A Theory of Retrieval with Incremental Memory: Causal Influence of Writes and Predictive Gains

Hiroyuki Kasai
Waseda University, Tokyo, Japan
hiroyuki.kasai@waseda.jp

Abstract

Incremental-memory predictors store information as they process inputs, and a write can affect later queries and writes. We ask how omitting one write changes expected prediction loss and how to evaluate the predictive gain of whole memory against a fixed reference predictor. We develop TRIM (Theory of Retrieval with Incremental Memory) for causal query–read–update–write systems. We decompose signed single-write influence exactly into direct-deletion, regeneration, and query-change contributions. Under suitable conditions, the direct-deletion contribution vanishes as memory grows, but the regeneration contribution can remain nonzero even as the selected-write weight vanishes. For a fixed omitted write, we give sufficient conditions for the regeneration contribution to vanish. We also bound approximation errors for whole-memory and regenerated-memory gains, even when queries depend on stored values.

1 Introduction

Incremental-memory predictors store new information as they process inputs. A stored write can affect later retrieval, queries, and writes. Memory networks, memory-augmented language models, and associative memories, including Memory Mosaics, motivate this setting [14, 13, 20, 19]. Attention weights need not reliably reflect the importance of attended inputs for prediction [4, 9]. We ask two questions. How does omitting one write change expected prediction loss at a later time? How much does the full memory help prediction compared with a fixed reference predictor?

We develop TRIM (Theory of Retrieval with Incremental Memory) for causal query–read–update–write systems with append-only memory and bounded softmax retrieval. We omit one append after the state update. We then recompute later states, queries, and writes with the same external inputs and writer randomness. We keep the original biases of surviving writes fixed. Unlike methods for memory credit assignment [17, 16, 5], which estimate credit signals for memory operations, we analyze the signed change in expected prediction loss under this omission.

Our first result is an exact decomposition of signed single-write influence into direct-deletion, regeneration, and query-change contributions (Section 4.1). This matters as the context grows. Under the stated fixed-system conditions, the direct-deletion contribution is $O(T^{-1})$, but the regeneration contribution can remain nonzero even when the selected-write weight vanishes (Example 4.5). The omitted write can change many later writes. For a fixed omitted write, we give sufficient conditions for regeneration to vanish. These conditions do not ensure decay of the query-change contribution (Corollary 4.6). Our second result bounds approximation errors for the whole-memory and regenerated-memory gains against the same fixed reference (Theorems 5.1 and 5.2). Under the original-write statistical model, these bounds allow queries to depend on stored values without imposing that model on regenerated writes. When both gains are finite, the whole-memory gain minus the regenerated-memory gain equals signed single-write influence. Neither gain is guaranteed to be positive. Numerical examples show cancellation among contributions and changes in the signs

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of predictive gains (Section 6). Selected trajectories of a pretrained Memory Mosaics model show indirect readout changes larger than direct deletion after retrieval-time exclusion. These pathwise observations are not expected-loss results or evidence of typical behavior.

Ramesh [8] and Yao et al. [18] study exact recovery after record or observation omission, rather than signed prediction-loss bounds under a post-update append omission. Xiao et al. [15] study memory and feedback pathways of error propagation; Tang et al. [10] study memory-update utility. TRIM instead quantifies the specified write-omission influence and predictive gains. Detailed comparisons appear in Appendix A. All proofs are in the appendix. The assumptions for the main theoretical results are summarized in Appendix F.

2 Model and assumptions

Let $p \geq 0$ and $d \geq 1$ be integers, and let $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ be a measurable input space. We use the Euclidean inner products and norms on $\mathbb{R}^p$ and $\mathbb{R}^d$, with $\mathbb{R}^0 := \{0\}$; the ambient space is inferred from the arguments. Write $\mathbb{N} := \{1, 2, \ldots\}$, $\mathbb{B}_p(r) := \{h \in \mathbb{R}^p : \|h\| \leq r\}$ and $\mathbb{B}_d(r) := \{u \in \mathbb{R}^d : \|u\| \leq r\}$ for $r > 0$. For a sequence (x_i), write $x_{1:n} := (x_1, \ldots, x_n)$. All random objects share a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that a claim holds *pathwise* if it holds at every realization satisfying its hypotheses.

2.1 Causal memory model

Ordered memory and retrieval. Let $\mathcal{M}$ be the set of all finite ordered sequences of triples $(k_i, v_i, \beta_i) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}$, where β_i is a query-independent bias. Equip $\mathcal{M}$ with the disjoint-union Borel σ-algebra. An *ordered memory* $M \in \mathcal{M}$ has length $|M|$; $()$ denotes the empty memory, and $M \oplus (k, v, \beta)$ denotes appending the triple (k, v, β) to M. Fix $\tau > 0$ (the *temperature*). In the following readout definition, write $M = ((k_i, v_i, \beta_i))_{i=1}^{|M|}$ for $M \neq ()$: $i = 1, \ldots, |M|$ indexes memory *positions*, not the labels introduced below. This use of (k_i, v_i, β_i) is local to the generic-memory definition. For $q \in \mathbb{R}^d$, define the *readout map* $R_\tau : \mathbb{R}^d \times \mathcal{M} \to \mathbb{R}^p$, the *softmax weights*, and the *scores*, respectively, by

$$R_\tau(q; M) := \sum_{i=1}^{|M|} \pi_i(q; M) v_i, \quad \pi_i(q; M) := \frac{\exp(s_i(q; M))}{\sum_{\ell=1}^{|M|} \exp(s_\ell(q; M))}, \quad s_i(q; M) := -\frac{\|q - k_i\|^2}{2\tau} + \beta_i.$$

Set $R_\tau(q; ()) := 0$. The readout, softmax weights on nonempty memories, memory append, and fixed-position deletion are measurable on their respective domains. The letter s also denotes a time index below; a score is denoted by $s_i(q; M)$.

Causal Q–R–U–W loop. The inputs $z_t \in \mathcal{X}$, initial state $h_1 \in \mathbb{R}^p$, and exogenous writer randomness $\zeta_t \in \Xi$, where $(\Xi, \mathcal{F}_\Xi)$ is a measurable space, are random objects. Fix a measurable *query map* $\phi : \mathbb{R}^p \to \mathbb{R}^d$, measurable *update maps* $U_t : \mathbb{R}^p \times \mathcal{X} \times \mathbb{R}^p \to \mathbb{R}^p$, and measurable *writers* $W_t : (\mathbb{R}^p)^{t+1} \times \mathcal{X}^t \times \mathcal{M} \times \Xi^t \to \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}$, $t \geq 1$; write $W_t^{\mathrm{KV}} := \mathrm{pr}_{\mathrm{KV}} \circ W_t$ where $\mathrm{pr}_{\mathrm{KV}}(k, v, \beta) := (k, v)$. Starting from $M_0 := ()$, each step $t \geq 1$ uses the following order:

1. **Query (Q):** $q_t := \phi(h_t)$.
2. **Read (R):** $r_t := R_\tau(q_t; M_{t-1})$.
3. **Update (U):** $h_{t+1} := U_t(h_t, z_t, r_t)$.
4. **Write (W):** form the *post-update, pre-write history* $\mathsf{hist}_t := (h_{1:t+1}, z_{1:t}, M_{t-1}, \zeta_{1:t})$, set $(k_t, v_t, \beta_t) := W_t(\mathsf{hist}_t)$, and append this entry to form $M_t := M_{t-1} \oplus (k_t, v_t, \beta_t)$.

The resulting trajectory is the *original run*; (k_t, v_t, β_t) is its *original write* at time t and, once appended, an *original entry*. Unadorned write triples refer to original writes unless entries of a generic memory are explicitly under discussion. The writer is *causal*: W_t uses only hist_t, not z_s or ζ_s for $s > t$; this restricts its arguments, not statistical independence. Its domain includes *off-path* histories. Here "causal" describes the Q–R–U–W chronology; in single-write comparisons it refers to omitting one write. Since writing follows the update, a write cannot affect q_t, r_t, or h_{t+1}, but can affect later retrieval, queries, and writer outputs through subsequent memories and states.

Causal memory model and its base conditions. A *causal memory model* is a Q–R–U–W loop satisfying the four *base conditions* (B1)–(B4) below. Fix deterministic $K, V > 0$, $\kappa_\beta \geq 0$, and $H > 0$. Call $\mathbb{B}_p(H)$ the *state ball* and $\mathbb{B}_d(K)$ the *query ball*.

(B1) Original key–value bounds. Assume $\|k_i\| \le K$ and $\|v_i\| \le V$ hold almost surely (a.s.) for all original entries.

(B2) Uniform bias-width bound. Assume the original biases satisfy $\sup_{i,\ell\ge 1} |\beta_i - \beta_\ell| \le \kappa_\beta$ a.s. For later bounds, set $A := 2K^2/\tau + \kappa_\beta$.

(B3) Uniform query-map bound. Assume $\sup_{h\in\mathbb{B}_p(H)} \|\phi(h)\| \le K$.

(B4) Original state-trajectory boundedness. Assume $h_t \in \mathbb{B}_p(H)$ for all $t \ge 1$ on a common measurable probability-one event. Thus, by (B3), $q_t \in \mathbb{B}_d(K)$ on that event.

Original submemories. An entry appended at time t carries the *write-time label* t; this applies to original entries and to those constructed in Section 3. For a finite *deterministic* set $I \subseteq \mathbb{N}$, let the *original submemory* M_I consist of the original entries with write-time labels in I, in increasing label order and with their original biases. Put $I_m := \{1, \dots, m\}$ for $m \ge 1$ (with $I_0 := \varnothing$); for $m \ge 0$ and $j \in \mathbb{N}$, put $I_m^{-j} := I_m \setminus \{j\}$. For nonempty I and $q \in \mathbb{R}^d$, define the weight of write-time label $i \in I$ directly by

$$\pi_i^I(q) := \frac{\exp(-\|q - k_i\|^2/(2\tau) + \beta_i)}{\sum_{\ell\in I} \exp(-\|q - k_\ell\|^2/(2\tau) + \beta_\ell)}.$$

Set $R_I(q) := R_\tau(q; M_I) = \sum_{i\in I} \pi_i^I(q) v_i$ for nonempty I, and $R_\varnothing(q) := 0$. In particular, $M_\varnothing = ()$ and $M_{I_t} = M_t$ for every $t \ge 0$. The indices in π_i^I are write-time labels, not memory positions.

Application to particular architectures requires separate verification (Appendix B.1).

2.2 Original-write statistical model and prediction task

Let $(\Theta, \mathcal{F}_\Theta)$ be a measurable latent-condition space, θ a Θ-valued random element (the *latent condition*), and ϑ a generic deterministic element of Θ. Fix a measurable *target map* $f : \Theta \times \mathcal{X} \to \mathbb{R}^p$; set $y_t := f(\theta, z_t)$ for $t \ge 1$ and $\ell_y(r) := \|r - y\|^2$ (the *squared loss*) for $r, y \in \mathbb{R}^p$. At evaluation time T, the original *pre-update readout* r_T (computed before z_T enters the update) predicts y_T and is evaluated by ℓ_{y_T}; comparison readouts are defined in Section 3 and evaluated in Section 4.1. Conditions (S1)–(S4) form the *original-write statistical model*; they apply only where stated. Conditions (S1) and (S4) concern the shared inputs and targets, while (S2) and (S3) constrain original writes.

(S1) Conditional-i.i.d. input law. The inputs $(z_t)_{t\ge 1}$ are conditionally i.i.d. given θ.

(S2) Original-write observation rule. Let $\Phi_{\mathrm{K}} : \mathcal{X} \to \mathbb{B}_d(K)$ and $\Phi_{\mathrm{V}} : \mathcal{X} \to \mathbb{B}_p(V)$ be measurable *observation maps*. The *original-write observation rule* holds if $(k_i, v_i) = (\Phi_{\mathrm{K}}(z_i), \Phi_{\mathrm{V}}(z_i))$ for all *original* writes a.s. (S2) does not constrain W_i^{KV} on off-path histories. Since the observation maps take values in the radius-K and radius-V balls, (S2) implies (B1) for original writes; (B1) remains a separate base condition for results not assuming (S2).

(S3) Original-bias key measurability. For $m \ge 1$, write $\mathcal{G}_m := \sigma(\theta, k_{1:m})$ for the information in the latent condition and the first m original keys. *Original-bias key measurability* holds if, for each $i \ge 1$, the *original* bias β_i is $\mathcal{G}_i$-measurable. This is a restriction on the information entering the original softmax weights, separate from the uniform bias-width bound (B2). It imposes no unconditional independence between the biases and the original values.

(S4) Square-integrable prediction target. Assume $\mathbb{E}\|f(\theta, z_1)\|^2 < \infty$.

Remark (Target second moments). Under (S1), $y_T = f(\theta, z_T)$ and y_1 have the same law for every $T \ge 1$. Together with (S4), this gives

$$\mathbb{E}\|y_T\|^2 = \mathbb{E}\|y_1\|^2 = \mathbb{E}\|f(\theta, z_1)\|^2 < \infty \qquad (T \ge 1). \tag{1}$$

In particular, the target second moments are uniformly bounded over evaluation times.

Under (S1)–(S2), original key–value pairs are conditionally i.i.d. given θ, whereas original biases need not be. The original-write statistical model does not impose the observation rule (S2), or the conditional-i.i.d. law, on writer-generated key–value outputs from off-path histories, including later regenerated keys and values in Section 3; both comparisons retain each surviving label's original bias. The pre-update readout precedes the entry of z_T into the update; observing a current test covariate before prediction, as in in-context regression, is not ensured by these assumptions alone.

Remark 2.1 (Fixed-system convention)**.** A *fixed system* retains the probability space, jointly defined $\theta, h_1, (z_t), (\zeta_t)$, and all model spaces, maps, and constants as T varies. This fixes the jointly defined random objects, not an individual sample realization. The runs defined in Section 3 use these same objects. Fixed-T bounds are *finite-horizon*; *long-context* statements take $T \to \infty$ within the system.

2.3 Conditional-mean–residual decomposition

We decompose original-write readouts into conditional-mean and residual aggregates and establish a uniform residual bound for adaptive queries.

Conditional mean and residuals. Under (S1)–(S2), fix a jointly measurable $g : \Theta \times \mathbb{R}^d \to \mathbb{R}^p$ satisfying $g(\theta, k_1) = \mathbb{E}[v_1 \mid \theta, k_1]$ a.s. and $\|g(\vartheta, k)\| \le V$ everywhere (Appendix B.2.1). Set $g_\theta(k) := g(\theta, k)$ and define the *value residual* $\epsilon_i := v_i - g_\theta(k_i)$ for every original write-time label i. By (S1)–(S2), (θ, k_i, v_i) has the same law for every i; hence $g_\theta(k_i) = \mathbb{E}[v_i \mid \theta, k_i]$, $\mathbb{E}[\epsilon_i \mid \theta, k_i] = 0$, and $\|\epsilon_i\| \le 2V$ a.s. for every i (Appendix B.2.2). The value residual ϵ_i is distinct from writer randomness ζ_i.

Readout aggregates. For nonempty finite deterministic $I \subseteq \mathbb{N}$ and $u \in \mathbb{R}^d$, define the *conditional-mean aggregate* $R_I^{\mathrm{mean}}(u) := \sum_{i \in I} \pi_i^I(u) g_\theta(k_i)$ and the *residual aggregate* $R_I^{\mathrm{res}}(u) := \sum_{i \in I} \pi_i^I(u) \epsilon_i$. Since $v_i = g_\theta(k_i) + \epsilon_i$, the decomposition holds pathwise:

$$R_I(u) = \sum_{i \in I} \pi_i^I(u) v_i = \sum_{i \in I} \pi_i^I(u)(g_\theta(k_i) + \epsilon_i) = R_I^{\mathrm{mean}}(u) + R_I^{\mathrm{res}}(u).$$

These aggregates use original writes and their original biases; R_I^{mean} need not be the conditional expectation of the readout evaluated at a residual-dependent random query.

Residual orthogonality and uniform bound. Under (S1)–(S3), for $m \ge 1$, deterministic nonempty $I \subseteq I_m$, and deterministic u, the weights $\pi_i^I(u)$ are $\mathcal{G}_m$-measurable, and the summands $\pi_i^I(u)\epsilon_i$ have zero expected inner products for distinct $i, \ell \in I$ (Lemma B.1, Appendix B.2.3). The following lemma extends the resulting second-moment bound to residual-dependent queries.

Lemma 2.2 (Uniform residual bound for adaptive queries)**.** *Under (S1)–(S3) and (B1)–(B2), set* $\alpha := 1/(d+2)$, $c_\epsilon := 8V^2 \exp(A) 3^d + 32K^4V^2/\tau^2$, *and* $\delta_n := \sqrt{c_\epsilon}\, n^{-\alpha}$ *for* $n \ge 1$. *Then, for every* $m \ge 1$, $\mathbb{E}\left[\sup_{u \in \mathbb{B}_d(K)} \|R_{I_m}^{\mathrm{res}}(u)\|^2\right] \le c_\epsilon m^{-2\alpha}$. *Consequently, any measurable, possibly residual-dependent* $U \in \mathbb{B}_d(K)$ *a.s. satisfies* $\mathbb{E}\|R_{I_m}^{\mathrm{res}}(U)\|^2 \le \delta_m^2$.

The proof is in Appendix B.2.4.

3 Single-write causal comparisons

This section defines the single-write omission intervention and the associated runs, using the two senses of "causal" distinguished in Section 2.1. Fix a deterministic *omitted write-time label* $j \ge 1$ and an evaluation time $T \ge \max\{3, j+1\}$; set $m := T - 1$. Then $j \le m$, and I_m^{-j} has $m - 1 \ge 1$ labels. For each s, labels in I_s^{-j} are *surviving labels*; entries carrying them in a given run's memory are its *surviving entries*.

An *intervention* selects the original write with label j (the *selected write*) and omits its append after the state update at time j in both comparisons. As this does not specify how later writes are handled, we define the primary *regenerated-write comparison* and auxiliary *frozen-write comparison*. Their trajectories, the *regenerated run* and *frozen run* (the *comparison runs*), carry superscripts R and F, respectively; both depend on j, which is suppressed in the notation. Both start from $(h_1^{\mathrm{R}}, M_0^{\mathrm{R}}) := (h_1, ())$ and $(h_1^{\mathrm{F}}, M_0^{\mathrm{F}}) := (h_1, ())$, share inputs and writer randomness with the original run, and recompute later states, queries, and readouts. Both retain each surviving label's original bias but differ in how surviving keys and values are formed.

Regenerated-write comparison. The regenerated run recomputes later key–value writes using the same writer while retaining the original biases β_s. The omission leaves input z_j and its state update unchanged. Later original biases, possibly dependent on the omitted key, retain their original labels

and are not recomputed or renumbered. For $s \geq 1$,

$$q_s^{\mathrm{R}} := \phi(h_s^{\mathrm{R}}), \qquad r_s^{\mathrm{R}} := R_\tau(q_s^{\mathrm{R}}; M_{s-1}^{\mathrm{R}}), \qquad h_{s+1}^{\mathrm{R}} := U_s(h_s^{\mathrm{R}}, z_s, r_s^{\mathrm{R}}),$$
$$M_s^{\mathrm{R}} := \begin{cases} M_{s-1}^{\mathrm{R}}, & s = j, \\ M_{s-1}^{\mathrm{R}} \oplus (k_s^{\mathrm{R}}, v_s^{\mathrm{R}}, \beta_s), & s \neq j, \end{cases}$$

where, for $s \neq j$, $(k_s^{\mathrm{R}}, v_s^{\mathrm{R}}) := W_s^{\mathrm{KV}}(\mathsf{hist}_s^{\mathrm{R}})$ with the post-update, pre-write history $\mathsf{hist}_s^{\mathrm{R}} := (h_{1:s+1}^{\mathrm{R}}, z_{1:s}, M_{s-1}^{\mathrm{R}}, \zeta_{1:s})$. For $s \neq j$, we call $(k_s^{\mathrm{R}}, v_s^{\mathrm{R}}, \beta_s)$ the *regenerated write* at time s and, once appended, a *regenerated entry*. At evaluation time T, the query q_T^{R} is the *regenerated query*. We say that the *regenerated key–value bounds* hold if $\|k_s^{\mathrm{R}}\| \leq K$ and $\|v_s^{\mathrm{R}}\| \leq V$ a.s. for $s \in I_m^{-j}$. The *regenerated value bounds* require only the value inequalities, for the same labels and almost-sure qualification.

Frozen-write comparison. To isolate later-write regeneration, the frozen run computes states h_s^{F}, queries q_s^{F}, and readouts r_s^{F} using the query, read, and update operations of Section 2.1, reading M_{s-1}^{F} at q_s^{F}. Instead of generating new entries, it retains the surviving original entries $M_s^{\mathrm{F}} := M_{I_s^{-j}}$ for $s \geq 1$ (Appendix C.2). The decomposition in Section 4.1 uses the frozen memory M_{T-1}^{F}, not the frozen state trajectory. The full frozen-run recursion is also used in the exact-reduction argument.

Common properties of comparison runs. At steps $s < j$, both comparison runs share the original states, queries, and memories, and the regenerated run also shares the original writes. The states of all three runs agree through h_{j+1}. We call these facts the *pre-omission coupling* (Appendix C, Sections C.1 and C.2). The states, queries, and memories of both comparison runs are measurable by finite recursion using the measurable maps and memory operations of Section 2.1. The base conditions (B3)–(B4) bound the original states and queries, but impose no corresponding bounds on $h_s^{\mathrm{F}}, q_s^{\mathrm{F}}, h_s^{\mathrm{R}}, q_s^{\mathrm{R}}$.

Run notation and mixed readouts. Upright O, F, and R denote run tags; italic R (as in R_τ and R_I) denotes a readout map, not a run. For the original run, set $q_s^{\mathrm{O}} := q_s$ and $M_s^{\mathrm{O}} := M_s$. For $X, Y \in \{\mathrm{O}, \mathrm{F}, \mathrm{R}\}$, define the *mixed readout* $r_s^{X|Y} := R_\tau(q_s^X; M_{s-1}^Y)$, where X identifies the query run and Y the memory run. In particular, $r_s^{\mathrm{O}|\mathrm{O}} = r_s$, $r_s^{\mathrm{F}|\mathrm{F}} = r_s^{\mathrm{F}}$, and $r_s^{\mathrm{R}|\mathrm{R}} = r_s^{\mathrm{R}}$. The superscript $X \mid Y$ is reserved for mixed readouts. For scalar quantities, signed single-write influences use the tag R, gains use O or R, contributions use R together with their contribution type, and readout deviations comparing frozen and regenerated memories use FR; none uses a query–memory pair. Dependence on j is suppressed in mixed readouts involving a comparison run. When $j = j(T)$ varies with T, comparison quantities at each evaluation time use runs omitting $j(T)$, not a single comparison trajectory across evaluation times. Single-write comparison quantities at evaluation time, such as influences, contributions, and deviations, carry the indices (j, T).

Exact reduction to the frozen-write comparison. If $k_s^{\mathrm{R}} = k_s$ and $v_s^{\mathrm{R}} = v_s$ a.s. for every surviving label s with $j < s \leq T - 1$, then, by induction on the common update recursion, the states, queries, readouts, and pre-read memories of the regenerated and frozen runs agree a.s. at each time $s \leq T$; in particular, $r_T^{\mathrm{R}|\mathrm{R}} = r_T^{\mathrm{F}|\mathrm{F}}$ a.s. (The induction is given in Appendix C.2.)

4 Signed single-write influence and long-context characterization

We quantify the signed prediction-loss effect of omitting one write through an exact decomposition, finite-horizon bounds, and conditional long-context results.

4.1 Decomposition of signed single-write influence

We derive an exact *three-term decomposition* of signed single-write influence by comparing four readouts at evaluation time T: $r_T^{\mathrm{O}|\mathrm{O}}$, $r_T^{\mathrm{R}|\mathrm{O}}$, $r_T^{\mathrm{R}|\mathrm{F}}$, and $r_T^{\mathrm{R}|\mathrm{R}}$. They form an *algebraic comparison sequence*: change the query while retaining original memory, delete write j at the regenerated query, and then replace surviving original key–value entries by regenerated ones. The middle two readouts share the regenerated query and use full original and frozen memory, respectively; neither uses q_T^{F}. This sequence decomposes an evaluation-time loss difference; it is not the chronology of the intervention. Each signed expectation below is defined only when its displayed loss difference is

integrable; individual expected losses need not be finite. Define the *signed single-write influence* directly through the loss difference

$$D^{\mathrm{R}}_{j,T} := \mathbb{E}[\ell_{y_T}(r_T^{\mathrm{R|R}}) - \ell_{y_T}(r_T^{\mathrm{O|O}})].$$

Positive (negative) $D^{\mathrm{R}}_{j,T}$ means that the expected signed change in loss on omitting write j is positive (negative).

Define the *direct-deletion contribution* $C^{\mathrm{R,del}}_{j,T}$ at the regenerated query, the *regeneration contribution* $C^{\mathrm{R,reg}}_{j,T}$ from later-write changes at the common regenerated query, and the *query-change contribution* $C^{\mathrm{R,qry}}_{j,T}$ between two queries on full original memory: $C^{\mathrm{R,del}}_{j,T} := \mathbb{E}[\ell_{y_T}(r_T^{\mathrm{R|F}}) - \ell_{y_T}(r_T^{\mathrm{R|O}})]$, $C^{\mathrm{R,reg}}_{j,T} := \mathbb{E}[\ell_{y_T}(r_T^{\mathrm{R|R}}) - \ell_{y_T}(r_T^{\mathrm{R|F}})]$, and $C^{\mathrm{R,qry}}_{j,T} := \mathbb{E}[\ell_{y_T}(r_T^{\mathrm{R|O}}) - \ell_{y_T}(r_T^{\mathrm{O|O}})]$. In $C^{\mathrm{R,type}}_{j,T}$, the run tag R specifies the regenerated-write comparison and $\mathrm{type} \in \{\mathrm{del}, \mathrm{reg}, \mathrm{qry}\}$ specifies deletion, regeneration, or query change, not the query–memory pair of $r_T^{X|Y}$. Adding and subtracting the losses at $r_T^{\mathrm{R|F}}$ and $r_T^{\mathrm{R|O}}$ gives the three-term decomposition

$$D^{\mathrm{R}}_{j,T} = C^{\mathrm{R,del}}_{j,T} + C^{\mathrm{R,reg}}_{j,T} + C^{\mathrm{R,qry}}_{j,T}. \tag{2}$$

The contributions depend on the intermediate readouts, but their sum depends only on the original and regenerated endpoints. The decomposition requires neither independence nor a conditional-mean–residual decomposition.

To establish finiteness, let a, b be any two of the four readouts. By (B1) and the regenerated value bounds, $\|a\|, \|b\| \le V$ a.s. The squared-loss identity gives

$$|\ell_{y_T}(a) - \ell_{y_T}(b)| = |\langle a - b, a + b - 2y_T\rangle| \le \|a - b\|(\|a\| + \|b\| + 2\|y_T\|) \le 4V(V + \|y_T\|).$$

Thus all displayed loss differences are integrable if $\mathbb{E}\|y_T\| < \infty$. Under (S1) and (S4), (1) and Cauchy–Schwarz give $\mathbb{E}\|y_T\| \le \sqrt{\mathbb{E}\|y_T\|^2} < \infty$.

4.2 Finite-horizon contribution bounds and conditional-mean-aggregate approximation

We establish finite-horizon bounds for direct-deletion and regeneration and control the error of the conditional-mean-aggregate approximation to the query-change contribution.

Lemma 4.1 (Squared-loss sensitivity)**.** *If $\mathbb{E}\|y_T\|^2 < \infty$, put $L_* := 2(V + \sqrt{\mathbb{E}\|y_T\|^2})$. For random predictions x, x' with $\|x\|, \|x'\| \le V$ a.s., $|\mathbb{E}\ell_{y_T}(x) - \mathbb{E}\ell_{y_T}(x')| \le L_*\sqrt{\mathbb{E}\|x - x'\|^2}$, without any independence requirement.*

The proof is in Appendix D.1.1. In general L_* depends on T. Under (S1) and (S4), (1) makes L_* independent of T.

Bounds and approximation for the three contributions. We now control the three contributions in (2) using estimates tailored to each contribution.

(i) Direct-deletion contribution: $C^{\mathrm{R,del}}_{j,T}$. For a nonempty memory M, its *score width* at a query q is $\max_{1\le i\le|M|} s_i(q; M) - \min_{1\le i\le|M|} s_i(q; M)$. Under (B1)–(B2), for each nonempty original submemory M_I and query in $\mathbb{B}_d(K)$, this width is at most A. The same upper bound applies to any other nonempty memory whose keys have norm at most K and whose bias width is at most κ_β; (B1)–(B2) alone do not ensure these conditions for regenerated memories. Set $w_n := \min\{1, \exp(A)/n\}$ for $n \ge 1$.

Lemma 4.2 (Direct-deletion retrieval bound)**.** *For any read time $s > j$ and $q \in \mathbb{R}^d$, let $p_{j,s}(q) := \pi_j^{I_{s-1}}(q)$ denote the* selected-write weight*, i.e., the softmax weight of original write-time label j. Without boundedness assumptions, $R_{I_{s-1}}(q) - R_{I_{s-1}^{-j}}(q) = p_{j,s}(q)\big(v_j - R_{I_{s-1}^{-j}}(q)\big)$. Under (B1), its norm is at most $2Vp_{j,s}(q)$. Under (B1)–(B2), simultaneously for every $q \in \mathbb{B}_d(K)$, a.s., $p_{j,s}(q) \le w_{s-1}$ and hence $\|R_{I_{s-1}}(q) - R_{I_{s-1}^{-j}}(q)\| \le 2Vw_{s-1}$.*

The identity shows that the readout change is small when the selected-write weight is small or the selected value is close to the surviving-memory readout. The proof, including the empty-memory case, is in Appendix D.1.2.

(ii) Regeneration contribution: $C_{j,T}^{\mathrm{R,reg}}$. At the common regenerated query, $r_T^{\mathrm{R|F}}$ and $r_T^{\mathrm{R|R}}$ differ through subsequent key–value regeneration. Define the *frozen–regenerated deviation* by $\mu_{j,T}^{\mathrm{FR}} := \mathbb{E}\|r_T^{\mathrm{R|R}} - r_T^{\mathrm{R|F}}\|$. Here FR denotes the frozen and regenerated memories evaluated at the common regenerated query, not a query–memory pair. This is a deterministic element of $[0,\infty]$, finite under the value bounds of Proposition 4.3(ii). Under those value bounds the readout difference has norm at most $2V$; its second moment is at most $2V\mu_{j,T}^{\mathrm{FR}}$. Lemma 4.1 converts this into the regeneration bound. Unlike the direct-deletion bound, this requires no key or bias-width assumption.

(iii) Query-change contribution: $C_{j,T}^{\mathrm{R,qry}}$. Here $r_T^{\mathrm{R|O}}$ and $r_T^{\mathrm{O|O}}$ use the same full original memory but different queries. The bounds above provide no decay guarantee for this contribution. Instead, under the original-write statistical model, introduce the *conditional-mean-aggregate query-change counterpart*

$$\widetilde{C}_{j,T}^{\mathrm{R,qry}} := \mathbb{E}\big[\ell_{y_T}(R_{I_m}^{\mathrm{mean}}(q_T^{\mathrm{R}})) - \ell_{y_T}(R_{I_m}^{\mathrm{mean}}(q_T))\big].$$

The tilde denotes replacement of the full original readout by its conditional-mean aggregate; it does not imply decay of the contribution. Apply Lemma 2.2 separately at q_T^{R} and q_T. By Lemma 4.1, each full-readout versus conditional-mean-aggregate loss difference is at most $L_*\delta_m$, and their signed difference is therefore at most $2L_*\delta_m$. The two-query comparison uses Lemma D.1 in Appendix D.1.3, with $(U, U') = (q_T^{\mathrm{R}}, q_T)$. The two queries need not be independent of the original residuals.

The preceding arguments yield the following three bounds under their respective assumptions.

Proposition 4.3 (Finite-horizon contribution bounds and conditional-mean-aggregate approximation)**.** *Fix $T \geq 3$, put $m := T-1$, fix a deterministic $j \in I_m$, and assume $\mathbb{E}\|y_T\|^2 < \infty$. The first two bounds do not require the original-write statistical model.*

(i) *Under (B1)–(B2) of Section 2.1, with $q_T^{\mathrm{R}} \in \mathbb{B}_d(K)$ a.s., $|C_{j,T}^{\mathrm{R,del}}| \leq 2VL_*w_m$.*

(ii) *If $\|v_s\|, \|v_s^{\mathrm{R}}\| \leq V$ a.s. for every surviving label $s \in I_m^{-j}$, then $|C_{j,T}^{\mathrm{R,reg}}| \leq L_*\sqrt{2V\mu_{j,T}^{\mathrm{FR}}}$; no key or bias-width bounds are needed.*

(iii) *Under the hypotheses of (i), conditions (B3)–(B4), and (S1)–(S4), $|C_{j,T}^{\mathrm{R,qry}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \leq 2L_*\delta_m$.*

The proof is in Appendix D.1.4. The original query is bounded under (B3)–(B4). These estimates require no bound on intermediate comparison states or statistical model for regenerated writes; the regenerated query may depend on the original residuals. Only the direct-deletion contribution has an explicit length-decaying magnitude bound; the query-change estimate controls approximation error rather than the magnitude of the contribution itself.

4.3 Long-context characterization

We combine the contribution bounds into conditional long-context equivalences, then demonstrate persistent regeneration and give sufficient conditions for its decay.

Theorem 4.4 (Finite-horizon bounds and long-context equivalences)**.** *At any evaluation time $T \geq 3$, put $m := T-1$ and fix a deterministic omitted write-time label $j \in I_m$. Assume the causal memory model (B1)–(B4) of Section 2.1, the regenerated value bounds defined for the regenerated-write comparison in Section 3, $q_T^{\mathrm{R}} \in \mathbb{B}_d(K)$ a.s., and $\mathbb{E}\|y_T\|^2 < \infty$. Use L_* in Section 4.2.* Finite-horizon bounds. *The following three inequalities hold.*

(i) Direct-deletion characterization. *Without invoking the original-write statistical model,* $\big|D_{j,T}^{\mathrm{R}} - (C_{j,T}^{\mathrm{R,reg}} + C_{j,T}^{\mathrm{R,qry}})\big| \leq 2VL_*w_m$.

(ii) Conditional-mean-aggregate refinement. *If, in addition, the original-write statistical model (S1)–(S4) of Section 2.2 holds, then* $\big|D_{j,T}^{\mathrm{R}} - C_{j,T}^{\mathrm{R,reg}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}\big| \leq 2VL_*w_m + 2L_*\delta_m$.

(iii) Frozen–regenerated deviation refinement. *Under the hypotheses of (ii), one also has* $\big|D_{j,T}^{\mathrm{R}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}\big| \leq L_*\big(2Vw_m + 2\delta_m + \sqrt{2V\mu_{j,T}^{\mathrm{FR}}}\big)$.

Long-context consequences. *Consider one fixed system (Remark 2.1) satisfying (B1)–(B4). For each $T \geq 3$, choose a deterministic $j = j(T) \in I_{T-1}$ and assume the regenerated value bounds*

and $q_T^{\mathrm{R}} \in \mathbb{B}_d(K)$ *a.s. at that evaluation time. Assume* $\sup_{T\geq 3}\mathbb{E}\|y_T\|^2 < \infty$ *(the* uniform target second-moment bound*). By part (i) and* (2)*,*

$$|C_{j,T}^{\mathrm{R,del}}| \leq 2VL_* w_{T-1} \leq \frac{2VL_*\exp(A)}{T-1} = O(T^{-1}),$$

where the last estimate follows from $\sup_{T\geq 3} L_* < \infty$ *under the uniform target second-moment bound. If the fixed system also satisfies (S1)–(S4), the right-hand side of (ii) tends to zero. If in addition* $\mu_{j,T}^{\mathrm{FR}} \to 0$*, the right-hand side of (iii) tends to zero. Consequently, when the original-write statistical model holds and* $\mu_{j,T}^{\mathrm{FR}} \to 0$*, the conclusions form one chain:*

$$D_{j,T}^{\mathrm{R}} \to 0 \iff C_{j,T}^{\mathrm{R,reg}} + C_{j,T}^{\mathrm{R,qry}} \to 0 \iff C_{j,T}^{\mathrm{R,reg}} + \widetilde{C}_{j,T}^{\mathrm{R,qry}} \to 0 \iff \widetilde{C}_{j,T}^{\mathrm{R,qry}} \to 0.$$

The first equivalence requires no original-write statistical model; the second additionally uses that model; the third is ensured by $\mu_{j,T}^{\mathrm{FR}} \to 0$*, which is a sufficient condition, not claimed to be necessary.*

The proof is in Appendix D.2.1. For part (i) alone, (B3)–(B4) are not used: (B1)–(B2), the regenerated value bounds, the stated regenerated query-ball condition, and $\mathbb{E}\|y_T\|^2 < \infty$ suffice. The stronger common hypotheses are retained for parts (ii)–(iii).

$\widetilde{C}_{j,T}^{\mathrm{R,qry}}$ is a theoretical quantity. The first two equivalences concern sums of signed terms, not the vanishing of each contribution individually. No part requires intermediate comparison-state bounds.

Persistence despite vanishing selected-write weight. The following example shows that vanishing selected-write weight does not necessarily imply vanishing regeneration contribution.

Example 4.5 (Persistent regeneration contribution (short version))**.** For every fixed $j \geq 2$, there exists one fixed system satisfying (B1)–(B4), (S1)–(S4), and the regenerated key–value bounds such that, for every $T \geq j+1$, the selected-write weight and regeneration contribution are $p_{j,T}(q_T^{\mathrm{R}}) = \frac{1}{T-1} \longrightarrow 0$, and $C_{j,T}^{\mathrm{R,reg}} = \left(\frac{T-j-1}{T-2}\right)^2 \longrightarrow 1$. Thus direct-deletion decay need not imply regeneration decay.

The full construction and verification are in Example D.2 of Appendix D.2.2.

Sufficient conditions for vanishing regeneration contribution. In this example, every post-omission value changes by unit magnitude, so the normalized sum of these changes does not vanish as $T \to \infty$. The next corollary shows that convergence in probability of a normalized sum of key–value differences suffices for regeneration decay; no converse is claimed.

Corollary 4.6 (Vanishing regeneration contribution under diminishing average key–value perturbations)**.** *Fix a deterministic* $j \geq 1$ *and one fixed system satisfying (B1)–(B2). At every* $T \geq \max\{3, j+1\}$*, assume the regenerated key–value bounds and* $q_T^{\mathrm{R}} \in \mathbb{B}_d(K)$ *a.s.; also assume* $\sup_T \mathbb{E}\|y_T\|^2 < \infty$ *(the uniform target second-moment bound). Define the* average key–value perturbation

$$\mathcal{A}_{j,T} := \frac{\exp(A)}{T-2}\sum_{s=j+1}^{T-1}\left(\|v_s^{\mathrm{R}} - v_s\| + \frac{4KV}{\tau}\|k_s^{\mathrm{R}} - k_s\|\right).$$

Then

$$\mu_{j,T}^{\mathrm{FR}} \leq \sqrt{\mathbb{E}[\min\{2V, \mathcal{A}_{j,T}\}^2]}, \qquad |C_{j,T}^{\mathrm{R,reg}}| \leq L_*\sqrt{\mathbb{E}[\min\{2V, \mathcal{A}_{j,T}\}^2]}.$$

Consequently, if $\mathcal{A}_{j,T} \to 0$ *in probability, then* $\mu_{j,T}^{\mathrm{FR}} \to 0$ *and* $C_{j,T}^{\mathrm{R,reg}} \to 0$*.*

If the system also satisfies (B3)–(B4) and (S1)–(S4), then

$$|D_{j,T}^{\mathrm{R}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \leq L_*\left(2Vw_{T-1} + 2\delta_{T-1} + \sqrt{\mathbb{E}[\min\{2V, \mathcal{A}_{j,T}\}^2]}\right),$$

and this difference tends to zero when $\mathcal{A}_{j,T} \to 0$ *in probability.*

In particular, $N_{j,T}/(T-2) \to 0$ *in probability suffices for these limits under their respective assumptions, where* $N_{j,T}$ *counts surviving labels with changed keys or values. If, almost surely, the sum defining* $\mathcal{A}_{j,T}$ *is bounded by a deterministic constant uniformly in* T*, then* $\mu_{j,T}^{\mathrm{FR}} = O(T^{-1})$ *and* $|C_{j,T}^{\mathrm{R,reg}}| = O(T^{-1})$*. This holds, in particular, when* $N_{j,T} \leq n_0$ *a.s. uniformly in* T *for some deterministic* $n_0 < \infty$*.*

The proof is in Appendix D.2.3. No state contraction or pointwise convergence of each regenerated write is required.

5 Predictive gains of whole and regenerated memory

While signed single-write influence measures the expected loss change caused by omitting a write, we study the predictive gains of whole and regenerated memory relative to a common fixed prediction and establish bounds for their approximations.

5.1 Whole-memory gain approximation

We bound the approximation error obtained by replacing the original readout with its conditional-mean-aggregate counterpart at the original query. For $T \geq 2$, put $m := T-1$ and use the original pre-update readout $r_T = R_\tau(q_T; M_{T-1}) = R_{I_m}(q_T)$ of Section 2.2 as prediction. Here *whole memory* means the full original memory, without omitting any write. Fix a common deterministic *reference predictor* $r_{\mathrm{ref}} \in \mathbb{R}^p$ and define the *whole-memory gain* and *whole-memory conditional-mean-aggregate gain* by

$$G_T^{\mathrm{O}} := \mathbb{E}\ell_{y_T}(r_{\mathrm{ref}}) - \mathbb{E}\ell_{y_T}(r_T^{\mathrm{O|O}}), \qquad \widetilde{G}_T^{\mathrm{O}} := \mathbb{E}\ell_{y_T}(r_{\mathrm{ref}}) - \mathbb{E}\ell_{y_T}(R_{I_m}^{\mathrm{mean}}(q_T)).$$

Both gains use the same target and reference; positive, zero, and negative values correspond to lower, equal, and higher expected loss than r_{ref}, respectively. The superscript O identifies the original run. The theoretical comparator $R_{I_m}^{\mathrm{mean}}(q_T)$ (Section 2.3) replaces original values v_i by $g_\theta(k_i)$, retaining the original query, labels, keys, and biases; it is not assumed observable.

Under (B1) and $\mathbb{E}\|y_T\|^2 < \infty$, both gains are finite; the moment condition follows from (S1) and (S4) by (1). Under the theorem's assumptions, (B3)–(B4) give $q_T \in \mathbb{B}_d(K)$ a.s. Since I_m is deterministic, the uniform residual bound of Lemma 2.2 and the squared-loss sensitivity of Lemma 4.1 apply even when q_T depends on the original value residuals.

Theorem 5.1 (Whole-memory gain approximation)**.** *Under the causal memory model (B1)–(B4) and the original-write statistical model (S1)–(S4) (using the decomposition of Section 2.3), for evaluation time $T \geq 2$, put $m := T-1$ and use the original query $q_T = \phi(h_T)$ in both gains. Then*

$$|G_T^{\mathrm{O}} - \widetilde{G}_T^{\mathrm{O}}| \leq L_* \delta_m.$$

The proof is in Appendix E.1.

5.2 Regenerated-memory gain approximation

We next control the approximation error obtained by using the original-memory conditional-mean aggregate at the regenerated query, accounting for deletion and regeneration effects. For $T \geq 3$, put $m := T-1$ and fix a deterministic omitted write-time label $j \in I_m$. Define the *regenerated-memory gain*, evaluated at the regenerated query q_T^{R}, by $G_{j,T}^{\mathrm{R}} := \mathbb{E}\ell_{y_T}(r_{\mathrm{ref}}) - \mathbb{E}\ell_{y_T}(r_T^{\mathrm{R|R}})$. When both gains are finite, the signed single-write influence satisfies $D_{j,T}^{\mathrm{R}} = G_T^{\mathrm{O}} - G_{j,T}^{\mathrm{R}}$. This is the case, for example, under (B1), the regenerated value bounds, and $\mathbb{E}\|y_T\|^2 < \infty$.

Theorem 5.2 (Whole-memory-to-regenerated-memory gain approximation)**.** *Under the hypotheses of Proposition 4.3 (i)–(ii) and (S1)–(S4),*

$$\left|G_{j,T}^{\mathrm{R}} - (\widetilde{G}_T^{\mathrm{O}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}})\right| \leq L_*\left(\delta_m + 2V w_m + \sqrt{2V\mu_{j,T}^{\mathrm{FR}}}\right).$$

The approximation centers in the two theorems are $\widetilde{G}_T^{\mathrm{O}}$ and $\widetilde{G}_T^{\mathrm{O}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}$, respectively. The proof is in Appendix E.2. Directly combining Theorem 5.1 with Theorem 4.4(iii) would yield $3L_*\delta_m$ rather than $L_*\delta_m$ in the error bound, under the stronger assumptions of the latter theorem. Indeed, the definitions give $\widetilde{G}_T^{\mathrm{O}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}} = \mathbb{E}\ell_{y_T}(r_{\mathrm{ref}}) - \mathbb{E}\ell_{y_T}(R_{I_m}^{\mathrm{mean}}(q_T^{\mathrm{R}}))$. The original-query terms cancel before applying the residual bound, so Lemma 2.2 is needed only at the regenerated query q_T^{R}, even when it depends on the original value residuals. The remaining direct-deletion and regeneration contributions are controlled by Proposition 4.3(i)–(ii), without imposing the original-write statistical model on regenerated writes.

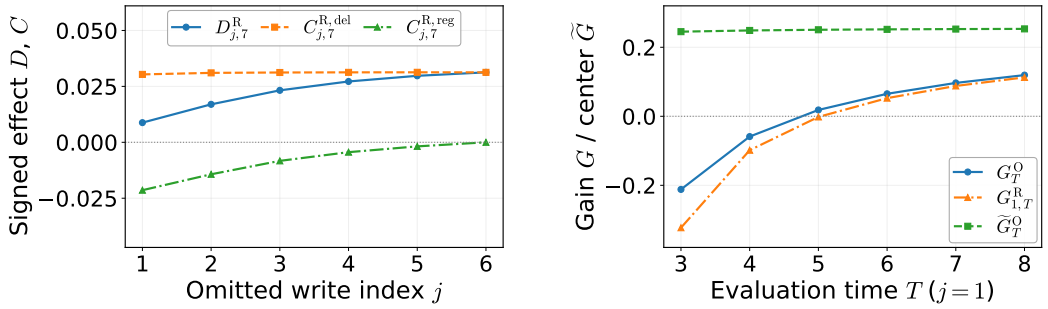
6 Numerical illustrations

TRIM decomposes signed single-write influence and bounds long-context effects and predictive gains, but does not determine system-specific finite-horizon magnitudes, signs, cancellation, or approximation accuracy. We therefore examine quantitative behavior rather than verify theorems.

We use a controlled Q–R–U–W model with a scalar latent variable and conditionally i.i.d. binary observations. Under squared-error prediction of the latent state, exact enumeration and quadrature yield signed contributions, predictive gains, and approximation errors (Appendix G.1).

We first vary the omitted write label to assess signed influence, then fix the label to examine predictive gains across evaluation times. At $T = 7$, direct-deletion contributions across $j = 1, \ldots, 6$ are nearly constant (0.03036–0.03129), while signed influence varies from 0.00879 to 0.03128; regeneration offsets about 71% of direct deletion at $j = 1$ (Figure 1(a); Appendix G.2). For $j = 1$, the first positive evaluated whole-memory gain occurs at $T = 5$, versus $T = 6$ for regenerated-memory gain; both are negative at $T = 3$ despite positive approximation centers (Figure 1(b); Appendix G.3). At $T = 8$, the whole-memory gain approximation error is 0.134, against the bound $L_* \delta_7 = 14.11$ (Appendix G.3). These are model-specific observations, not general cancellation or sign laws.

In selected pretrained Memory Mosaics trajectories, indirect readout discrepancies can exceed direct deletion, while regeneration and query-change vectors partly cancel (Appendix H). These pathwise observations are not expected-loss effects or evidence of typical behavior; the full architecture is not claimed to satisfy the model assumptions.



(a) Signed single-write influence.

(b) Predictive gains.

Figure 1: Finite-horizon numerical illustrations (exact enumeration and polynomial quadrature). (a) Signed single-write influence, direct-deletion contribution, and regeneration contribution at $T = 7$; the query-change contribution is in Appendix G.2. (b) Whole-memory and regenerated-memory gains at $j = 1$, with the approximation center $\widetilde{G}^{\mathrm{O}}_{T}$; the second approximation center $\widetilde{G}^{\mathrm{O}}_{T} - \widetilde{C}^{\mathrm{R,qry}}_{j,T}$ is specified in Theorem 5.2 and discussed in Appendix G.3.

7 Conclusions and Future Directions

The exact decomposition (2) separates signed single-write influence into direct-deletion, regeneration, and query-change contributions. Under the long-context assumptions of Theorem 4.4, direct deletion is $O(T^{-1})$, but regeneration can persist even as the selected-write weight vanishes (Example 4.5). Thus, vanishing weight need not eliminate downstream effects. For a fixed omitted label, Corollary 4.6 gives conditions for regeneration decay, not query-change decay. Whole-memory and regenerated-memory gains measure signed loss reductions against the same fixed reference predictor. Theorems 5.1 and 5.2 bound their conditional-mean-aggregate approximation errors. The latter uses the regenerated query and bounds residual, direct-deletion, and regeneration errors. Neither gain is guaranteed to be positive. These results hold under stated causal Q–R–U–W assumptions (Appendix F). Regenerated writes need not follow the original-write observation rule or a conditional-i.i.d. key–value law. Applying the theory to specific architectures requires separate verification (Appendix B.1).

Future work will study query-change decay and persistence and extend the theory to memory rewrites and other retrieval methods.

References

[1] Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and François-Xavier Vialard. Token sample complexity of attention. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 8899–8947. PMLR, 2026.

[2] Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 9392–9429. PMLR, 2026.

[3] Audrey Durand, Odalric-Ambrym Maillard, and Joelle Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *Journal of Machine Learning Research*, 19(17):1–34, 2018.

[4] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics, 2019.

[5] Haoxuan Jia, Yang Liu, Yingguang Yang, Yancheng Chen, Chongyang Zhang, Hao Zheng, Qian Li, Yulin Huang, Jianshen Zhang, Yongzhi Qi, Shang Luo, Kefu Xu, Hao Peng, Junyu Lu, Du Cheng, Philip S. Yu, and Bin Chong. Hindsight Memory-PRM: Supervising memory management with auditable hindsight credit. arXiv preprint arXiv:2608.29605, 2026.

[6] Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 3rd edition, 2021.

[7] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022.

[8] Vishwajith Ramesh. Exact record omission in delta attention: A transport criterion, its cost, and a replay certificate. arXiv preprint arXiv:2609.06872v2, 2026.

[9] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951. Association for Computational Linguistics, 2019.

[10] Jiaming Tang, Mingyan Liu, and Armin Sarabi. Learning what to remember: Long-horizon counterfactual memory optimization. arXiv preprint arXiv:2609.37930, 2026.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[12] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

[13] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[14] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representations (ICLR)*, 2015.

[15] Shuyao Xiao, Shengling Wang, Xuan Chen, Ke Chao, Ming Cui, Feifei Qian, Fanlin Meng, Chaoyang Mei, Chaoyong Jiang, Qi Ouyang, and Junxi Yi. When errors become memories: Causal pathway tracing in multi-turn memory-augmented LLMs. arXiv preprint arXiv:2608.30198, 2026.

[16] Jiangze Yan, Yi Shen, Wenjing Zhang, Jieyun Huang, Ping Chen, Zhaoxiang Liu, Ning Wang, Kai Wang, and Shiguo Lian. HiMPO: Hindsight-informed memory policy optimization for less-entangled credit in long-horizon agents. arXiv preprint arXiv:2606.16285, 2026.

[17] Sikuan Yan, Ahmed Bahloul, Ercong Nie, Susanna Schwarzmann, Riccardo Trivisonno, Volker Tresp, and Yunpu Ma. Memory-R2: Fair credit assignment for long-horizon memory-augmented LLM agents. arXiv preprint arXiv:2605.21768, 2026.

[18] Chao Yao, Yangbo Wei, Zhen Huang, Junhong Qian, Chenle Chen, Shaoqiang Lu, Chen Wu, and Lei He. Forgetting without restarting: Execution-state unlearning for stateful LLM agents. arXiv preprint arXiv:2609.04875, 2026.

[19] Jianyu Zhang and Léon Bottou. Memory mosaics at scale. In *Advances in Neural Information Processing Systems*, volume 38, pages 43556–43583, 2025.

[20] Jianyu Zhang, Niklas Nolte, Ranajoy Sadhukhan, Beidi Chen, and Léon Bottou. Memory mosaics. In *International Conference on Learning Representations (ICLR)*, 2025.

Supplementary Material:
A Theory of Retrieval with Incremental Memory: Causal Influence of Writes and Predictive Gains

A Related work and positioning (Section 1)

Exact record omission and downstream state changes. Ramesh [8] gives an if-and-only-if criterion for exact transport of an omission-boundary state difference in affine recurrent updates: the accumulated forcing from later updates must vanish. TRIM instead retains the input and state update, omits only the memory append, and bounds signed prediction-loss influence after later queries and key–value writes are regenerated. It makes no exact-state-recovery claim.

Counterfactual execution-state replay. Yao et al. [18] recover a target-omitted execution state by restoring a shared prefix and replaying sanitized observations, under deterministic transitions (or fixed randomness), matching counterfactual observations, and no committed post-target external side effects (with environment restoration ensuring reproducible observations). Their worst-case recomputation lower bound holds in a specified computational model. TRIM omits a post-update append rather than the target observation and studies predictive influence rather than execution-state unlearning.

Counterfactual pathways of error propagation. Xiao et al. [15] construct four counterfactual trajectories by varying whether an erroneous response enters memory updating and next-question feedback. TRIM instead fixes the external input sequence and omits one append. Its three-term decomposition is an exact identity for signed loss differences using specified intermediate readouts, not an evaluation of the pathway effects defined by Xiao et al.

Long-horizon marginal utility of memory updates. Tang et al. [10] aggregate utility differences between successive memory states on current and future targets and show that, under their formulation, the resulting return preserves the expected policy gradient of the factual objective. Their comparison does not define the continuation after omitting a single memory rewrite. TRIM instead regenerates subsequent internal computation and bounds the resulting signed loss differences.

Other counterfactual memory-credit methods. Memory-R2 [17] uses local rerollouts from a common intermediate memory state; HiMPO [16] compares target-answer log-likelihoods under the previous and updated memories; and Hindsight Memory-PRM [5] applies a controlled deletion-and-reanswer test to source-anchored probe questions. It also separates the presence counterfactual of a final memory entry from the action-level counterfactual, which would require replaying the continuation, and validates entry-level credit against replay on a sample of operations. These are prior methods for counterfactual memory credit. TRIM differs in deriving prediction-loss bounds for its specified write-omission intervention and in decomposing signed single-write influence exactly into direct-deletion, regeneration, and query-change contributions.

Statistical guarantees for softmax retrieval. Durand et al. [3] obtain confidence bounds uniform over evaluation points and times in streaming kernel regression. Boursier and Boyer [2] bound the finite-prompt concentration of softmax-attention outputs and gradients in L^2 over the query distribution, while Bohbot et al. [1] establish attention-map convergence uniformly over a bounded query set under their token assumptions. Thus, uniformity over queries is not itself new. Lemma 2.2 instead bounds the expected pathwise supremum of the original-write residual aggregate under (S1)–(S3) and (B1)–(B2), including residual-dependent queries, without assuming an analogous statistical law for regenerated writes.

B Additional details for the model and assumptions (Section 2)

This appendix gives architectural correspondences for the causal memory model and details the common conditional-mean–residual construction, original-residual orthogonality, and the uniform residual bound for adaptive queries. No conditional-i.i.d. law is imposed on regenerated memories.

B.1 Causal memory model: Architectural correspondences and applicability (Section 2.1)

Memory Mosaics. For Memory Mosaics (MM) [20, 19], M_{t-1} may represent inference-time context memory, while inference-fixed persistent memories may be incorporated into U_t. Applicability requires aligning the implementation with the Q–R–U–W chronology.

Autoregressive transformers. For a strictly causal single attention head [11], suppose positional score biases are, up to a read-time-common shift, of the form b_i. Then the dot-product scores $\tau^{-1}\langle q, k_i\rangle + b_i$ and the distance-based scores of Section 2.1 give identical softmax weights on a fixed prefix when $\beta_i = b_i + \|k_i\|^2/(2\tau)$. This fixed-prefix identity does not imply counterfactual equivalence: TRIM retains original biases even when later keys are regenerated. ALiBi [7] admits the stated bias representation, but its nonzero linear slope violates (B2) on unbounded write-time labels. Reading the current position or combining multiple heads requires separate treatment.

Neither correspondence alone establishes (S1)–(S4) or the required regenerated-write and query bounds.

B.2 Conditional-mean–residual decomposition (Section 2.3)

B.2.1 Bounded version of the conditional mean

Coordinatewise Doob–Dynkin factorization gives a jointly measurable map $g_0 : \Theta \times \mathbb{R}^d \to \mathbb{R}^p$ such that $g_0(\theta, k_1) = \mathbb{E}[v_1 \mid \sigma(\theta, k_1)]$ a.s.; see [6, Lemma 1.14]. By (S2), $\|v_1\| \le V$ a.s., so conditional Jensen's inequality yields $\|g_0(\theta, k_1)\| \le V$ a.s. Define $g := P_{\mathbb{B}_p(V)} \circ g_0$, where $P_{\mathbb{B}_p(V)}$ denotes metric projection onto the closed ball. The projection is continuous and fixes every point in the ball; hence g is jointly measurable, agrees with the conditional mean a.s., and satisfies $\|g(\vartheta, k)\| \le V$ for every $(\vartheta, k) \in \Theta \times \mathbb{R}^d$. Consequently, for every nonempty deterministic original label set I, $\|R_I^{\mathrm{mean}}(u)\| \le V$ for all u, pathwise. Since $\|v_i\| \le V$ a.s. for every original write-time label i by (S2), also $\|R_I(u)\| \le V$ a.s., simultaneously for all u. The common-version property across write-time labels and the stronger residual-centering statement are proved in Appendix B.2.2 and Lemma B.1, respectively.

B.2.2 Common conditional-mean version across original write-time labels

The same g works at every index. For bounded measurable $\chi_1 : \Theta \to \mathbb{R}$ and $\chi_2 : \mathcal{X} \to \mathbb{R}$, (S1) gives $\mathbb{E}[\chi_1(\theta)\chi_2(z_i) \mid \theta] = \chi_1(\theta)\mathbb{E}[\chi_2(z_1) \mid \theta] = \mathbb{E}[\chi_1(\theta)\chi_2(z_1) \mid \theta]$ a.s. By the monotone class theorem, the identity $\mathbb{E}[\chi(\theta, z_i) \mid \theta] = \mathbb{E}[\chi(\theta, z_1) \mid \theta]$ a.s. extends to every bounded measurable $\chi : \Theta \times \mathcal{X} \to \mathbb{R}$. Taking expectations shows that (θ, z_i) and (θ, z_1) have the same law. Since the common observation maps give $(k_i, v_i) = (\Phi_{\mathrm{K}}(z_i), \Phi_{\mathrm{V}}(z_i))$, it follows that (θ, k_i, v_i) and (θ, k_1, v_1) have the same law. For bounded measurable $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}[\psi(\theta, k_i)v_i] = \mathbb{E}[\psi(\theta, k_1)v_1] = \mathbb{E}[\psi(\theta, k_1)g(\theta, k_1)] = \mathbb{E}[\psi(\theta, k_i)g(\theta, k_i)].$$

To make the conditional-expectation step explicit, use $\epsilon_i = v_i - g(\theta, k_i)$ from Section 2.3. For every $B \in \sigma(\theta, k_i)$, the Doob–Dynkin factorization of $\mathbf{1}_B$ provides a bounded measurable test function ψ_B such that $\mathbf{1}_B = \psi_B(\theta, k_i)$ a.s. The preceding identity therefore gives $\mathbb{E}[\mathbf{1}_B\epsilon_i] = 0$ for every $B \in \sigma(\theta, k_i)$. Since ϵ_i is integrable, the defining test-set criterion for conditional expectation, applied coordinatewise in $\mathbb{R}^p$, yields $\mathbb{E}[\epsilon_i \mid \theta, k_i] = 0$. Thus $\mathbb{E}[\epsilon_i \mid \theta, k_i] = 0$ a.s. and $\|\epsilon_i\| \le 2V$ a.s., simultaneously for all $i \ge 1$. This is a decomposition, not an extra independence assumption.

B.2.3 Orthogonality of original residuals

The uniform residual bound (Lemma 2.2) needs only the following fixed-query second-moment bound. Write $X_\ell := (k_\ell, v_\ell)$ for original pairs and, for $i \in I_m$, $X_{-i} := (X_\ell)_{\ell \in I_m^{-i}}$ (empty if $m = 1$).

Lemma B.1 (Orthogonality of original residuals)**.** *Assume (S1)–(S2) of Section 2.2. For $m \geq 1$ and $i \in I_m$, put $\mathcal{F}_m^{-i} := \sigma(\theta, k_i, X_{-i}) = \sigma\big(\theta, k_{1:m}, (v_\ell)_{\ell \in I_m^{-i}}\big)$. Then $\mathbb{E}[\epsilon_i \mid \mathcal{F}_m^{-i}] = 0$ a.s. If, in addition, (S3) holds, then for every deterministic nonempty $I \subseteq I_m$ and deterministic $u \in \mathbb{R}^d$,*

$$\mathbb{E}\|R_I^{\mathrm{res}}(u)\|^2 = \sum_{i \in I} \mathbb{E}\big[(\pi_i^I(u))^2 \|\epsilon_i\|^2\big] \leq 4V^2\, \mathbb{E} \sum_{i \in I} (\pi_i^I(u))^2.$$

Proof of Lemma B.1. Fix $m \geq 1$ and $i \in I_m$. By (S1), z_i and $(z_\ell)_{\ell \in I_m^{-i}}$ are conditionally independent given θ. Applying the common measurable observation maps in (S2) shows that X_i and X_{-i} are conditionally independent given θ. Conditional independence given θ is preserved when θ is adjoined to both sides, so $\sigma(\theta, X_i)$ and $\sigma(\theta, X_{-i})$ are conditionally independent given θ. Since $\epsilon_i = v_i - g(\theta, k_i)$ is $\sigma(\theta, X_i)$-measurable, for bounded ψ measurable with respect to $\sigma(\theta, k_i)$ and bounded ψ' measurable with respect to $\sigma(\theta, X_{-i})$, this gives, coordinatewise,

$$\mathbb{E}[\psi\psi'\epsilon_i \mid \theta] = \mathbb{E}[\psi\epsilon_i \mid \theta]\mathbb{E}[\psi' \mid \theta] = 0. \tag{B.1}$$

The last equality follows from $\mathbb{E}[\psi\epsilon_i \mid \theta] = \mathbb{E}[\psi\mathbb{E}[\epsilon_i \mid \theta, k_i] \mid \theta] = 0$. Taking ψ and ψ' to be indicators, their products form a π-system generating $\mathcal{F}_m^{-i} = \sigma(\theta, k_i, X_{-i})$. The π–λ theorem and integrability of ϵ_i therefore give $\mathbb{E}[\epsilon_i \mid \mathcal{F}_m^{-i}] = 0$.

Now fix deterministic nonempty $I \subseteq I_m$ and $u \in \mathbb{R}^d$. By (S3) and $\mathcal{G}_j \subseteq \mathcal{G}_m$ for $j \leq m$, all original biases with labels in I are $\mathcal{G}_m$-measurable. The corresponding keys are $\mathcal{G}_m$-measurable by definition. For fixed u, the weights $\pi_j^I(u)$, $j \in I$, are therefore $\mathcal{G}_m$-measurable. Since $\mathcal{G}_m \subseteq \mathcal{F}_m^{-i}$, they are also $\mathcal{F}_m^{-i}$-measurable. For distinct $i, \ell \in I$, ϵ_ℓ is also $\mathcal{F}_m^{-i}$-measurable, so

$$\mathbb{E}[\pi_i^I(u)\pi_\ell^I(u)\langle\epsilon_i, \epsilon_\ell\rangle] = \mathbb{E}\left\langle \mathbb{E}[\epsilon_i \mid \mathcal{F}_m^{-i}], \pi_i^I(u)\pi_\ell^I(u)\epsilon_\ell \right\rangle = 0.$$

Expanding the squared norm of $R_I^{\mathrm{res}}(u) = \sum_{i \in I} \pi_i^I(u)\epsilon_i$ and using $\|\epsilon_i\| \leq 2V$ gives

$$\mathbb{E}\|R_I^{\mathrm{res}}(u)\|^2 = \sum_{i \in I} \mathbb{E}[(\pi_i^I(u))^2 \|\epsilon_i\|^2] \leq 4V^2\, \mathbb{E} \sum_{i \in I} (\pi_i^I(u))^2.$$

The argument uses deterministic u; residual-dependent queries are handled by Lemma 2.2. □

B.2.4 Proof of Lemma 2.2

Proof of Lemma 2.2. Fix deterministic nonempty $I \subseteq I_m$ and put $n := |I|$. For deterministic $u \in \mathbb{B}_d(K)$ and $i, \ell \in I$, (B1) gives $0 \leq \|u - k_i\|^2 \leq 4K^2$ and (B2) gives $|\beta_i - \beta_\ell| \leq \kappa_\beta$, a.s. Hence the exponents $-\|u - k_i\|^2/(2\tau) + \beta_i$ in the definition of $\pi_i^I(u)$ differ pairwise by at most $A = 2K^2/\tau + \kappa_\beta$. For each $i \in I$, the n terms in the denominator of $\pi_i^I(u)$ are at least $\exp(-A)$ times its numerator, so $\max_{i \in I} \pi_i^I(u) \leq \exp(A)/n$. By Lemma B.1 and $\sum_i \pi_i^I(u) = 1$,

$$\mathbb{E}\|R_I^{\mathrm{res}}(u)\|^2 \leq 4V^2 \mathbb{E} \sum_{i \in I} (\pi_i^I(u))^2 \leq 4V^2 \exp(A)/n.$$

Direct differentiation gives

$$\nabla_u \pi_i^I(u) = \frac{\pi_i^I(u)}{\tau}\left(k_i - \sum_{\ell \in I} \pi_\ell^I(u) k_\ell\right), \qquad \|\nabla_u \pi_i^I(u)\| \leq \frac{2K}{\tau}\pi_i^I(u).$$

Using $\|\epsilon_i\| \leq 2V$, we obtain $\|DR_I^{\mathrm{res}}(u)[w]\| \leq (4KV/\tau)\|w\|$. Since the query ball is convex, R_I^{res} is $(4KV/\tau)$-Lipschitz on it, pathwise.

With α as defined in Section 2.3, set $\lambda_n := Kn^{-\alpha}$. A deterministic maximal λ_n-separated set $\mathcal{T}_n \subseteq \mathbb{B}_d(K)$ is a λ_n-net [12, Lemma 4.2.6], with $|\mathcal{T}_n| \leq (1 + 2K/\lambda_n)^d \leq 3^d n^{d\alpha}$ by the Euclidean packing bound [12, Proposition 4.2.12 and Corollary 4.2.13]. Consequently, pathwise,

$$\sup_{u \in \mathbb{B}_d(K)} \|R_I^{\mathrm{res}}(u)\|^2 \leq 2 \max_{\bar{u} \in \mathcal{T}_n} \|R_I^{\mathrm{res}}(\bar{u})\|^2 + 2(4KV/\tau)^2 \lambda_n^2.$$

Taking expectations, using $\max \leq \sum$ on the deterministic net, and applying the fixed-query bound yields

$$\begin{aligned}
\mathbb{E}\left[\sup_{u \in \mathbb{B}_d(K)} \|R_I^{\mathrm{res}}(u)\|^2\right] &\leq \frac{8V^2 \exp(A)}{n}|\mathcal{T}_n| + 2(4KV/\tau)^2\lambda_n^2 \\
&\leq \left(8V^2 \exp(A) 3^d + \frac{32K^4V^2}{\tau^2}\right) n^{-2\alpha} = c_\epsilon n^{-2\alpha}.
\end{aligned}$$

The supremum is measurable because continuity reduces it to a countable supremum of measurable random variables over a deterministic dense subset of $\mathbb{B}_d(K)$. Fixed-u measurability and continuity also imply joint measurability, so $R_I^{\mathrm{res}}(U)$ is measurable for any measurable query U. For $U \in \mathbb{B}_d(K)$ a.s., its second moment is bounded by the preceding expected supremum, hence by $c_\epsilon n^{-2\alpha} = \delta_n^2$. Taking $I = I_m$ and $n = m$ proves the lemma. □

C Single-write causal comparisons (Section 3)

C.1 Pre-omission coupling for the regenerated run

Both runs start from the same state and empty memory and share the inputs and writer randomness. Induction over $s < j$ in the Q–R–U–W recursion shows that their post-update, pre-write histories coincide, hence

$$h^{\mathrm{R}}_{1:s+1} = h_{1:s+1}, \qquad M^{\mathrm{R}}_s = M_s \quad (s < j),$$

since the writer outputs agree and the original biases are retained. At $s = j$, the state update still agrees, but the append is omitted. Thus $h^{\mathrm{R}}_{j+1} = h_{j+1}$ and $M^{\mathrm{R}}_j = M_{j-1}$.

C.2 Frozen-write recursion and proof of exact reduction

The full frozen-run recursion, starting from the common initial state and memory of Section 3, is, for $s \geq 1$,

$$q^{\mathrm{F}}_s := \phi(h^{\mathrm{F}}_s), \quad r^{\mathrm{F}}_s := R_\tau(q^{\mathrm{F}}_s; M^{\mathrm{F}}_{s-1}), \quad h^{\mathrm{F}}_{s+1} := U_s(h^{\mathrm{F}}_s, z_s, r^{\mathrm{F}}_s), \quad M^{\mathrm{F}}_s = M_{I_s^{-j}}.$$

For $s \leq j$, the pre-read memory is M_{s-1} in both comparison runs. The same update induction as above gives $h^{\mathrm{F}}_s = h_s$ through $s = j + 1$; the first possibly altered read is at $s = j + 1$, and the first possibly altered updated state is h^{F}_{j+2}.

Proof of the exact reduction stated in Section 3. Assume $(k^{\mathrm{R}}_s, v^{\mathrm{R}}_s) = (k_s, v_s)$ a.s. for every surviving label s with $j < s \leq T - 1$. On a common probability-one event, the pre-omission coupling gives $h^{\mathrm{R}}_{j+1} = h^{\mathrm{F}}_{j+1} = h_{j+1}$ and $M^{\mathrm{R}}_j = M^{\mathrm{F}}_j = M_{j-1}$. Suppose at some $j < s \leq T - 1$ that $h^{\mathrm{R}}_s = h^{\mathrm{F}}_s$ and $M^{\mathrm{R}}_{s-1} = M^{\mathrm{F}}_{s-1}$. The common query map and the same pre-read memory yield $q^{\mathrm{R}}_s = q^{\mathrm{F}}_s$ and $r^{\mathrm{R}}_s = r^{\mathrm{F}}_s$. The common input z_s and update map then give $h^{\mathrm{R}}_{s+1} = h^{\mathrm{F}}_{s+1}$. By hypothesis the regenerated key and value equal the original ones, and both runs retain the original bias β_s, so their post-write memories also agree: $M^{\mathrm{R}}_s = M^{\mathrm{F}}_s$. This closes the induction through time T. In particular, the evaluation queries and memories coincide, so $r_T^{\mathrm{R}|\mathrm{R}} = r_T^{\mathrm{F}|\mathrm{F}}$ a.s.

D Signed single-write influence and long-context characterization (Section 4)

D.1 Finite-horizon contribution bounds and conditional-mean-aggregate approximation (Section 4.2)

D.1.1 Proof of Lemma 4.1

Proof of Lemma 4.1. We make the integrability and the two uses of Cauchy–Schwarz explicit. Since $\|x\|, \|x'\| \leq V$ a.s., for either $z \in \{x, x'\}$,

$$0 \leq \ell_{y_T}(z) = \|z - y_T\|^2 \leq (\|z\| + \|y_T\|)^2 \leq 2\|z\|^2 + 2\|y_T\|^2 \leq 2V^2 + 2\|y_T\|^2.$$

Thus both losses are integrable since $\mathbb{E}\|y_T\|^2 < \infty$. The polarization identity holds for every outcome:

$$\ell_{y_T}(x) - \ell_{y_T}(x') = \|x\|^2 - \|x'\|^2 - 2\langle x - x', y_T\rangle = \langle x - x', x + x' - 2y_T\rangle.$$

The ordinary inner-product Cauchy–Schwarz inequality and the triangle inequality first give the pointwise bound

$$|\ell_{y_T}(x) - \ell_{y_T}(x')| \leq \|x - x'\| \, (\|x\| + \|x'\| + 2\|y_T\|) \leq 2\|x - x'\|(V + \|y_T\|).$$

Now apply the scalar Cauchy–Schwarz inequality to $\|x - x'\| \, \|y_T\|$ in expectation:

$$\begin{aligned}|\mathbb{E}[\ell_{y_T}(x) - \ell_{y_T}(x')]| &\le \mathbb{E}|\ell_{y_T}(x) - \ell_{y_T}(x')| \le 2V\mathbb{E}\|x - x'\| + 2\mathbb{E}[\|x - x'\| \, \|y_T\|] \\ &\le 2V\sqrt{\mathbb{E}\|x - x'\|^2} + 2\sqrt{\mathbb{E}\|x - x'\|^2}\sqrt{\mathbb{E}\|y_T\|^2} \\ &= L_*\sqrt{\mathbb{E}\|x - x'\|^2}.\end{aligned}$$

The first moment bound uses $\mathbb{E}\|x - x'\| \le (\mathbb{E}\|x - x'\|^2)^{1/2}$. No independence of the predictions and the target is needed.

□

Under the conditional-i.i.d. input law (S1) and the square-integrable prediction target (S4), the required square-integrability of y_T follows from (1).

D.1.2 Proof of Lemma 4.2

Proof of Lemma 4.2. Fix $q \in \mathbb{R}^d$ and set $M := M_{I_{s-1}}$. We first derive the *softmax add-one identity* exactly, then bound its value difference under (B1), and finally bound the selected-write weight under (B1)–(B2). The identity itself uses neither bounded keys nor bounded values. Since $M = M_{I_{s-1}}$ is the full original prefix, its memory positions agree with original write-time labels. For each $i \in I_{s-1}$, the score $s_i(q; M)$ is defined in Section 2.1. The exponentiated score of the selected write and the sum of exponentiated scores over surviving labels are

$$a_j := \exp(s_j(q; M)) > 0, \qquad Z_- := \sum_{i \in I_{s-1}^{-j}} \exp(s_i(q; M)).$$

Step 1: softmax add-one identity. Suppose first that $Z_- > 0$. Since deletion only removes the label j, all surviving keys and biases (hence their scores) remain unchanged. The deleted-memory readout is therefore

$$R_{I_{s-1}^{-j}}(q) = \frac{\sum_{i \in I_{s-1}^{-j}} \exp(s_i(q; M))v_i}{Z_-}.$$

By definition, $p_{j,s}(q) = a_j/(a_j + Z_-)$. Substitute the last equality into the full readout, and then subtract the deleted-memory readout:

$$\begin{aligned}R_{I_{s-1}}(q) &= \frac{a_j v_j + \sum_{i \in I_{s-1}^{-j}} \exp(s_i(q; M))v_i}{a_j + Z_-} = \frac{a_j}{a_j + Z_-}v_j + \frac{Z_-}{a_j + Z_-}R_{I_{s-1}^{-j}}(q) \\ &= p_{j,s}(q)v_j + (1 - p_{j,s}(q))R_{I_{s-1}^{-j}}(q),\end{aligned}$$

which yields $R_{I_{s-1}}(q) - R_{I_{s-1}^{-j}}(q) = p_{j,s}(q)\big(v_j - R_{I_{s-1}^{-j}}(q)\big)$. If $Z_- = 0$, the original memory has exactly one entry ($s = 2, j = 1$). By the empty-memory convention, $R_{I_1}(q) = v_1$, $R_\varnothing(q) = 0$, and $p_{1,2}(q) = 1$; hence the same identity holds as $v_1 - 0 = 1 \cdot (v_1 - 0)$.

Step 2: value bound. For any nonempty memory N with $\|v_i\| \le V$ for every entry, its softmax weights are nonnegative and sum to one, so

$$\|R_\tau(q; N)\| = \left\|\sum_i \pi_i(q; N)v_i\right\| \le \sum_i \pi_i(q; N)\|v_i\| \le V\sum_i \pi_i(q; N) = V.$$

The bound also holds for the empty-memory readout, which is zero. Thus $\|v_j\| \le V$ and $\|R_{I_{s-1}^{-j}}(q)\| \le V$; the triangle inequality gives $\|v_j - R_{I_{s-1}^{-j}}(q)\| \le 2V$. Applying the identity from Step 1 yields, without any restriction on the query,

$$\|R_{I_{s-1}}(q) - R_{I_{s-1}^{-j}}(q)\| = p_{j,s}(q)\|v_j - R_{I_{s-1}^{-j}}(q)\| \le 2Vp_{j,s}(q).$$

Step 3: uniform softmax-weight bound. For the score-width and weight bound, now restrict $q \in \mathbb{B}_d(K)$. Work on the common probability-one event of (B1)–(B2). The distance range follows from $\|q - k_i\| \le \|q\| + \|k_i\| \le 2K$, so $0 \le \|q - k_i\|^2 \le 4K^2$; this applies to each of the $s - 1$ original labels, including the label about to be deleted. For any two labels $i, \ell \in I_{s-1}$ (a set that

includes the omitted write-time label j), the squared distances lie in $[0, 4K^2]$ and the original bias difference has absolute value at most κ_β. Thus

$$s_i(q;M) - s_\ell(q;M) = -\frac{\|q-k_i\|^2 - \|q-k_\ell\|^2}{2\tau} + (\beta_i - \beta_\ell),$$
$$|s_i(q;M) - s_\ell(q;M)| \le \frac{4K^2}{2\tau} + \kappa_\beta = A.$$

Put $n := s-1$. There are exactly n labels before deletion. The score-width bound gives $s_i(q;M) - s_j(q;M) \ge -A$ for every $i \in I_{s-1}$. Exponentiating and summing over these n labels yields

$$\sum_{i\in I_{s-1}} \exp(s_i(q;M) - s_j(q;M)) \ge \sum_{i\in I_{s-1}} \exp(-A) = n\exp(-A),$$

$$p_{j,s}(q) = \frac{\exp(s_j(q;M))}{\sum_{i\in I_{s-1}} \exp(s_i(q;M))} = \left(\sum_{i\in I_{s-1}} \exp(s_i(q;M) - s_j(q;M))\right)^{-1} \le \frac{1}{n\exp(-A)} = \frac{\exp(A)}{n}.$$

Since also $p_{j,s}(q) \le 1$, this proves $p_{j,s}(q) \le w_{s-1}$. Combining it with Step 2 gives $\|R_{I_{s-1}}(q) - R_{I_{s-1}^{-j}}(q)\| \le 2Vw_{s-1}$. All estimates use the same probability-one event and constants independent of q, so they hold simultaneously for every $q \in \mathbb{B}_d(K)$. The denominator counts the $s-1$ labels *before* deletion, not the $s-2$ surviving labels.

□

D.1.3 Conditional-mean comparison at two adaptive queries

Unlike the general squared-loss sensitivity bound (Lemma 4.1), this two-query comparison also uses the original-write uniform residual bound of Lemma 2.2.

Lemma D.1 (Conditional-mean comparison at two adaptive queries)**.** *Under the original-write statistical model (S1)–(S4) of Section 2.2 and conditions (B1)–(B2) of Section 2.1, for every evaluation time $T \ge 1$, every $m \ge 1$, and any two measurable, possibly residual-dependent queries U, U' that belong to $\mathbb{B}_d(K)$ a.s., one has*

$$\left|\mathbb{E}[\ell_{y_T}(R_{I_m}(U)) - \ell_{y_T}(R_{I_m}(U'))] - \mathbb{E}[\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U)) - \ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U'))]\right| \le 2L_*\delta_m.$$

Proof of Lemma D.1. At each query $R_{I_m} = R^{\mathrm{mean}}_{I_m} + R^{\mathrm{res}}_{I_m}$ pathwise. Both the readout and its conditional-mean aggregate have norm at most V (Appendix B.2.1). The uniform residual bound and squared-loss sensitivity give (Lemmas 2.2 and 4.1)

$$\left|\mathbb{E}\ell_{y_T}(R_{I_m}(U)) - \mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U))\right| \le L_*\delta_m.$$

Repeating the same inequality at U' is legitimate because the deterministic set I_m and the hypotheses of Lemma 2.2 are unchanged: its estimate holds for *every* measurable query taking values in the query ball, without requiring that the two queries have the same law. To see the cancellation and signs explicitly, put $e(U) := \mathbb{E}[\ell_{y_T}(R_{I_m}(U)) - \ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U))]$ and define $e(U')$ analogously. Both quantities are finite and each has absolute value at most $L_*\delta_m$. The quantity in the lemma is exactly $e(U) - e(U')$, and hence $|e(U) - e(U')| \le |e(U)| + |e(U')| \le L_*\delta_m + L_*\delta_m = 2L_*\delta_m$. □

D.1.4 Proof of Proposition 4.3

Proof of Proposition 4.3. Fix $T \ge 3$, put $m := T-1$, and fix the deterministic omitted write-time label $j \in I_m$. Write $r := r_T^{\mathrm{O|O}}$, $c := r_T^{\mathrm{R|O}}$, $a := r_T^{\mathrm{R|F}}$, and $b := r_T^{\mathrm{R|R}}$. For part (i), assume its stated hypotheses. At the common query q_T^{R}, the softmax add-one identity is $c - a = p_{j,T}(q_T^{\mathrm{R}})(v_j - a)$ (Lemma 4.2), where $p_{j,T}(q_T^{\mathrm{R}}) \le w_m$. The lemma applies at the (possibly adaptive) regenerated query because its bound holds simultaneously at all points of the query ball. The deleted memory is the frozen memory at evaluation time; no change to later surviving writes is involved in this

particular comparison. The original values and a have norm at most V; hence $\|c-a\| \le 2Vw_m$ a.s. Since w_m is deterministic, the same pathwise bound controls the squared mean and its square root:

$$(\mathbb{E}\|c-a\|^2)^{1/2} \le (\mathbb{E}(2Vw_m)^2)^{1/2} = 2Vw_m,$$
$$|C_{j,T}^{\mathrm{R,del}}| = |\mathbb{E}[\ell_{y_T}(a) - \ell_{y_T}(c)]| \le L_* \sqrt{\mathbb{E}\|c-a\|^2} \le 2VL_*w_m.$$

The loss inequality uses Lemma 4.1, applicable because $\|a\|, \|c\| \le V$ a.s. (both read original values); the sign of the loss difference does not affect its bound.

For part (ii), assume its stated hypotheses. At the regenerated query, $\|a\|, \|b\| \le V$ a.s., and thus $\|b-a\| \le 2V$. Consequently

$$\mathbb{E}\|b-a\|^2 \le 2V\mathbb{E}\|b-a\| = 2V\mu_{j,T}^{\mathrm{FR}}.$$

The final inference in (ii) is the following composition of the two separate estimates:

$$|C_{j,T}^{\mathrm{R,reg}}| = |\mathbb{E}[\ell_{y_T}(b) - \ell_{y_T}(a)]| \le L_* \sqrt{\mathbb{E}\|b-a\|^2} \le L_* \sqrt{2V\mu_{j,T}^{\mathrm{FR}}}.$$

Squared-loss sensitivity at (a, b) proves part (ii) (Lemma 4.1). No distributional condition on regenerated writes is used.

For part (iii), assume its stated hypotheses. Both q_T and q_T^{R} lie in $\mathbb{B}_d(K)$ a.s., and r, c read the same full original memory. At $(U, U') = (q_T^{\mathrm{R}}, q_T)$, the two-query conditional-mean estimate gives $|C_{j,T}^{\mathrm{R,qry}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \le 2L_*\delta_m$ (Lemma D.1), proving part (iii). The uniform residual estimate is an expectation of a supremum on the query ball, so this application does not require independence of the queries from the original residuals.

□

D.2 Long-context characterization (Section 4.3)

D.2.1 Proof of Theorem 4.4

Proof of Theorem 4.4. The finite-horizon argument compares three successive approximations of the same signed loss difference: first omit the selected write, then replace the original query-change readouts by their conditional-mean aggregates, and finally control changes in surviving writes. Each approximation has its own hypothesis and error term; the long-context equivalences follow by showing those errors vanish, not by separately forcing all contributions to vanish. Fix an arbitrary evaluation time $T \ge 3$ and put $m := T-1$. Fix the specified deterministic label $j \in I_m$. The four readouts $r_T^{\mathrm{O|O}}$, $r_T^{\mathrm{R|O}}$, $r_T^{\mathrm{R|F}}$, and $r_T^{\mathrm{R|R}}$ have norm at most V a.s.: the first three use original values, and the last uses regenerated values satisfying the imposed bound. Since $\mathbb{E}\|y_T\|^2 < \infty$, their loss expectations are finite, so (2) holds with finite terms.

Part (i). By (2), $D_{j,T}^{\mathrm{R}} - (C_{j,T}^{\mathrm{R,reg}} + C_{j,T}^{\mathrm{R,qry}}) = C_{j,T}^{\mathrm{R,del}}$, and Proposition 4.3(i) gives $|C_{j,T}^{\mathrm{R,del}}| \le 2VL_*w_m$. The finite-horizon bound requires no fixed-system convention.

Part (ii). The two queries q_T^{R} and q_T are measurable. The former belongs to $\mathbb{B}_d(K)$ a.s. by hypothesis, and the latter does so by (B3)–(B4). Both query-change readouts use the full original memory. Under the additional original-write statistical model (S1)–(S4), Proposition 4.3(iii) gives $|C_{j,T}^{\mathrm{R,qry}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \le 2L_*\delta_m$ even for residual-dependent queries. By the exact decomposition (2),

$$\begin{aligned}\left|D_{j,T}^{\mathrm{R}} - C_{j,T}^{\mathrm{R,reg}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}\right| &= \left|C_{j,T}^{\mathrm{R,del}} + (C_{j,T}^{\mathrm{R,qry}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}})\right| \le |C_{j,T}^{\mathrm{R,del}}| + |C_{j,T}^{\mathrm{R,qry}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \\ &\le 2VL_*w_m + 2L_*\delta_m,\end{aligned}$$

using the direct bound independently established in part (i). This proves Theorem 4.4(ii).

Part (iii). The value part of (B1) and the regenerated value bounds supply the hypothesis of Proposition 4.3(ii), which gives $|C_{j,T}^{\mathrm{R,reg}}| \le L_*\sqrt{2V\mu_{j,T}^{\mathrm{FR}}}$. Combining this with Theorem 4.4(ii) gives

$$|D_{j,T}^{\mathrm{R}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \le |D_{j,T}^{\mathrm{R}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}} - C_{j,T}^{\mathrm{R,reg}}| + |C_{j,T}^{\mathrm{R,reg}}| \le L_*(2Vw_m + 2\delta_m + \sqrt{2V\mu_{j,T}^{\mathrm{FR}}}),$$

which proves Theorem 4.4(iii).

In all three parts the loss comparisons are between bounded readouts, but only (ii) uses the statistical model to control an adaptive-query residual. Part (i) is a direct softmax deletion bound; (iii) additionally uses the frozen–regenerated deviation from part (ii) of the proposition. For the asymptotic assertions, work with one fixed system (Remark 2.1) and the uniform target second-moment bound across evaluation times. Then $w_m \le \exp(A)/m \to 0$ and $\sup_T L_* < \infty$, giving the $O(T^{-1})$ direct-deletion bound in (i). Under the additional original-write statistical model (S1)–(S4), $\delta_m = \sqrt{c_\epsilon}\, m^{-\alpha} \to 0$ with $\alpha = 1/(d+2) > 0$, proving vanishing of the right-hand side of part (ii). If also $\mu_{j,T}^{\mathrm{FR}} \to 0$, then the right-hand side of part (iii) vanishes as well.

For the equivalence chain, it is important to compare the *signed sums* rather than assert that their individual summands vanish. Define

$$S_{1,T} := C_{j,T}^{\mathrm{R,reg}} + C_{j,T}^{\mathrm{R,qry}}, \qquad S_{2,T} := C_{j,T}^{\mathrm{R,reg}} + \widetilde{C}_{j,T}^{\mathrm{R,qry}}, \qquad S_{3,T} := \widetilde{C}_{j,T}^{\mathrm{R,qry}}.$$

The proved bounds, the triangle inequality and $\mu_{j,T}^{\mathrm{FR}} \to 0$ give, successively,

$$\begin{aligned}
&|D_{j,T}^{\mathrm{R}} - S_{1,T}| \le 2VL_* w_m \to 0,\\
&|S_{1,T} - S_{2,T}| = |C_{j,T}^{\mathrm{R,qry}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \le 2L_*\delta_m \to 0,\\
&|S_{2,T} - S_{3,T}| = |C_{j,T}^{\mathrm{R,reg}}| \le L_*\sqrt{2V\mu_{j,T}^{\mathrm{FR}}} \to 0.
\end{aligned}$$

For any sequences of real numbers x_T, x'_T, $|x_T - x'_T| \to 0$ implies $x_T \to 0 \iff x'_T \to 0$: this follows from $|x_T| \le |x_T - x'_T| + |x'_T|$ and the reverse inequality with x_T, x'_T interchanged. Apply this elementary fact to the three displayed differences. The first equivalence holds without (S1)–(S4); the second uses them; and the third uses the additional deviation-decay condition. The bounds in Proposition 4.3(i) and (iii) have numerical right-hand sides independent of the deterministic j; the bound in (ii) may depend on j through $\mu_{j,T}^{\mathrm{FR}}$. All three apply to any deterministic $j = j(T)$ satisfying their respective hypotheses at every evaluation time. □

D.2.2 A persistent regeneration contribution

Example D.2 (Persistent regeneration contribution)**.** Fix $j \ge 2$. Consider one fixed system with $p = d = K = V = H = \tau = 1$, $\kappa_\beta = 0$, $\beta_s = 0$, $\Xi := \{0\}$, and $\zeta_s = 0$. Let $\Theta := \{-1, 1\}$, let θ be uniform on Θ, $\mathcal{X} := \{0\} \times \{-1, 1\}$, and $z_s := (0, \theta)$ for all s; $z^{(2)}$ denotes the second coordinate of $z \in \mathcal{X}$. Set $\Phi_{\mathrm{K}}(z) := 0$, $\Phi_{\mathrm{V}}(z) := z^{(2)}$, $f(\vartheta, z) := \vartheta$, $h_1 := 0$, $\phi \equiv 0$, and $U_s \equiv 0$.

For every history $\overline{\mathsf{hist}}_s = (\bar{h}_{1:s+1}, \bar{z}_{1:s}, \bar{M}, \bar{\zeta}_{1:s})$, define the measurable writer

$$W_s(\overline{\mathsf{hist}}_s) := \begin{cases} (0, \bar{z}_s^{(2)}, 0), & |\bar{M}| = s - 1,\\ (0, 0, 0), & |\bar{M}| \ne s - 1. \end{cases}$$

Original writes have $(k_s, v_s, \beta_s) = (0, \theta, 0)$. After omitting j, the regenerated values are $v_s^{\mathrm{R}} = \theta$ for $s < j$ and $v_s^{\mathrm{R}} = 0$ for $s > j$. All keys, biases, states, and queries are zero. Thus (B1)–(B4) and the regenerated key–value bounds hold. Conditions (S1)–(S4) hold because the inputs are conditionally deterministic, the original writes follow the observation maps, the biases are zero, and $y_T = \theta$ is bounded. Moreover, $g(\vartheta, k) = \vartheta$ is a bounded measurable original conditional mean.

All softmax weights are uniform. Hence, for every $T \ge j + 1$,

$$p_{j,T}(q_T^{\mathrm{R}}) = \frac{1}{T-1}, \qquad r_T^{\mathrm{R|F}} = \theta, \qquad r_T^{\mathrm{R|R}} = \frac{j-1}{T-2}\theta.$$

Since $y_T = \theta$ and $\theta^2 = 1$,

$$C_{j,T}^{\mathrm{R,reg}} = \mathbb{E}\Big[\ell_\theta(r_T^{\mathrm{R|R}}) - \ell_\theta(r_T^{\mathrm{R|F}})\Big] = \left(\frac{T-j-1}{T-2}\right)^2 \longrightarrow 1,$$

while $p_{j,T}(q_T^{\mathrm{R}}) \to 0$. Finally, $\|v_s^{\mathrm{R}} - v_s\| = 1$ for $s > j$, so $\mathcal{A}_{j,T} = \exp(A)(T-j-1)/(T-2) \to \exp(A) > 0$.

D.2.3 Proof of Corollary 4.6

Proof of Corollary 4.6. Put $n := T - 2$. By pre-omission coupling (Appendix C.1), the frozen and regenerated memories have the same surviving labels and biases, and their keys and values agree for $s < j$. Work on the common probability-one event of the stated bounds.

For $u \in \mathbb{B}_d(K)$, let σ^{F} and σ^{R} be their score vectors and put $\sigma(\lambda) := (1-\lambda)\sigma^{\mathrm{F}} + \lambda\sigma^{\mathrm{R}}$. For a score vector σ indexed by I_{T-1}^{-j}, write $\omega_s(\sigma) := \exp(\sigma_s)/\sum_{\ell \in I_{T-1}^{-j}} \exp(\sigma_\ell)$. The endpoint score widths are at most A by (B1)–(B2) and the key part of the regenerated key–value bounds, with the original biases retained. Since each score difference along the segment is a convex combination of its endpoint differences, the width of $\sigma(\lambda)$ is at most A; hence $\omega_s(\sigma(\lambda)) \le \exp(A)/n$. With $\rho(\sigma; v) := \sum_s \omega_s(\sigma) v_s$, softmax differentiation gives

$$\frac{d}{d\lambda}\rho(\sigma(\lambda); v) = \sum_s \omega_s(\sigma(\lambda))(v_s - \rho(\sigma(\lambda); v))(\sigma_s^{\mathrm{R}} - \sigma_s^{\mathrm{F}}).$$

Using $\|v_s - \rho(\sigma(\lambda); v)\| \le 2V$ and $|\sigma_s^{\mathrm{R}} - \sigma_s^{\mathrm{F}}| \le (2K/\tau)\|k_s^{\mathrm{R}} - k_s\|$, split the readout difference at a fixed query u as

$$R_\tau(u; M_{T-1}^{\mathrm{R}}) - R_\tau(u; M_{T-1}^{\mathrm{F}}) = \rho(\sigma^{\mathrm{R}}; v) - \rho(\sigma^{\mathrm{F}}; v) + \sum_{s \in I_{T-1}^{-j}} \omega_s(\sigma^{\mathrm{R}})(v_s^{\mathrm{R}} - v_s).$$

Integrating the preceding derivative bound controls the first term by $\frac{4KVe^A}{\tau n}\sum_{s=j+1}^{T-1}\|k_s^{\mathrm{R}} - k_s\|$. The second term is bounded by $\frac{e^A}{n}\sum_{s=j+1}^{T-1}\|v_s^{\mathrm{R}} - v_s\|$. Their sum is $\mathcal{A}_{j,T}$; both readouts also have norm at most V. Thus

$$\sup_{u \in \mathbb{B}_d(K)} \|R_\tau(u; M_{T-1}^{\mathrm{R}}) - R_\tau(u; M_{T-1}^{\mathrm{F}})\| \le Z_T := \min\{2V, \mathcal{A}_{j,T}\}. \tag{D.1}$$

The uniform estimate (D.1) applies to q_T^{R}, which lies in $\mathbb{B}_d(K)$ a.s., even when the query depends on residuals. Therefore $\mu_{j,T}^{\mathrm{FR}} \le \sqrt{\mathbb{E}Z_T^2}$, and Lemma 4.1 gives $|C_{j,T}^{\mathrm{R,reg}}| \le L_*\sqrt{\mathbb{E}Z_T^2}$. Under the additional hypotheses, Theorem 4.4(ii) and the triangle inequality give

$$|D_{j,T}^{\mathrm{R}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}| \le L_*\big(2Vw_{T-1} + 2\delta_{T-1} + \sqrt{\mathbb{E}Z_T^2}\big).$$

For every $\eta > 0$, $\mathbb{E}Z_T^2 \le \eta^2 + 4V^2\mathbb{P}(\mathcal{A}_{j,T} > \eta)$. Thus $\mathcal{A}_{j,T} \to 0$ in probability implies $\sqrt{\mathbb{E}Z_T^2} \to 0$. The uniform target second-moment bound gives $\sup_T L_* < \infty$; under the additional hypotheses, $w_{T-1}, \delta_{T-1} \to 0$ by Theorem 4.4. Hence the stated limits follow.

Finally, each summand defining $\mathcal{A}_{j,T}$ vanishes when both the key and the value are unchanged. For a changed label, the key–value bounds give $\|v_s^{\mathrm{R}} - v_s\| + (4KV/\tau)\|k_s^{\mathrm{R}} - k_s\| \le 2V + 8K^2V/\tau$. There are at most $N_{j,T}$ such labels. Hence

$$\mathcal{A}_{j,T} \le B\frac{N_{j,T}}{T-2}, \qquad B := \exp(A)(2V + 8K^2V/\tau).$$

This proves the sufficient condition involving $N_{j,T}/(T-2)$. If, almost surely, the sum defining $\mathcal{A}_{j,T}$ is bounded by a deterministic constant uniformly in T, then $\mathcal{A}_{j,T} = O(T^{-1})$ a.s. The bound $Z_T \le \mathcal{A}_{j,T} = O(T^{-1})$ holds almost surely with a deterministic constant. Hence $\sqrt{\mathbb{E}Z_T^2} = O(T^{-1})$. The bounds above for $\mu_{j,T}^{\mathrm{FR}}$ and $|C_{j,T}^{\mathrm{R,reg}}|$, together with $\sup_T L_* < \infty$, give both stated rates. A uniform deterministic bound on $N_{j,T}$ suffices for this condition. □

E Predictive gains of whole and regenerated memory (Section 5)

We prove the gain-approximation bounds for whole and regenerated memory, including the residual and frozen–regenerated deviation terms.

E.1 Proof of Theorem 5.1 (Section 5.1)

Proof of Theorem 5.1. By the definitions of the two gains, the common reference loss cancels. More explicitly, with the common reference predictor r_{ref} defined in Section 5.1,

$$\begin{aligned}&\big(\mathbb{E}\ell_{y_T}(r_{\mathrm{ref}}) - \mathbb{E}\ell_{y_T}(R_{I_m}(q_T))\big) - \big(\mathbb{E}\ell_{y_T}(r_{\mathrm{ref}}) - \mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(q_T))\big)\\ &\qquad = \mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(q_T)) - \mathbb{E}\ell_{y_T}(R_{I_m}(q_T)).\end{aligned}$$

The identity $R_{I_m}(q_T) = R^{\mathrm{mean}}_{I_m}(q_T) + R^{\mathrm{res}}_{I_m}(q_T)$ holds pathwise even at this adaptive query. Thus

$$|G^{\mathrm{O}}_T - \widetilde{G}^{\mathrm{O}}_T| = |\mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(q_T)) - \mathbb{E}\ell_{y_T}(R_{I_m}(q_T))| \le L_*\sqrt{\mathbb{E}\|R^{\mathrm{res}}_{I_m}(q_T)\|^2} \le L_*\delta_m,$$

where the first inequality is squared-loss sensitivity (Lemma 4.1) and the second is the uniform residual bound (Lemma 2.2); the two predictions are bounded by V (the original readout a.s. and the conditional-mean aggregate everywhere); the query $q_T \in \mathbb{B}_d(K)$ a.s. by original state-trajectory boundedness (B4) and the uniform query-map bound (B3) of Section 2.1. □

E.2 Proof of Theorem 5.2 (Section 5.2)

The conditional-mean aggregate remains based on original writes (Appendix B.2); regenerated readouts are controlled under the value and evaluation-query conditions of Proposition 4.3(i)–(ii).

Proof of Theorem 5.2. Put $U := q^{\mathrm{R}}_T$ and abbreviate $c := r^{\mathrm{R|O}}_T = R_{I_m}(U)$, $a := r^{\mathrm{R|F}}_T = R_{I_m^{-j}}(U)$, and $b := r^{\mathrm{R|R}}_T = R_\tau(U; M^{\mathrm{R}}_{T-1})$. Thus $R^{\mathrm{mean}}_{I_m}(U)$ evaluates the conditional-mean aggregate at the regenerated query. By the definitions of $\widetilde{G}^{\mathrm{O}}_T$ and $\widetilde{C}^{\mathrm{R,qry}}_{j,T}$,

$$\widetilde{G}^{\mathrm{O}}_T - \widetilde{C}^{\mathrm{R,qry}}_{j,T} = \mathbb{E}\ell_{y_T}(r_{\mathrm{ref}}) - \mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U)).$$

The definition of $G^{\mathrm{R}}_{j,T}$ therefore gives the exact identity

$$G^{\mathrm{R}}_{j,T} - (\widetilde{G}^{\mathrm{O}}_T - \widetilde{C}^{\mathrm{R,qry}}_{j,T}) = \mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U)) - \mathbb{E}\ell_{y_T}(b).$$

Inserting the losses at c and a yields

$$\begin{aligned}&\mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U)) - \mathbb{E}\ell_{y_T}(b)\\ &\qquad = \big(\mathbb{E}\ell_{y_T}(R^{\mathrm{mean}}_{I_m}(U)) - \mathbb{E}\ell_{y_T}(c)\big) + \big(\mathbb{E}\ell_{y_T}(c) - \mathbb{E}\ell_{y_T}(a)\big) + \big(\mathbb{E}\ell_{y_T}(a) - \mathbb{E}\ell_{y_T}(b)\big).\end{aligned}$$

By the definitions in Section 4.1, $C^{\mathrm{R,del}}_{j,T} = \mathbb{E}[\ell_{y_T}(a) - \ell_{y_T}(c)]$ and $C^{\mathrm{R,reg}}_{j,T} = \mathbb{E}[\ell_{y_T}(b) - \ell_{y_T}(a)]$. Thus the second and third differences above are $-C^{\mathrm{R,del}}_{j,T}$ and $-C^{\mathrm{R,reg}}_{j,T}$, respectively. By Proposition 4.3(i)–(ii), the second and third differences have absolute values at most $2VL_*w_m$ and $L_*\sqrt{2V\mu^{\mathrm{FR}}_{j,T}}$, respectively. For the first difference, $c - R^{\mathrm{mean}}_{I_m}(U) = R^{\mathrm{res}}_{I_m}(U)$, the label set I_m is deterministic, and $U \in \mathbb{B}_d(K)$ a.s. under the present hypotheses; both c and $R^{\mathrm{mean}}_{I_m}(U)$ have norm at most V. Lemmas 2.2 and 4.1 therefore bound its absolute value by $L_*\delta_m$, even if U depends on the original value residuals. The three bounds apply to the three bracketed terms in exactly the same order as they appear in the loss difference. Keeping the signs from the definitions, one gets

$$\Big|G^{\mathrm{R}}_{j,T} - (\widetilde{G}^{\mathrm{O}}_T - \widetilde{C}^{\mathrm{R,qry}}_{j,T})\Big| \le \underbrace{L_*\delta_m}_{\text{residual at } U} + \underbrace{2VL_*w_m}_{\text{direct deletion}} + \underbrace{L_*\sqrt{2V\mu^{\mathrm{FR}}_{j,T}}}_{\text{regeneration}}.$$

The triangle inequality gives the stated bound. □

F Hypotheses for the main results

Table F.1 summarizes the hypotheses for the principal results of Sections 4 and 5. Regenerated conditions apply only when explicitly stated; the original-write statistical model is not imposed on regenerated writes. For bounds involving expected losses, $\mathbb{E}\|y_T\|^2 < \infty$ is either assumed explicitly or follows from (S1) and (S4) by (1). Long-context conclusions without (S1)–(S4) retain their separately stated uniform target second-moment assumption. Theorem 4.4(i) also holds under the weaker conditions stated after that theorem.

Table F.1: Hypotheses for the listed results. Base lists the applicable base conditions (B1)–(B4); "Stat." denotes (S1)–(S4); "Reg." denotes regenerated key–value bounds ("values" means only regenerated value bounds); and "Eval." denotes $q_T^{\mathrm{R}} \in \mathbb{B}_d(K)$ a.s. A checkmark indicates that the corresponding condition is assumed.

Result	Base	Stat. (S)	Reg.	Eval.	Notes
Lem. 4.2	(B1)–(B2)				Identity: no bounds; readout-difference bound: (B1); weight bound: (B1)–(B2).
Prop. 4.3(i)	(B1)–(B2)			✓	Direct-deletion bound.
Prop. 4.3(ii)			values		Also assumes $\|v_s\| \le V$ a.s. for surviving labels; no key or bias-width bounds.
Prop. 4.3(iii)	(B1)–(B4)	✓		✓	Conditional-mean-aggregate approximation.
Thm. 4.4(i)	(B1)–(B4)		values	✓	Common theorem hypotheses listed; part (i) also holds under the weaker conditions stated after the theorem.
Thm. 4.4(ii)–(iii)	(B1)–(B4)	✓	values	✓	The additional condition $\mu_{j,T}^{\mathrm{FR}} \to 0$ is used for the long-context vanishing of the right-hand side of (iii) and for the final equivalence; it is not required for the finite-horizon inequalities.
Cor. 4.6 (regeneration bounds)	(B1)–(B2)		✓	✓	Fixed system and uniform target second-moment bound; no statistical model. Decay follows if $\mathcal{A}_{j,T} \to 0$ in probability.
Cor. 4.6 (additional bound)	(B1)–(B4)	✓	✓	✓	Additional bound for $D_{j,T}^{\mathrm{R}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}$; retains all hypotheses of the preceding corollary row.
Thm. 5.1	(B1)–(B4)	✓			Whole memory only.
Thm. 5.2	(B1)–(B2)	✓	values	✓	Requires hypotheses of Proposition 4.3 (i)–(ii).

Under a deterministic uniform bound on the sum defining $\mathcal{A}_{j,T}$, Corollary 4.6 also gives $O(T^{-1})$ rates for $\mu_{j,T}^{\mathrm{FR}}$ and $|C_{j,T}^{\mathrm{R,reg}}|$.

G Controlled Q–R–U–W experiment (Section 6)

This controlled experiment specifies the observation law and each Q–R–U–W map analytically, without learning a predictor. It evaluates model-specific finite-horizon magnitudes not fixed by the general theorems. Its purpose is neither to validate the exact decomposition (2) numerically nor to establish general monotonicity, cancellation, or sign claims. The signed single-write influence and predictive gains are those of Sections 4 and 5; no new performance measure is introduced here.

G.1 Model and computation

Let $\theta \sim \mathrm{Unif}[-1,1]$. Conditional on θ, the original observations $(X_t, Y_t) \in \{-1,+1\}^2$ are independent and identically distributed with

$$\mathbb{P}(X_t = x, Y_t = y \mid \theta) = \frac{1 + y\theta(1/2 + x\theta/4)}{4}.$$

Set the target $y_T := (\theta, 0)$ and the deterministic reference $r_{\mathrm{ref}} := (0,0)$; the loss is squared Euclidean error. The original keys and values are specified by $k_t := X_t$ and $v_t := (Y_t, 0)$, with zero score biases. The original conditional mean is $g_\theta(k) = (\theta(1/2 + k\theta/4), 0)$. We take $(\tau, \gamma, \rho, \lambda) = (2, 1/2, 1/2, 1)$, $K = V = 1$, $\kappa_\beta = 0$, $H = \sqrt{2}$, $d = 1$, and $p = 2$. The Q–R–U–W maps of Section 2.1, intervention, and measurements are specified as follows.

1. **Query (Q).** The state $h_t = (a_t, b_t)$ starts at $(0,0)$, and $\phi(a,b) := \gamma\,\mathrm{clip}_{[-1,1]}(b)$ gives the scalar query $q_t = \phi(h_t)$.
2. **Read (R).** The pre-update readout is $r_t = R_\tau(q_t; M_{t-1})$ with $\tau = 2$, original keys $k_i = X_i$, and zero score biases.
3. **Update (U).** The fixed update map uses the shared observation Y_t and the readout r_t:

$$a_{t+1} := \frac{(t-1)a_t + Y_t}{t}, \qquad b_{t+1} := \mathrm{clip}_{[-1,1]}\big((1-\rho)b_t + \rho(r_t)_1\big).$$

4. **Write (W).** After this update, the writer appends key X_t and a value $(\widehat{v}_t, 0)$, where

$$\widehat{v}_t := \mathrm{clip}_{[-1,1]}\left(\frac{n^2 Y_t + \lambda(n a_{t+1} - U_{t-1}^{\mathrm{mem}})}{n^2 + \lambda}\right),$$

   n is the number of surviving writes after the append, and U_{t-1}^{mem} is the sum of surviving first-coordinate memory values before it. This is the minimizer on $[-1,1]$ of $(v - Y_t)^2 + \lambda((U_{t-1}^{\mathrm{mem}} + v)/n - a_{t+1})^2$. The original run has $\widehat{v}_t = Y_t$; later keys remain the shared X_t, whereas compensatory values may change after omission.
5. **Intervention.** Write j is omitted while (X_t, Y_t) is shared. The same Q–R–U–W maps then operate on the surviving memory, with no cancellation sign imposed.
6. **Measurement.** The squared-loss evaluation includes signed single-write influence, its three contributions, and predictive gains. Define the *direct-only proxy error* by $\Delta_{j,T}^{\mathrm{dir}} := D_{j,T}^{\mathrm{R}} - C_{j,T}^{\mathrm{R,del}}$. By (2), $\Delta_{j,T}^{\mathrm{dir}} = C_{j,T}^{\mathrm{R,reg}} + C_{j,T}^{\mathrm{R,qry}}$.

For the primary finite-horizon calculations, we enumerate all 4^{T-1} observation histories and use $(T+2)$-node Gauss–Legendre quadrature in θ. The resulting polynomial integrals are exact up to floating-point arithmetic: with $m := T - 1$, the integrands have degree at most $2m + 4$, whereas $m + 3$ nodes integrate through degree $2m + 5$. The absolute-loss values $\mathbb{E}|\ell_{y_T}(r) - \ell_{y_T}(r')|$ of Appendix G.2, for pairs of readouts r, r', are the exception, because their integrands need not be polynomial in θ. All readouts in this model have zero second coordinate, so $\ell_{y_T}(r) - \ell_{y_T}(r') = ((r)_1 - (r')_1)((r)_1 + (r')_1 - 2\theta)$. For each history, the θ-integral is therefore split at $\theta = ((r)_1 + (r')_1)/2$, and the same quadrature is applied on each piece, where it is again exact. We evaluate $j = 1$ at $T = 3, \ldots, 8$, and all $j = 1, \ldots, 6$ at $T = 7$; whole-memory gain also includes $T = 2$. Long-horizon calculations at $T \in \{16, 32, 64, 128\}$ with $j = 1$ use 16 independent batches of 16,384 paths each (262,144 per time), with parent random seed 20261103. Error bars are pointwise two-sided 95% t intervals from the 16 batch means (15 degrees of freedom), not simultaneous intervals. Prefixes of the same paths are used across times, so estimates at different times are dependent.

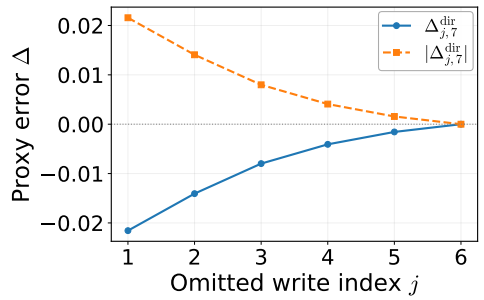


Figure G.1: Signed and absolute direct-only proxy errors at $T = 7$.

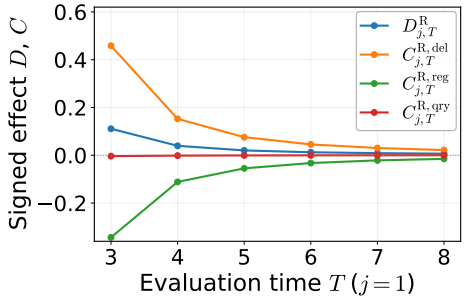


Figure G.2: Finite-horizon signed single-write influence and three contributions for $j = 1$.

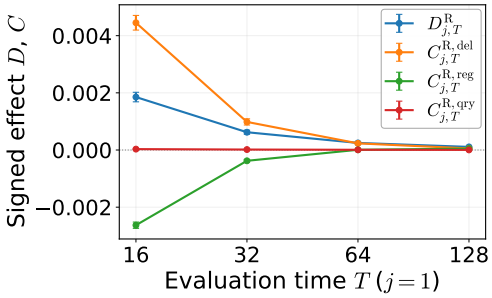


Figure G.3: Long-horizon signed single-write influence and three contributions for $j = 1$, with pointwise 95% confidence intervals.

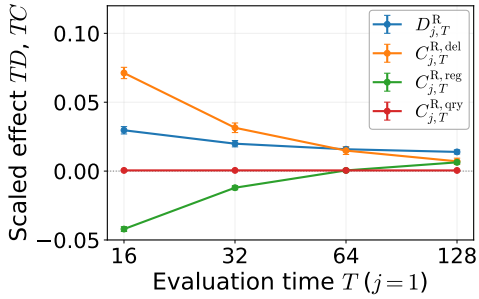


Figure G.4: The same long-horizon estimates and pointwise intervals multiplied by the evaluation time T.

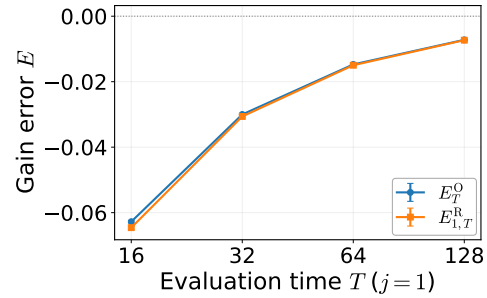


Figure G.5: Long-horizon predictive-gain approximation errors with pointwise 95% confidence intervals ($j = 1$).

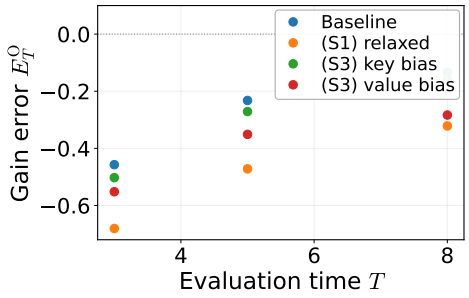


Figure G.6: Exploratory predictive-gain errors under alternative models; points mark evaluated horizons without interpolation.

G.2 Signed single-write influence and direct-only proxy error

At $T = 7$, the direct-deletion contribution is 0.03036–0.03129 as the omitted write-time label j varies, while signed single-write influence ranges from 0.00879 to 0.03128 (Figure 1(a)). For $j = 1$, regeneration contributes -0.02141, offsetting about 71% of the direct-deletion contribution; for $j = 6 = T-1$, there are no later writes to regenerate, and the pre-omission coupling gives $q^{\mathrm{R}}_{T} = q_T$; thus both regeneration and query change vanish. This dependence on j does not establish a general monotonicity law.

At $(T, j) = (8, 1)$, the direct-only proxy error is $\Delta^{\mathrm{dir}}_{j,T} = 0.006610 - 0.021751 = -0.015141$; Figure G.1 shows the errors as j varies. The time sweep and long-horizon estimates are in Figures G.2 to G.4. At $T = 128$, the regeneration estimate is $(4.961 \pm 0.842) \times 10^{-5}$ (pointwise 95% interval halfwidth), while its $T = 64$ interval includes zero. Neither observation determines an asymptotic sign or sign-change time.

At $(T, j) = (8, 1)$, the direct-deletion contribution has magnitude 0.021751, versus 0.100942, 0.464023, and 2.450104 at successive absolute-loss, readout-norm, and uniform-bound stages. For regeneration, the corresponding values are 0.015087, 0.048143, and 0.243744; two other bounds are 0.664006 and 1.097117, with no general ordering between them. Thus measured contributions are much smaller than the general bounds. The regenerated keys equal the original keys. The model is one fixed system (Remark 2.1) evaluated at every T. With $K = V = 1$, zero biases, and values clipped to $[-1, 1]$, it satisfies (B1)–(B2) with $\kappa_\beta = 0$ and the regenerated key–value bounds; also $|q_T^{\mathrm{R}}| \le \gamma \le K$, and $|\theta| \le 1$ gives the uniform target second-moment bound. Because clipping is nonexpansive, the writer rule gives, for $s > j$,

$$\|v_s^{\mathrm{R}} - v_s\| \le \frac{\lambda}{(s-1)^2 + \lambda} \left(2 + \sum_{r=j+1}^{s-1} \|v_r^{\mathrm{R}} - v_r\| \right).$$

Adding $2 + \sum_{r=j+1}^{s-1} \|v_r^{\mathrm{R}} - v_r\|$ to both sides of the recurrence gives

$$2 + \sum_{r=j+1}^{s} \|v_r^{\mathrm{R}} - v_r\| \le \left(1 + \frac{\lambda}{(s-1)^2 + \lambda} \right) \left(2 + \sum_{r=j+1}^{s-1} \|v_r^{\mathrm{R}} - v_r\| \right).$$

Iterating from $s = j + 1$, where the sum on the right-hand side is empty, gives

$$\begin{aligned} \sum_{s=j+1}^{T-1} \|v_s^{\mathrm{R}} - v_s\| &\le 2 \left\{ \prod_{s=j+1}^{T-1} \left(1 + \frac{\lambda}{(s-1)^2 + \lambda} \right) - 1 \right\} \\ &\le 2 \left\{ \exp\left(\sum_{s=j+1}^{\infty} \frac{\lambda}{(s-1)^2 + \lambda} \right) - 1 \right\} < \infty. \end{aligned}$$

The summability of the coefficients makes this deterministic bound uniform in T. Thus the sum defining $\mathcal{A}_{j,T}$ is also uniformly bounded, almost surely. Corollary 4.6 then yields $\mu_{j,T}^{\mathrm{FR}} = O(T^{-1})$ and $|C_{j,T}^{\mathrm{R,reg}}| = O(T^{-1})$ for fixed j. These rates do not automatically extend to the query-change contribution or signed single-write influence.

G.3 Predictive gains and approximation errors

Both gains use $r_{\mathrm{ref}} = (0, 0)$ and their respective realized queries. The centers $\widetilde{G}_T^{\mathrm{O}}$ and $\widetilde{G}_T^{\mathrm{O}} - \widetilde{C}_{j,T}^{\mathrm{R,qry}}$ are comparison quantities, not observation-only predictors. Figure 1(b) shows that for $j = 1$ the whole-memory gain first becomes positive at evaluated $T = 5$ and the regenerated-memory gain at $T = 6$. At $T = 3$, both centers are positive, but both gains are negative: a positive approximation center alone does not establish positive gain without a sufficiently small approximation-error bound.

At $T = 8$, put $E_8^{\mathrm{O}} := G_8^{\mathrm{O}} - \widetilde{G}_8^{\mathrm{O}}$. Its magnitude and bounds are $|E_8^{\mathrm{O}}| = 0.133907 \le L_* \sqrt{\mathbb{E}\|R_{I_7}^{\mathrm{res}}(q_8)\|^2} = 1.132368 \le L_* \delta_7 = 14.113323$. At every evaluated horizon, the error bound $L_* \delta_m$ exceeds the model's value-range bound 4; at $T = 128$, it is about 5.3711 versus a measured error magnitude of 0.007217. Failure to certify a gain does not imply zero gain. The readout decomposition and the gain definitions give

$$\begin{aligned} G_T^{\mathrm{O}} - \widetilde{G}_T^{\mathrm{O}} = &- \mathbb{E}\|R_{I_m}^{\mathrm{res}}(q_T)\|^2 \\ &- 2\mathbb{E}\langle R_{I_m}^{\mathrm{mean}}(q_T) - y_T, R_{I_m}^{\mathrm{res}}(q_T) \rangle. \end{aligned}$$

At $T = 8$, the signed error decomposes as $E_8^{\mathrm{O}} = -0.128842 - 0.005065 = -0.133907$; the residual-square term accounts for 96.2% of its magnitude. The cross term vanishes in the fixed-query $q = 0$ control, not necessarily for adaptive queries. Figure G.5 reports long-horizon errors; finite-horizon trends alone do not establish asymptotic rates.

G.4 Exploratory assumption sensitivity and limits

These post-hoc diagnostics do not establish the necessity of any assumption. To vary (S1), reuse the preceding observation with probability χ instead of drawing afresh conditional on θ. To examine

(S3), set the score bias to $\beta_i := X_i/2$ (key-dependent control) or $\beta_i := Y_i/2$ (value-dependent comparison), leaving the writer fixed; the baseline has $\chi = 0$ and $\beta_i = 0$. At $(T, j) = (8, 1)$, the baseline has $(D^{\mathrm{R}}_{1,8}, G^{\mathrm{O}}_{8}, G^{\mathrm{R}}_{1,8}) = (0.006610, 0.119374, 0.112764)$. Reusing observations with $\chi = 1/2$ makes both gains negative ($-0.072136, -0.088583$), while value-dependent bias $\beta_i = Y_i/2$ makes signed single-write influence and both gains negative ($-0.018471, -0.031682, -0.013212$). Key-dependent bias preserves their positive signs. These are comparisons between fully specified models, not isolated effects of individual assumptions.

At $T = 8$, the whole-memory gain approximation error changes from -0.133907 (baseline) to -0.321544 (temporal dependence) and -0.283322 (value-dependent bias); see Figure G.6. The adaptive-query residual mean is approximately zero at baseline, 0.010273 under key-dependent bias, and 0.366798 under value-dependent bias. Thus (S3) alone does not imply a zero adaptive-query residual mean; baseline sign symmetry also matters. Bounds requiring (S1)–(S4) are not claimed for relaxed models.

The bounded, compensatory model yields no general cancellation law, sign-change time, or asymptotic rate from the finite-horizon data. Long-horizon intervals are pointwise. Some post-hoc diagnostics ($T = 7$, $j = 2, \ldots, 6$) have only one independent implementation check. These limitations do not alter the proved theorems.

H Memory Mosaics: single-write retrieval exclusion and readout changes

H.1 Model and computation

Model and Q–R–U–W correspondences. In contrast to the controlled model's specified maps (Appendix G.1), we use a pretrained Vanilla Memory Mosaics (MM) language model [20] with fixed weights. We follow the same six items, interpreting Q, R, and W through a selected contextual-memory head without asserting that the full architecture realizes the abstract model.

1. **Query (Q).** In place of the specified scalar map ϕ, the learned key-generation mechanism maps the current token history to the query key of layer 0, head 0 (dimension 64).
2. **Read (R).** The selected head uses inner-product softmax to average earlier contextual-memory values instead of the distance-based readout. For the normalized keys, its softmax weights can also be evaluated by squared-distance retrieval, up to a score offset independent of the memory index; the numerical correspondence was checked.
3. **Update (U).** Instead of an explicit state recursion, the full MM network (including other heads and layers) forms next-token scores; a shared position-wise Gumbel vector selects the next token. This computation is not identified with a single abstract U_t.
4. **Write (W).** The learned key–value mechanisms form the target head's contextual-memory entry using the newly generated token for its value. If the generated token histories diverge, subsequent keys and values can both change, unlike the controlled experiment's shared later keys and compensatory value updates.
5. **Intervention.** At the selected write j, the target head's entry is excluded from subsequent retrievals by setting its search score to $-\infty$ and renormalizing the softmax. This is *retrieval-time exclusion*, rather than omission of the writer's append operation as in the controlled experiment. Other heads, persistent memories, and learned parameters are not directly modified. Original (O) and regenerated (R) runs share the initial context and position-wise Gumbel random vectors but may generate different subsequent tokens; later queries, keys, and values are computed from each run's own token history.
6. **Measurement.** We compare mixed readouts of the selected head before output projection (dimension 64), not signed expected-loss differences, predictive gains, or the final next-token prediction.

For each evaluation time $T > j$, put $L := T - j$. We use 16 distinct 12-token initial contexts, separately excluding $j \in \{13, 29, 45\}$ (one-based TRIM indices) in each context, and generate through position 65. The first token-divergence time is not imposed.

Mixed readouts and measured quantities. At time $T > j$, write M^{O}_{T-1} for the original target-head memory, M^{F}_{T-1} for its restriction with label j removed, and M^{R}_{T-1} for the surviving memory formed along the regenerated run. We construct the four readouts of Section 4.1: $r^{\mathrm{O|O}}_{T}$ (original query,

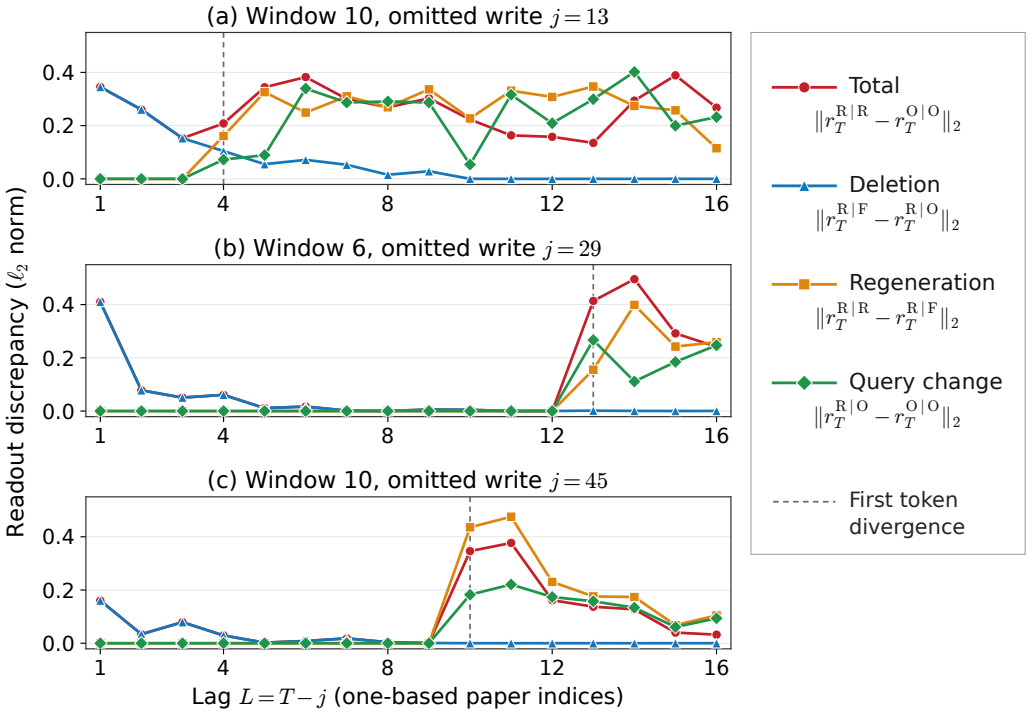


Figure H.7: Four readout discrepancies along three regenerated trajectories after excluding one contextual-memory entry from retrieval. The horizontal axis is $L = T - j$; vertical dashed lines indicate the first token divergence ($L = 4, 13, 10$ in panels (a)–(c)). Full readout-norm expressions appear in the legend. All panels display pathwise quantities rather than signed expected-loss contributions.

original memory), $r_T^{\mathrm{R}|\mathrm{O}}$ (regenerated query, original memory), $r_T^{\mathrm{R}|\mathrm{F}}$ (regenerated query, original memory without write j), and $r_T^{\mathrm{R}|\mathrm{R}}$ (regenerated query and memory). Only two generation runs are required; the mixed readouts are computed from their stored queries and keys–values. Figure H.7 plots the four Euclidean readout discrepancies, with their complete formulas given in the legend. These readout differences telescope as vectors, but their Euclidean norms do not in general add. These pathwise norms are *not* the signed expected-loss contributions $C_{j,T}^{\mathrm{R,del}}$, $C_{j,T}^{\mathrm{R,reg}}$, $C_{j,T}^{\mathrm{R,qry}}$, or the influence $D_{j,T}^{\mathrm{R}}$.

H.2 Readout changes and interpretation

Observed readout changes. The original design generated 16 original and 48 regenerated trajectories and evaluated 192 (j, L) pairs for $L \in \{2, 4, 8, 16\}$. Five regenerated trajectories diverged within the generated horizon; three had diverged by the displayed $L \leq 16$ window. We reconstructed all 16 time points per displayed trajectory from saved causal-prefix keys, values, and queries, using double-precision retrieval without further MM generation. At the prespecified points, the maximum discrepancy from the original recorded readout norms was 3.16×10^{-6}. Before token divergence, the remaining target-head keys, values, and query coincide across runs up to numerical error, so only the direct-deletion discrepancy is present. After divergence, the indirect discrepancies can dominate. In panel (a), at $L = 16$, the direct-deletion norm is 5.98×10^{-5} but the total is 0.2673. Panel (b) reaches a transient total discrepancy of 0.4954 at $L = 14$. In panel (c), regeneration and query-change difference vectors largely oppose each other; at $L = 16$, their norms are 0.1043 and 0.0937 while the total is 0.0321.

Interpretation and scope. The three paths were selected after observing divergence and do not estimate a typical effect over all runs. The experiment isolates readout-level propagation through

subsequent generated tokens, memories, and queries; it neither measures signed expected-loss contributions nor establishes that the full MM architecture satisfies the assumptions of the TRIM causal memory model. A near-tied next-token selection in one case also limits claims of numerical robustness of that branch.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the scope of the causal write-omission analysis, the exact direct-deletion, regeneration, and query-change decomposition, the conditional long-context conclusions, and predictive-gain approximation bounds. Section 6 presents model-specific numerical illustrations and distinguishes selected Memory Mosaics readout changes from signed expected-loss effects.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 and Appendix F delimit the causal Q–R–U–W model and result-specific assumptions; applicability to trained architectures is not asserted without separate verification. Appendices G.4 and H explain the limits of the numerical comparisons and the selected Memory Mosaics trajectories.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Sections 2–5 state or reference the assumptions for the principal results, and Appendix F summarizes them result by result. The corresponding formal proofs and auxiliary lemmas appear in Appendices B–E.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Appendix G specifies the controlled model and its calculations, and Appendix H identifies the Memory Mosaics execution settings and the checkpoint fingerprint. However, third-party access to the exact checkpoint and matching validation-data instance has not been established, so there is no verified path for others to regenerate the Memory Mosaics trajectories. Checks of stored numerical results do not constitute an independent rerun of model generation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide an openly accessible reproduction code package or the complete experimental inputs and outputs. Appendices G and H report the numerical methods and observations, but these descriptions are not equivalent to an open release of code and data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix G.1 specifies the controlled stochastic model, numerical integration, observation histories, and long-horizon sampling. Appendix H identifies the fixed-weight Memory Mosaics checkpoint by its checksum and records the model configuration, source revision, validation data fingerprint, context-selection procedure, intervention, random-number generation, evaluation times, and numerical precision. No model training was conducted for this study; the checkpoint's prior training provenance was not independently established.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The finite-horizon controlled calculations use exact enumeration and polynomial quadrature; for long horizons, Appendix G.1 specifies 16 independent batches and pointwise two-sided 95% t intervals from batch means. Appendix H treats three post-selected Memory Mosaics trajectories as descriptive cases, not population-level statistical estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Both reported experiments were small-scale, and each took less than one hour in the author's runs. Appendices G and H describe the numerical procedures and experiment scale, but the paper does not fully report hardware types, memory requirements, or total compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research develops mathematical results and reports controlled synthetic calculations and offline examinations of a pretrained language model (Section 6). It does not conduct human-subject research or deploy an experimental intervention on users.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: This is a theoretical study with controlled and offline numerical illustrations, not a deployed system. Section 7 discusses the scope and limitations, but the paper does not explicitly analyze both potential positive and negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new pretrained model, dataset, or deployed system; the reported experiments are controlled calculations and offline evaluations (Section 6). Thus safeguards for releasing high-risk assets are not applicable to this work.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper credits the Memory Mosaics work and identifies the implementation revision and the checkpoint and data fingerprints in Appendix H. However, the licenses and applicable terms for all used code, model, and data assets are not explicitly documented.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, benchmark, pretrained model, or reusable software package as a research asset; its numerical results are presented as illustrations rather than a standalone asset release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work includes no crowdsourcing, user studies, or interaction with human research participants. Its experiments comprise controlled synthetic computations and offline analysis of a pretrained model.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not conduct research with human subjects or crowdsourcing, so participant-risk disclosures and institutional review board approval are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The research questions, the TRIM framework, and the central ideas of the main theorems were conceived by the author. Generative AI tools were used in developing and refining the mathematical formulations, claims, and proof arguments, including proposing candidate arguments and proof steps. They also assisted with literature review, manuscript preparation, figures, and supporting code. The author independently rederived and verified the final mathematical claims and proofs, checked the cited sources and bibliographic information, reviewed the figures and supporting code for consistency with the reported results, and takes full responsibility for the correctness and originality of the paper. Generative AI is not a component of the TRIM framework itself and was not used as a computational component in the reported experiments; its use is declared because it contributed to the development of the core theoretical results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.