{pp}$ at $K = 12$, and a recency-weighted target with a six-month half-life leaves it ahead at every K (Appendix A). Only a target that samples refit states *exclusively* favors the hierarchy. Redefining the target as the endpoint is therefore a change of question, not a repair of the estimate.

Table 6: **Sparse reporting rules at equal budget.** Seed-level cells of the Llama-1B warm and fresh panels (ranks 8 and 32; ties $|\mu| < 0.5\,\mathrm{pp}$ excluded). Sign errors count disagreements with the full-history winner; MAE is the mean absolute margin error in percentage points. "Consolidation states only" scores one state per cycle ($C = T/K$); the random row averages 200 seeded draws per cell.

		Warm panel		Fresh panel	
Rule	Checkpoints	Sign errors	MAE	Sign errors	MAE
Endpoint ($t = T$)	1	16/36	13.90	16/33	10.04
Consolidation states only	C	14/36	11.07	18/33	10.61
Final-K (last cycle)	K	4/36	2.90	0/33	1.19
Spread-K (evenly spaced)	K	0/36	1.30	2/33	1.29
Random K states (mean of 200 draws)	K	1.8/36	2.28	2.5/33	1.83

5 Related work

Temporal knowledge and temporal evaluation. TemporalWiki and EvolvingQA evaluate facts as the world changes (Jang et al., 2022; Kim et al., 2024); OAKS streams densely annotated evolving knowledge for online adaptation (Kim et al., 2026), and CRAFT aligns models with real-time edits (Tang et al., 2026). Retrieval-augmented generation (Lewis et al., 2020) maintains knowledge externally and is the complementary route under knowledge drift (Liu et al., 2026); our question is conditional on updating parameters. Temporal knowledge editing asks whether a fact was valid at a given *world time* (Yin et al., 2024). Our axis is different: we score the *model state after each update* and ask which updater a declared mixture of those states favors. Stability-gap work shows that task-boundary evaluation hides intermediate forgetting (De Lange et al., 2023); evaluation of evolving external memory makes a related case for continuous observation (Dong et al., 2026). We build on both by auditing a paired *method decision* for parametric updates once replay initialization and query form are made explicit.

What an editing score measures. Knowledge-editing benchmarks and lifelong editors test retention, locality, and generalization beyond a single edited statement (Cohen et al., 2024; Mitchell et al., 2022; Wang et al., 2024; Hartvigsen et al., 2023; Wang et al., 2025; Liu et al., 2025; Lyu et al., 2026). Pohl et al. (2025) find evaluation-dependent editor rankings and discuss string-matching limitations; O'Neill (2026) separate recital of injected statements from their use under other prompts. Our MEM score is trained-form multiple choice and GEN is one held-out paraphrase of the same fact; in the warm panel the two forms answer the reversal question differently, with the training form at ceiling.

Replay, capacity, and consolidation. Experience replay and gradient-constrained memory are established continual-learning tools (Rolnick et al., 2019; Lopez-Paz & Ranzato, 2017). LoRA rank, fact storage, and supervision form have been studied directly (Back et al., 2026; Pletenev et al., 2025; Jiang et al., 2024). Consolidation, partitioned edit memory, and adapter interference motivate the fast/slow testbed (Schwarz et al., 2018; Wang et al., 2023; Fang et al., 2025; Le & Venkatesh, 2026; Han et al., 2025); these updater families are the natural next controls, and the completed reversal test compares one hierarchy with cumulative replay.

6 Protocol, limitations, and conclusion

A reporting protocol. (i) Declare the target: the final model only, or accuracy across update states under a stated query mix and cohort weighting. (ii) Report the paired margin trajectory together with the declared target and the endpoint, keeping query forms separate and noting any form at ceiling. (iii) When cumulative replay is a baseline, include fresh and warm initialization and a rank sensitivity. (iv) Report unchanged-fact preservation and training exposure separately from updated-fact retention. (v) If only a few checkpoints can be scored, spread them over the deployment interval and state the approximation; a last-cycle average removes phase error but carries no general sign guarantee, and consolidation-aligned checkpoints alone are biased toward the consolidating method.

Limits of the evidence. The stream is retrospective, relation-concentrated, and of unmeasured pretraining overlap; a no-adapter baseline is needed to bound the latter. Warm replay was trained only on Llama-1B; the fresh sweep spans two families and a small 3B check. The main score uses one paraphrase template and fixed distractors. The warm recipe re-trains six epochs on cumulative data every month, so the late decline that drives drift may be recipe-dependent. Epoch and learning-rate sensitivity is the first experiment we would run next, followed by one external updater family on the same stream and a hierarchy variant that warm-starts its slow adapter. Hyperparameters are fixed rather than tuned under a common budget, costs are compared only by a training-exposure proxy, and historical software revisions are incompletely recorded.

Conclusion. Choose the reporting window to match the intended use, and test whether a winner survives reasonable baseline and query-form choices. In this case study a warm replay baseline changes the full-history winner, consolidation-aligned endpoints select the opposite paraphrase winner, and last-cycle averaging fails whenever late drift outweighs the margin. The phase–drift decomposition explains all three, and spreading a fixed checkpoint budget across the stream is the cheap rule it recommends.

AI USE STATEMENT

We used generative AI assistants to propose and refine hypotheses, critique experimental designs, help implement and inspect analysis code, organize exported numerical records, interpret completed results, develop the phase–drift formulation and draft its proofs, identify and summarize related work, check bibliographic details, translate research notes, produce figures and tables, and draft and edit manuscript sections. The training and evaluation records were selected from Wikidata; no synthetic dataset was generated. The authors reviewed all AI-assisted work against the saved observations and primary sources, regenerated every numerical table and figure from the included exports, and checked the mathematical identities algebraically and with deterministic property tests. The authors are responsible for the final text, code, figures, references, proofs, and empirical claims.

REPRODUCIBILITY STATEMENT

The anonymous supplementary package regenerates all tables and vector figures from the original and updated aggregate exports and 21 saved Llama trajectories, with provenance hashes and deterministic algebra checks. It reproduces the reported analysis; raw training data, model checkpoints, and complete environment locks are outside the package. Appendix B records the training settings and the unresolved historical provenance.

ETHICS STATEMENT

The experiments concern public factual records and evaluation methodology and do not validate an autonomous or safety-critical factual service. Historical extraction can preserve outdated associations, and retention of a selected statement does not establish that it is the correct current answer. Reporting uncertainty, query-form dependence, and preservation losses is necessary to avoid overstating the reliability of deployed updates.

REFERENCES

Seungju Back, Dongwoo Lee, Naun Kang, Taehee Lee, S. K. Hong, Youngjune Gwon, and Sungjin Ahn. Understanding LoRA as knowledge memory: An empirical analysis. *arXiv preprint arXiv:2603.01097*, 2026. URL https://arxiv.org/abs/2603.01097.

Heejin Choi. Beyond endpoint scores: Time- and capacity-conditioned evaluation of continual knowledge updating. *arXiv preprint arXiv:2609.03900*, 2026. URL https://arxiv.org/abs/2609.03900.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/.

Matthias De Lange, Gido M. van de Ven, and Tinne Tuytelaars. Continual evaluation for lifelong learning: Identifying the stability gap. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zy350cRstc6.

Songwei Dong, Zihan Chen, Chengshuai Shi, Peng Wang, Jundong Li, and Cong Shen. Is one score enough? rethinking the evaluation of sequentially evolving LLM memory. *arXiv preprint arXiv:2605.15384*, 2026. URL https://arxiv.org/abs/2605.15384.

Quntian Fang, Zhen Huang, Zhiliang Tian, Minghao Hu, Dongsheng Li, Yiping Yao, Xinyue Fang, Menglong Lu, and Guotong Geng. Hippocampal-like sequential editing for continual knowledge updates in large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0305. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0d230e3015b806e7c9cc9afd5ac2c4a9-Abstract-Conference.html.

Yujie Feng, Jian Li, Xiaoyu Dong, Pengfei Xu, Xiaohui Zhou, Yujia Zhang, Zexin Lu, Yasha Wang, Alan Zhao, Xu Chu, and Xiao-Ming Wu. AIMMerging: Adaptive iterative model merging using training trajectories for language model continual learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13420–13437, 2025. doi: 10.18653/v1/2025.emnlp-main.678. URL https://aclanthology.org/2025.emnlp-main.678/.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15202–15232, 2024. doi: 10.18653/v1/2024.findings-acl.902. URL https://aclanthology.org/2024.findings-acl.902/.

Xiao Han, Zimo Zhao, Wanyu Wang, Maolin Wang, Zitao Liu, Yi Chang, and Xiangyu Zhao. Data efficient adaptation in large language models via continuous low-rank fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5506. URL https://papers.nips.cc/paper_files/paper/2025/hash/f14c87dcab0d2869a28d6400b2c5e084-Abstract-Conference.html.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/95b6e2ff961580e03c0a662a63a71812-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://aclanthology.org/2022.emnlp-main.418/.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. MoRA: High-rank updating for parameter-efficient fine-tuning. *arXiv preprint arXiv:2405.12130*, 2024. URL https://arxiv.org/abs/2405.12130.

Jiyeon Kim, Hyunji Lee, Dylan Zhou, Sue Hyun Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Sungmin Cha, and Minjoon Seo. Can large language models keep up? benchmarking online adaptation to continual knowledge streams. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42240–42272, 2026. doi: 10.18653/v1/2026.acl-long.1956. URL https://aclanthology.org/2026.acl-long.1956/.

Yujin Kim, Jaehong Yoon, Seonghyeon Ye, Sangmin Bae, Namgyu Ho, Sung Ju Hwang, and Se-Young Yun. Carpe diem: On the evaluation of world knowledge in lifelong language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5401–5415, 2024. doi: 10.18653/v1/2024.naacl-long.302. URL https://aclanthology.org/2024.naacl-long.302/.

Hung Le and Svetha Venkatesh. Continual fine-tuning of large language models via program memory. *arXiv preprint arXiv:2605.13162*, 2026. URL https://arxiv.org/abs/2605.13162.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.11401.

Qi Li and Xiaowen Chu. Can we continually edit language models? on the knowledge attenuation in sequential model editing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5438–5455, 2024. doi: 10.18653/v1/2024.findings-acl.323. URL https://aclanthology.org/2024.findings-acl.323/.

Hanbing Liu, Lang Cao, and Yang Li. RAG or learning? understanding the limits of LLM adaptation under continuous knowledge drift in the real world. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. doi: 10.18653/v1/2026.findings-acl.546. URL https://aclanthology.org/2026.findings-acl.546/.

Jinzhe Liu, Junshu Sun, Shufan Shen, Chenxue Yang, and Shuhui Wang. Edit less, achieve more: Dynamic sparse neuron masking for lifelong knowledge editing in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4770. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d1bfaf1bbefb344df643997689b46613-Abstract-Conference.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Sicheng Lyu, Yu Gu, Xinyu Wang, Jerry Huang, Sitao Luan, Yufei Cui, Xiao-Wen Chang, and Peng Lu. EvoEdit: Evolving null-space alignment for robust and efficient knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1520–1540, 2026. doi: 10.18653/v1/2026.findings-acl.75. URL https://aclanthology.org/2026.findings-acl.75/.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831, 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Charles O'Neill. Can a language model learn facts continually in its weights? *arXiv preprint arXiv:2607.11020*, 2026. URL https://arxiv.org/abs/2607.11020.

Sergey Pletenev, Maria Marina, Daniil Moskovskiy, Vasily Konovalov, Pavel Braslavski, Alexander Panchenko, and Mikhail Salnikov. How much knowledge can you pack into a LoRA adapter without harming LLM? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4309–4322, 2025. doi: 10.18653/v1/2025.findings-naacl.243. URL https://aclanthology.org/2025.findings-naacl.243/.

Sebastian Pohl, Max Ploner, and Alan Akbik. Towards a principled evaluation of knowledge editors. In *Proceedings of the First Workshop on Large Language Model Memorization*, pp. 47–60, 2025. doi: 10.18653/v1/2025.l2m2-1.4. URL https://aclanthology.org/2025.l2m2-1.4/.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single LoRA continual learning via continual merging. *arXiv preprint arXiv:2512.23017*, 2025. URL https://arxiv.org/abs/2512.23017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4528–4537, 2018. URL https://proceedings.mlr.press/v80/schwarz18a.html.

Chenming Tang, Yutong Yang, Kexue Wang, and Yunfang Wu. Aligning language models with real-time knowledge editing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 363–378, 2026. doi: 10.18653/v1/2026.acl-long.14. URL https://aclanthology.org/2026.acl-long.14/.

Ke Wang, Yiming Qin, Nikolaos Dimitriadis, Alessandro Favero, and Pascal Frossard. MEMOIR: Lifelong model editing with minimal overwrite and informed retention for LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/809a560bf60984be05180c071c352eaa-Abstract-Conference.html.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/60960ad78868fce5c165295fbd895060-Abstract-Conference.html.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715/.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.01708.

Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. History matters: Temporal knowledge editing in large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19413–19421, 2024. doi: 10.1609/aaai.v38i17.29912. URL https://ojs.aaai.org/index.php/AAAI/article/view/29912.

A PROOFS AND INTERPRETATION OF REPORTING ERROR

The mathematical statements concern a deterministic, observed margin sequence. They do not model gradient dynamics, infer causation from an oscillation, or establish a new continual-learning convergence rate. Their role is to make the reporting recommendation falsifiable: phase balancing and long-horizon representativeness are different conditions.

A.1 GENERAL REPORTING WEIGHTS

Let $p, q \in \mathbb{R}^T$ be probability vectors over update states. The target margin is $\mu_p = \sum_t p_t m_t$ and the reported estimate is $\widehat{\mu}_q = \sum_t q_t m_t$. A final-only target has $p_T = 1$, while our benchmark uses $p_t = 1/T$. A last-cycle report assigns $q_t = 1/K$ on its final K positions and zero elsewhere.

Lemma 1 (Reporting error and sign preservation)**.** *For $a = \min_t m_t$, $b = \max_t m_t$,*

$$e_{q,p} = \widehat{\mu}_q - \mu_p = (q - p)^\top m, \qquad |e_{q,p}| \le \frac{b-a}{2}\|q - p\|_1. \tag{5}$$

For nonzero μ_p, a strict sign reversal occurs exactly when $\mu_p(\mu_p + e_{q,p}) < 0$. If an independently justified bound $|e_{q,p}| \le \varepsilon$ is available, either $|\mu_p| > \varepsilon$ or $|\widehat{\mu}_q| > \varepsilon$ is sufficient to rule out a reversal.

Proof. The identity follows by subtraction. Since $\sum_t (q_t - p_t) = 0$, subtracting $(a + b)/2$ from each m_t does not change the inner product. Applying the triangle inequality and $|m_t - (a + b)/2| \le (b - a)/2$ proves the bound. The reversal condition is the product of the two margins. If two numbers have opposite signs, their distance is at least the absolute value of either; a strictly smaller error bound excludes this possibility. □

For an endpoint and a uniform target, the distribution mismatch is large: $\|q - p\|_1 = 2(1 - 1/T)$. Without information about the temporal sequence, the resulting bound may be too loose to certify any winner. This is why measuring or assuming additional temporal structure is necessary. It is also why redefining the target to be the endpoint is not a repair of an across-time estimate; it is a change of question.

A.2 PROOF OF PROPOSITION 1

By definition, $m_{c,\phi} = u_c + v_{c,\phi}$ and $\sum_\phi v_{c,\phi} = 0$. Averaging all CK entries gives

$$\mu = \frac{1}{C}\sum_{c=1}^{C} u_c = \bar{u}.$$

The endpoint is $u_C + v_{C,K}$ and the final-cycle mean is u_C. Subtracting $\bar{u}$ proves both identities. There is no assumption that the within-cycle pattern repeats across cycles or that the cycle means are constant.

A.3 PROOF OF PROPOSITION 2

Let $\Delta_c = u_{c+1} - u_c$. Each Δ_k appears in $u_C - u_c$ for exactly k of the C cycle-indexed summands. Hence the exact identity and its two bounds are

$$u_C - \bar{u} = \frac{1}{C}\sum_{c=1}^{C-1} c\Delta_c, \qquad |u_C - \bar{u}| \le \frac{1}{C}\sum_{c=1}^{C-1} c|\Delta_c| \le \frac{L(C-1)}{2}.$$

If the last-cycle estimate has greater magnitude than a justified upper bound on its distance to μ, its sign must match μ. Equality in the last bound holds for linear cycle means $u_c = u_1 + (c - 1)L$, including bounded accuracy-margin sequences whenever their total drift fits within $[-1, 1]$. The empirical L in Table 5 is computed using all observed cycles; it verifies a retrospective condition rather than certifying unobserved future data.

A simple counterexample. Take four cycles with means $(-0.20, -0.15, -0.10, +0.05)$ and zero within-cycle residual. Their full-history margin is -0.10, whereas the final-cycle margin is $+0.05$. Averaging a whole cycle cannot help here: there is no phase component to remove. This is a mathematical example, not an additional experimental result. The measured warm-replay example in Section 4 has both components and demonstrates the same limitation in actual saved runs.

Unequal query traffic. Uniform phase balance is useful only for a target that weights phases uniformly. If user queries concentrate after refresh, a uniform cycle average may itself be misaligned with the application. The general-weight identity above remains valid; the simplified cycle identities apply to the uniform benchmark target. This paper does not estimate a real application's query-traffic distribution, but Table 7 shows how far the paraphrase winner moves under stylized traffic models: refit states ($t \in \{K, 2K, \dots\}$) weighted w times as heavily as all other states, and a recency model with geometric half-life h months. Warm replay stays ahead at $K = 6$ and $K = 12$ under every model, reaches a tie at $K = 3$ only at $w = 12$, and loses only when refit states are sampled exclusively.

Table 7: Rank-32 warm replay, GEN, three-seed mean margins (hierarchy minus replay, pp) under alternative target weights.

	Uniform	Refit states weighted $w\times$			Recency, half-life h		Refit only	Endpoint
K		w=3	w=6	w=12	h=12	h=6		
3	-4.60	-2.28	-0.98	-0.05	-2.68	-0.73	+1.19	+11.15
6	-11.11	-7.75	-5.01	-2.42	-9.21	-7.30	+2.31	+11.15
12	-17.73	-14.61	-11.30	-7.28	-15.84	-13.76	+4.12	+11.15

Sparse reporting rules (Table 6). Each rule is a reporting weight vector q applied to the same seed-level margin trajectory $m \in \mathbb{R}^{24}$. *Endpoint* puts all weight on $t = 24$. *Consolidation states only* averages m_t over $t \in \{K, 2K, \dots, 24\}$, one state per cycle. *Final-K* averages $t \in \{25 - K, \dots, 24\}$. *Spread-$K$* averages K evenly spaced states $t_i = \mathrm{round}\left(24\,(i + \frac{1}{2})/K\right)$, $i = 0, \dots, K - 1$, that is $\{4, 12, 20\}$ for $K = 3$, $\{2, 6, 10, 14, 18, 22\}$ for $K = 6$, and the odd months for $K = 12$; the grid mixes phases and covers all cycles, and it was fixed before the counts were computed. Sign errors are counted against the full-history winner of the same cell, excluding reference ties $|\mu| < 0.5\,\mathrm{pp}$ (36 warm and 33 fresh non-tie cells). The two fresh-panel spread-K errors are rank 8, $K = 12$, GEN, seed 0 ($\mu = +0.72\,\mathrm{pp}$, estimate $-0.28\,\mathrm{pp}$) and rank 32, $K = 3$, MEM, seed 2 ($\mu = -0.62\,\mathrm{pp}$, estimate $+0.04\,\mathrm{pp}$). The four warm-panel final-K errors are the three rank-32, $K = 3$, GEN seeds and rank 32, $K = 6$, GEN, seed 2. Shifting every grid point by -2 to $+2$ months leaves the spread rule with at most one warm-panel and two fresh-panel sign errors. All rules share the same trained runs; the comparison is descriptive.

B Experimental protocol and provenance

B.1 Data semantics and sampling

The stream builder assigns selected temporal Wikidata records to monthly event-time buckets. This differs from TemporalWiki's differences between consecutive snapshots (Jang et al., 2022). The current data audit supports the record counts, selected item overlap, and stored cohort definitions. It does not establish that a fact was first knowable in its assigned month, that it was absent from pretraining, or that every repeated subject–relation pair has an audited obsolete-value policy. Those stronger claims are not made.

Training records are capped at 200 per month in the original recipe. The multiple-choice evaluator constructs eligible same-relation distractors before selecting at most 100 facts per cohort. The selected facts are reused across methods and update states. The unchanged-control evaluation uses a cap of 300 records. Its exact benchmark accuracy is reported, rather than extrapolated to general abilities, safety, or unrelated downstream tasks.

The two surfaces are paired prompts for the same fact. MEM refers to the learned canonical form; GEN is a held-out paraphrase template. This is a limited form of query transfer. The single held-out

phrasing and fixed candidate construction do not certify paraphrase robustness or open-ended question answering.

B.2 TRAINING IMPLEMENTATION

Table 8: Recorded recipe and checked implementation. Historical environment completeness is discussed below.

Item	Setting or behavior
Primary backbones	Llama-3.2-1B; Qwen2.5-1.5B; limited Llama-3.2-3B scale check
Original training cap	Up to 200 records per update
Epochs / learning rate	6 / 3×10^{-4}
Batch / maximum text length	16 / 64 tokens in the checked training implementation
Optimizer / schedule	AdamW; cosine schedule with 0.03 warmup ratio; recreated for each training call
Gradient clipping	Norm 1 in the checked implementation
LoRA targets	Attention q, k, v, o projections and MLP gate/up/down projections
LoRA scale / dropout	$\alpha/r = 2$ in main sweep; dropout 0
Hierarchy	Slow rank 64, fast rank 8; $K \in \{3, 6, 12\}$
Warm replay	Prior adapter weights loaded; no saved Adam moments carried
Backbone update	No trained backbone weights on disk; slow adapters can be merged transiently for composing a fast level
Hyperparameter selection	Fixed recipes; no shared validation-based search establishing globally optimal baselines

The fast hierarchy state is a single adapter whose weights continue within a cycle. It is not a collection of separate historical fast adapters requiring simultaneous loading. When the slow adapter is replaced, the fast adapter restarts. Consequently the prior slow plus the previous month's fast adapter reconstructs the recorded previous-month state in the generation probe. It is still not an instrumented same-month, immediately-pre-consolidation counterfactual.

Limits of recipe equivalence. Fresh and warm replay intentionally differ in inherited adapter weights. They otherwise follow the matched data and per-update schedule in the checked code paths. This isolates an initialization/history choice within the implementation; it does not imply equal information accumulated in those weights. The hierarchy has a different replay schedule and total training exposure. We therefore separate baseline sensitivity from a fully compute-matched Pareto claim.

Current code and saved metrics support the reported settings, but not every historical launcher argument or immutable model/tokenizer revision is recoverable from each legacy result JSON. The source package marks these as verification requirements rather than retrospectively inventing pins. Device agreement across differently seeded historical runs cannot, on its own, bound a hardware effect.

Training-exposure proxy. If every update contained the maximum 200 records, six epochs of cumulative replay after all 24 updates would expose $6 \cdot 200 \sum_{t=1}^{24} t = 360{,}000$ record-epochs. For the $K = 6$ hierarchy, slow refits at $t = 1, 6, 12, 18, 24$ and a six-epoch monthly fast fit give $6 \cdot 200(1 + 6 + 12 + 18 + 24 + 24) = 102{,}000$ record-epochs, about 28% of this replay proxy. The actual stream contains 4,739 rather than 4,800 records, and this accounting excludes sequence length, adapter size, evaluation, and hardware effects. It describes different training exposures, not measured wall-clock or compute efficiency.

B.3 COVERAGE AND INDEPENDENT UNITS

The frozen source-corrected corpus has 44 aggregate cells and uses 47 distinct stored run files, not 44×3 independent training runs: hierarchy runs are shared across replay ranks and both surfaces come from the same checkpoints. The current 44-cell panel replaces the two Qwen rank-64 surface

aggregates with a completed three-seed extension. The complete warm extension adds six trained trajectories and reuses corresponding hierarchy/fresh anchors. The 36 warm comparison cells likewise reuse these runs across periods and surfaces.

Accordingly, $16/18$ paraphrase reversals is a descriptive coverage statement. It is not a binomial sample with 18 independent trials. The gain table uses sample standard deviations over three paired training seeds; no significance claim is inferred from the count or from a standard deviation. Qwen rank-64 now has three seeds; Llama-3B rank-72 remains one-seed. The preserved original export allows its earlier one-seed count to be reproduced.

B.4 Evidence snapshot and analysis corrections

The included numerical snapshot was exported after verifying the corrected original run-file hashes and completed warm trajectories. Every primary trajectory has 24 ordered monthly records. Tables are regenerated from the saved accuracies rather than copied from evolving prose. The original audit corrected an analysis-loader collision between streams and adapter types; this manuscript uses the corrected 44-cell corpus, not the earlier contaminated 34-cell summary.

The package contains a compact 21-run Llama export underlying all new warm comparisons, the frozen and updated 44-cell aggregate tables with the three-seed Qwen source JSON, and the generation-probe values with cohort hashes. The complete source-file manifest remains associated with the research archive. Analysis reproducibility from these exports is distinct from reproducing training from raw data.

Winner conventions. The primary tie rule is on the *reference* margin: $|\mu| < 0.005$. For a non-tie reference, an estimate's raw sign determines disagreement. In the warm rank-32, $K = 3$, seed-0 GEN cell, $\mu = -0.069854$ and final-$K = +0.002500$. The original rule counts this as an error even though the estimate is small. Applying an additional estimate-side band changes warm GEN final-K errors from four to three out of 18; we disclose that sensitivity rather than replacing the primary definition. The rule does not classify an estimate as statistically significant.

Reference-window sensitivity. In separately recorded analyses, excluding both calendar Januaries from the reference gives one final-K disagreement out of 42 non-tie aggregate cells. Excluding the newly learned cohort at each update gives two out of 43. These are changes in the evaluation target using saved results, not date-corrected retraining or new independent datasets. They limit the portability of the original zero-error final-K result.

Endpoint estimate-band sensitivity. In the updated fresh-replay panel, four of 18 reference-only endpoint reversals have $|m_T| < 0.005$: three Llama-1B rank-64 GEN cells share the same near-zero endpoint across $K = 3, 6, 12$, and one Qwen-1.5B rank-64 MEM cell has $m_T = +0.0004$. Treating those estimates as unresolved gives 14/43 reversals under the additional estimate-side band. The warm replay endpoint count remains 16/18 on GEN under either convention.

Secondary analyses not promoted to primary claims. The available endpoint-prefix sweep is descriptive over overlapping prefixes. Its raw phase index should be stratified by K; pooling phase 0 across several periods and comparing it to phase 3 does not isolate phase alone. Similarly, a mean of cohort-time accuracies is not a fact-micro average without denominator checks, and a quantile of the *paired margin* is not the difference of the methods' accuracy quantiles. We do not use these summaries to claim absolute deployment reliability. An always-positive bias baseline must also be considered before treating high bias-sign prediction accuracy as a useful forecasting result.

C All aggregate and seed-level results

C.1 Complete warm panel, aggregate over three seeds

All entries are hierarchy minus replay, in percentage points. "Fresh" and "Warm" are full-history margins; "End" and "Final-K" use warm replay. Warm replay wins 11 of 12 full-history aggregates; four non-tie aggregate conclusions switch from fresh to warm. The $r = 32$, $K = 3$, GEN fresh margin is a reference tie, not a fifth non-tie reversal.

r	K	Query	Fresh	Warm	End	Final-K
8	3	MEM	+9.69	-4.66	-2.12	-5.14
8	3	GEN	+13.84	+1.09	+9.44	+2.54
8	6	MEM	+3.96	-10.40	-2.12	-9.92
8	6	GEN	+7.33	-5.42	+9.44	-4.69
8	12	MEM	-3.30	-17.66	-2.12	-15.90
8	12	GEN	+0.71	-12.05	+9.44	-10.53
32	3	MEM	-1.01	-4.94	-1.57	-4.61
32	3	GEN	+0.23	-4.60	+11.15	+4.29
32	6	MEM	-6.75	-10.68	-1.57	-9.44
32	6	GEN	-6.28	-11.11	+11.15	-2.77
32	12	MEM	-14.01	-17.94	-1.57	-15.55
32	12	GEN	-12.90	-17.73	+11.15	-12.37

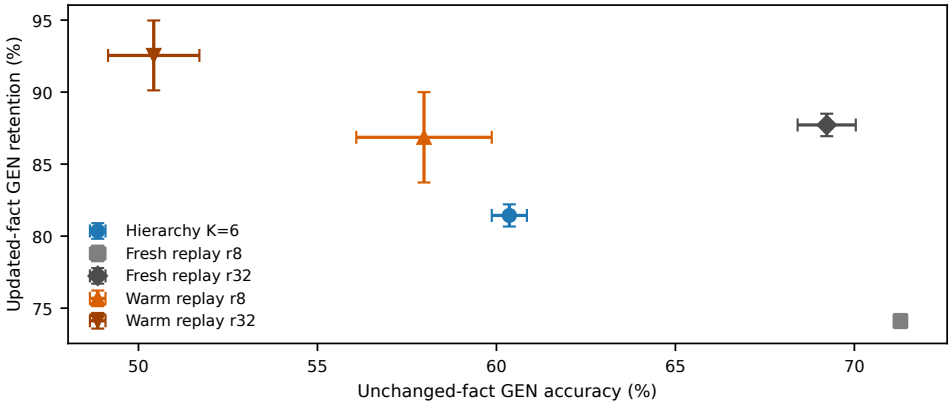

Figure 4: **Retained facts and unchanged facts are different objectives.** Llama-1B full-history GEN means, with sample standard deviations over three seeds. Warm replay's improved updated-fact retention coexists with reduced unchanged-fact accuracy. This is not a compute-matched frontier.

C.2 Warm seed-level margins

Table 9: Warm-panel individual seeds. Units are percentage points.

r	K	Query	Seed	Fresh avg.	Warm avg.	End	Final-K
8	3	MEM	0	+9.11	-5.35	-1.67	-5.75
8	3	MEM	1	+9.73	-4.34	-2.08	-4.77
8	3	MEM	2	+10.23	-4.30	-2.62	-4.91
8	3	GEN	0	+13.90	+3.75	+10.08	+5.93
8	3	GEN	1	+13.54	+1.82	+12.58	+3.98
8	3	GEN	2	+14.09	-2.31	+5.67	-2.29
8	6	MEM	0	+2.94	-11.52	-1.67	-10.39
8	6	MEM	1	+4.07	-10.00	-2.08	-9.86
8	6	MEM	2	+4.85	-9.67	-2.62	-9.53
8	6	GEN	0	+6.76	-3.40	+10.08	-2.21
8	6	GEN	1	+6.95	-4.76	+12.58	-2.58
8	6	GEN	2	+8.29	-8.11	+5.67	-9.28
8	12	MEM	0	-4.08	-18.55	-1.67	-16.38
8	12	MEM	1	-3.26	-17.33	-2.08	-15.71
8	12	MEM	2	-2.57	-17.10	-2.62	-15.60
8	12	GEN	0	+0.72	-9.44	+10.08	-6.14
8	12	GEN	1	+0.31	-11.41	+12.58	-8.57
8	12	GEN	2	+1.10	-15.29	+5.67	-16.87

r	K	Query	Seed	Fresh avg.	Warm avg.	End	Final-K
32	3	MEM	0	-1.41	-5.48	-1.21	-4.99
32	3	MEM	1	-1.01	-4.81	-1.54	-4.53
32	3	MEM	2	-0.62	-4.54	-1.96	-4.30
32	3	GEN	0	-0.07	-6.99	+7.96	+0.25
32	3	GEN	1	-0.25	-4.78	+10.25	+4.16
32	3	GEN	2	+1.00	-2.03	+15.25	+8.47
32	6	MEM	0	-7.58	-11.65	-1.21	-9.82
32	6	MEM	1	-6.67	-10.47	-1.54	-9.62
32	6	MEM	2	-6.00	-9.92	-1.96	-8.89
32	6	GEN	0	-7.22	-14.13	+7.96	-9.61
32	6	GEN	1	-6.83	-11.36	+10.25	-3.70
32	6	GEN	2	-4.80	-7.83	+15.25	+4.99
32	12	MEM	0	-14.61	-18.67	-1.21	-15.90
32	12	MEM	1	-14.00	-17.80	-1.54	-15.53
32	12	MEM	2	-13.42	-17.34	-1.96	-15.23
32	12	GEN	0	-13.25	-20.17	+7.96	-15.91
32	12	GEN	1	-13.48	-18.01	+10.25	-12.77
32	12	GEN	2	-11.98	-15.01	+15.25	-8.42

C.3 Updated source-corrected comparison corpus

All margins in Table 10 compare the hierarchy with *fresh* replay. These are kept separate from the warm extension. The completed Qwen rank-64 three-seed extension replaces the two original one-seed aggregates; the updated panel has 44 cells and 43 non-ties, with 18/43 endpoint errors and 0/43 final-K errors. The original export remains in `data/cells.csv` and has 17/43 and 0/43 under the same rule. The extra displayed precision comes from averaging stored monthly accuracies, not from greater raw measurement precision.

Table 10: Updated aggregate fresh-replay results, in percentage points. n is the seed count.

Backbone	K	r	Query	n	Average	Endpoint	Final-K
Llama-1B	3	8	GEN	3	+13.84	+18.60	+13.59
Llama-1B	3	8	MEM	3	+9.69	+13.13	+10.10
Llama-1B	3	16	GEN	3	+7.92	+12.11	+6.84
Llama-1B	3	16	MEM	3	+3.20	+6.67	+3.61
Llama-1B	3	32	GEN	3	+0.23	+2.74	-0.79
Llama-1B	3	32	MEM	3	-1.01	+1.92	-0.96
Llama-1B	3	64	GEN	3	-5.90	+0.14	-5.88
Llama-1B	3	64	MEM	3	-3.29	-0.47	-3.39
Llama-1B	3	72	GEN	3	-6.71	-2.63	-6.76
Llama-1B	3	72	MEM	3	-3.59	-0.40	-3.45
Llama-1B	6	8	GEN	3	+7.33	+18.60	+7.93
Llama-1B	6	8	MEM	3	+3.96	+13.13	+5.25
Llama-1B	6	16	GEN	3	+1.41	+12.11	+1.61
Llama-1B	6	16	MEM	3	-2.53	+6.67	-1.24
Llama-1B	6	32	GEN	3	-6.28	+2.74	-6.39
Llama-1B	6	32	MEM	3	-6.75	+1.92	-5.87
Llama-1B	6	64	GEN	3	-12.41	+0.14	-12.37
Llama-1B	6	64	MEM	3	-9.02	-0.47	-8.23
Llama-1B	6	72	GEN	3	-13.22	-2.63	-12.97
Llama-1B	6	72	MEM	3	-9.32	-0.40	-8.38
Llama-1B	12	8	GEN	3	+0.71	+18.60	+2.84
Llama-1B	12	8	MEM	3	-3.30	+13.13	-0.83
Llama-1B	12	16	GEN	3	-5.21	+12.11	-3.34
Llama-1B	12	16	MEM	3	-9.79	+6.67	-7.32

Table 11: **Fixed-fact free-generation accuracy (%).** A single saved training run, not four independent training replications. "Slow" is the refreshed slow adapter; "Stack" adds the new fast adapter. These numbers must not be compared directly with the multiple-choice accuracies in Table 3.

		MEM			GEN		
Boundary	n	Pre	Slow	Stack	Pre	Slow	Stack
2024-06	100	22.0	86.0	74.0	4.0	54.0	52.0
2024-12	220	23.2	89.1	86.8	9.5	56.4	48.6
2025-06	340	24.4	89.1	78.2	10.0	51.8	39.1
2025-12	460	23.7	90.2	85.0	6.5	35.7	29.8

Backbone	K	r	Query	n	Average	Endpoint	Final-K
Llama-1B	12	32	GEN	3	-12.90	+2.74	-11.89
Llama-1B	12	32	MEM	3	-14.01	+1.92	-11.87
Llama-1B	12	64	GEN	3	-19.03	+0.14	-18.18
Llama-1B	12	64	MEM	3	-16.28	-0.47	-14.17
Llama-1B	12	72	GEN	3	-19.85	-2.63	-18.74
Llama-1B	12	72	MEM	3	-16.58	-0.40	-14.38
Llama-3B	6	8	GEN	3	+7.58	+16.61	+9.05
Llama-3B	6	8	MEM	3	+0.63	+9.28	+2.30
Llama-3B	6	72	GEN	1	-10.31	-2.25	-9.68
Llama-3B	6	72	MEM	1	-8.61	-0.25	-7.18
Qwen-1.5B	6	8	GEN	3	+9.07	+27.28	+13.83
Qwen-1.5B	6	8	MEM	3	+5.02	+17.36	+7.10
Qwen-1.5B	6	16	GEN	3	+4.05	+20.58	+8.25
Qwen-1.5B	6	16	MEM	3	-1.92	+10.22	-0.57
Qwen-1.5B	6	32	GEN	3	-4.04	+10.79	-1.93
Qwen-1.5B	6	32	MEM	3	-7.98	+3.48	-7.09
Qwen-1.5B	6	64	GEN	3	-12.20	+3.19	-12.44
Qwen-1.5B	6	64	MEM	3	-11.31	+0.04	-10.23
Qwen-1.5B	6	72	GEN	3	-10.87	+4.89	-9.65
Qwen-1.5B	6	72	MEM	3	-11.62	-0.25	-10.46

D AUXILIARY FIXED-FACT GENERATION PROBE

Question and setup. Do the hierarchy's saved states differ on a *fixed fact set*, or could a change in the evaluation population explain the apparent recovery? We reuse Llama-1B, $K = 6$, seed-0 checkpoints at four boundaries. At each boundary, 20 eligible facts from each previous month form one fixed cohort. We compare the previous month's slow-plus-fast state ("pre"), the refreshed slow adapter alone, and its composition with the new fast adapter.

This probe uses free generation, not the likelihood-based multiple-choice metric above. Greedy continuations are capped at 16 new tokens and scored against normalized labels/aliases with word-boundary prefix matching. It measures bounded answer generation under that grader; it is not an independent human assessment. The fact IDs and prompts are held identical across the three states within each boundary.

Result and interpretation. On the identical within-boundary fact sets, the stacked post-update state exceeds the previous state by 52.0–63.6 points on MEM. Adding the new fast adapter to the refreshed slow adapter lowers prior-fact accuracy in all eight boundary–query-form cells of this cohort design. This is direct evidence about *state composition*. It is not an isolated causal effect of consolidation: the previous-month and post-boundary models also differ in their training data and optimization history, and there is no same-data no-consolidation counterfactual. Nor does the probe measure any benefit on newly arriving facts.

Controls constrain the mechanism claim. Reusing the same earliest 100 facts across all boundaries yields slow-only GEN accuracies 54, 56, 56, and 27%; an age-matched design yields 64, 66, 56, and 48%. The fixed-fact result is flat until the final boundary, not evidence of monotone deterioration with stream length. Furthermore, on the fixed cohort at the final boundary, the fast adapter *helps* GEN by three points. Thus fast-adapter interference is frequent here, not universal. A planned equal-quota team-membership/position-held control was unavailable because this primary stream lacks eligible position-held (P39) records.

These probes and the ranking panel share the stream and hierarchy, so they are complementary analyses, not independent replications. Together they motivate testing query-form robustness and adapter composition; they do not establish a universal mechanism for the paraphrase reversals.

D.1 Additional controls and boundaries of inference

The main probe samples 20 eligible prior facts per monthly cohort. At each boundary, pre, slow-only, and stacked use exactly the same prompts and fact IDs. The pre state is reconstructed from the previous slow checkpoint and the previous month's complete fast checkpoint. The output metric is *greedy free generation* with normalized alias-prefix matching, so it must not be pooled with the main multiple-choice accuracy.

Table 12: Slow-only generation on the controls. The nested cohort is the same 100 facts at every boundary. Age matching uses 25 facts in each of the 1–3 and 4–6 month bins.

Design/query	2024-06	2024-12	2025-06	2025-12
Fixed-100 MEM	86	93	94	90
Fixed-100 GEN	54	56	56	27
Age-matched MEM	86	88	92	92
Age-matched GEN	64	66	56	48

The fixed cohort holds content and relation membership constant, but the facts age as training continues. The age-matched cohort controls the age-bin counts, but its content changes. Agreement in the direction of the final decline under these two checks is evidence against a simple growing-pool explanation; it does not identify stream length as the cause. In particular, the fixed cohort is stable at the first three boundaries and drops at the fourth. A checkpoint-specific effect or training-state difference remains possible.

The fast-composition effect is also conditional. On the fixed 100 facts, stacked-minus-slow GEN changes are $-2, -8, -17, +3$ percentage points. The final positive value rules out universal harm from adding the fast adapter. The original, growing-cohort design gives eight negative changes across four boundaries and two query forms; these are correlated evaluations of one training run.

A P54/P39 equal-quota check was unavailable because there were no eligible P39 records in the primary stream. This failed quota does not prove that all possible relation-balancing schemes are impossible. It establishes only that the specified control was not run. Relation-diversity generalization remains limited.

What would establish a stronger mechanism? A same-data intervention that withholds consolidation while keeping the update data, evaluation facts, optimizer budget, and remaining state changes controlled would address causal attribution more directly. Evaluating new facts under the pre-update model would separately measure acquisition. Neither result is supplied by the current checkpoint ablation, and neither is claimed.

E Evidence boundaries

The main results are completed observations and the formal results are proved for the stated definitions. The following boundaries specify what the evidence supports.

Historical reproducibility. Immutable model/tokenizer revisions and complete launcher arguments for every legacy run were not recorded. Current code defaults identify the model families but do

not establish every historical environment. The saved hashes identify analyzed outputs rather than a complete reproducible software image.

Dataset semantics. Repeated-change/obsolete-value handling, exact extraction revisions, and pretraining overlap were not independently audited. The experiment measures retrospective sequential fact learning rather than a verified post-cutoff, real-time update benchmark.

Scope extension. Paired multi-form training, broader held-out prompt/distractor evaluation, and independent updater-family confirmation were not completed for the warm comparison. Prior TIES, SLAO, and AIMMerging screens do not establish an independent generality result; these methods are relevant comparison candidates (Yadav et al., 2023; Qiao & Mahdavi, 2025; Feng et al., 2025).

Free-generation validity. The free-generation alias-prefix grader has no blinded manual validation or decoding-length sensitivity check. A short cap or permissive alias may affect the automatic score; the main ranking claim relies on the separately defined multiple-choice evaluator.

Artifact availability. The accompanying anonymous analysis package regenerates the reported tables and figures. It does not contain the complete training corpus, checkpoints, or historical software environment, and an independently hosted training artifact has not been verified.