TEST-TIME MEMORY FROM THE ONLINE LEARNING PERSPECTIVE: ASSOCIATIVE PREDICTION GAME AND REGRET OPTIMALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time training (TTT) layers and Delta-style memories learn key-value (KV) associations as a sequence arrives. We study which update rules learn these association well and when the resulting memory can answer queries that differ from its update keys. We measure learning through regret, which is the extra prediction loss relative to a fixed comparison map. For exact recursive least squares (RLS), we connect prediction regret and retrieval error through the same regularized key covariance. Specifically, ordinary RLS has logarithmic regret when key norms and prediction residuals remain uniformly bounded. A classical Vovk–Azoury–Warmuth (VAW) readout achieves the optimal worst-case rate in the stated bounded prediction game without the residual condition. A separate construction shows that constant-step Delta can incur much larger regret. Our retrieval guarantees identify additional conditions on query coverage and how values and noise are generated. Furthermore, we analyze approximate updates and memories that forget old observations. Extensive experiments on WikiText streams show that RLS performs better in the evaluated stable settings, while resetting and discounting old information helps against controlled drifts. Meanwhile, a LongBench-v2 setting shows that selected Delta reconstructs native attention slightly better, while RLS performs better when attention is restricted to the stored context. These findings distinguish guarantees for learning KV associations from performance on query-time reconstruction.

1 INTRODUCTION

Modern sequence models, especially recurrent alternatives to attention models, increasingly rely on fast-weight memories that are updated online through KV interactions. In a test-time training (TTT) layer, the hidden state is a model updated at test time by a self-supervised learning rule (Sun et al., 2020; Liu et al., 2021). In Delta-rule memories (Yang et al., 2024; 2025) and related linear recurrent layers (Katharopoulos et al., 2020), the state is a fast associative map that binds keys to values. Thus, these mechanisms can be viewed as online learners, in that they receive sequential data, then update an internal state and make future predictions using it. However, this viewpoint raises an important question that is not answered by the architecture alone: *which update rule is the most appropriate for the prediction problem being solved ?*

We investigate this question through a simple online associative prediction game that captures the KV memory behavior of TTT and Delta-style layers. The goal is to understand how different recurrent updates behave when they are used to learn an associative map from streaming KV pairs. We also examine when good performance on the KV training objective leads to good retrieval for later queries, which may be different from the keys used during updating. This distinction is important. as low KV prediction error does not, by itself, imply accurate query-time retrieval. Therefore, our analysis separates two things: how well the memory learns from the past KV pairs, and how well the learned memory answers later queries.

The squared KV loss makes this distinction concrete. Delta-style updates (Schlag et al., 2021) ans single-step TTT-Linear-SGD (Liu et al., 2026) are first-order gradient methods. Instead, exact RLS uses the covariance of past keys and admits a data-dependent stability bound, which is turned

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

into logarithmic regret by controlled RLS residuals. Applying the classical Vovk-Azoury-Warmuth forecaster (Vovk, 2001; Azoury and Warmuth, 2001) as a key-dependent output gate helps to achieve the bounded-game minimax rate without the residual assumption. We also look at unit-norm keys, on which every constant scalar Delta step has expected regret of order $\sqrt{T}$, versus $O(\log T)$ of VAW. This serves as a limitation of one fixed learning rate, rather than all first-order algorithms.

The same ridge state connects prediction to retrieval, determining the weights of the causal sequence operator and measuring whether a query follows directions supported by observed keys. We turn this connection into retrieval bounds in two cases: noiseless observations and noisy models. A covariance condition makes query coverage checkable, while an indistinguishability argument identifies an unavoidable error in unseen directions. Together, these results explain what is needed beyond a small loss of KV learning. Approximation and forgetting raises different problems. A curvature-aware approximation, for example online Newton (Hazan et al., 2007) and fully preconditioned TTT (Wang et al., 2025), inherits an RLS guarantee only when its prediction error is controlled in the relevant loss geometry. Meanwhile, forgetting changes the comparator problem. Scalar decay can bias a memory away from a fixed target, whereas resets or discounting can reduce the influence of obsolete associations. We retain these distinctions, rather than trying to assign every gated or second-order architecture the same guarantee.

As seen, our major contribution is theory that connects different memory updates to the online learning problem they implicitly solve. Finally, our diagnostic experiments aim to test these mechanisms, rather than rank complete sequence models. On fixed-feature WikiText streams (Merity et al., 2016), exact RLS achieves lower static regret than Delta across all evaluated paired streams. On controlled drifting maps, reset and discounted memories improve dynamic regret in the reported comparisons, while scalar gating can hurt when it removes useful information. Meanwhile, on LongBench-v2 examples (Bai et al., 2023), selected Delta outperforms RLS on native-attention reconstruction, while RLS is slightly better when attention is limited to the stored context.

2 Related Work

Test-time training. TTT updates model parameters during inference, allowing a model to adapt after its initial training. Early works used this approach to address distribution shifts, where test inputs differ from training data (Sun et al., 2020; Liu et al., 2021; Gandelsman et al., 2022). TTT has been used in computer vision (Han et al., 2025), including 3D reconstruction (Chen et al., 2024; Yuan et al., 2026). Recently, Sun et al. (2024) introduced TTT layers whose hidden state is itself a model, updated through self-supervised learning as inputs arive.

Associative regression and linear attention. Wang et al. (2025) developed a test-time regression framework connecting several sequence models through online associative regression: learning mappings from keys to values as new pairs arrive. This framework includes linear attention (Katharopoulos et al., 2020) and variants that transform inputs through feature maps (Peng et al., 2021; Qin et al., 2022; Kasai et al., 2021; Aksenov et al., 2024). It also connect state-space models (Gu and Dao, 2023), DeltaNet (Schlag et al., 2021), gated DeltaNet (Yang et al., 2025), Longhorn (Liu et al., 2025), alongside TTT. Liu et al. (2026) aruged that a broad class of KV binding mechanisms can be expressed as learned linear attention during inference. We adopt the corresponding regression perspective: the memory is a model in a feature space, and its output forms a causal operator over earlier values. Within this representation, we study which guarantees follow from particular updates and when KV learning transfers to retrieval.

Delta-rule memories and forgetting. Delta-rule memories and fast-weight programmers (Schlag et al., 2021; Yang et al., 2025) have a long history in sequence modeling, and recent DeltaNet-style architectures make these ideas scalable with parallel implementations (Yang et al., 2024). Gated DeltaNet (Yang et al., 2025) adds decay or erasure mechanisms to the delta update that reduce the influence of older information. These mechanisms are useful architecturally, but it mathematically changes the reference predictor (referred to as a *comparator*): a decayed state is naturally aligned with discounted or drifting objectives. Meanwhile, preconditioned DeltaNet (Tumma et al., 2026) takes a different direction by transforming the updates toward the geometry of ridge regression. Our analysis turns these distinctions into upper/lower regret bounds, together with conditions under which approximate updates retain the guarantee of exact RLS.

Online learning guarantees. Our mathematical tools come from online convex optimization. Here, regret measures the additional loss relative to a comparator. Online gradient descent (OGD) provides a standard first-order method and its associated regret bounds (Zinkevich, 2003). We establish logarithmic regret in suitable square-loss and exp-concave settings with competitive online statistics (Vovk, 2001), online ridge regression, and online Newton analysis (Hazan et al., 2007). The ridge argument utilizes Follow-The-Regularized-Leader (FTRL) and Bregman geometry (McMahan, 2011). The noisy retrieval analysis uses a self-normalized martingale bound (Abbasi-Yadkori et al., 2011), while dynamic learning and discounting have related guarantees in drifting environments (Hazan and Seshadhri, 2009; Yuan and Lamperski, 2020; Jacobsen and Cutkosky, 2024). With the tools, we identify the prediction problem represented by each memory update.

3 Associative Online Prediction

We study a memory that learns KV associations one at a time. This follows the regression view of associative memory (Wang et al., 2025). Our two main questions: are *Which update rules achieve optima prediction regret?* and *When does learned memory also answer queries accurately?*. We separate these questions, as a memory can be evaluated on queries differing from its update keys.

The game. We consider an interaction sequence of horizon T. Before round $t \in [T]$, the state $W_t \in \mathbb{R}^{m\times d}$ depends only on earlier KV pairs. The learner observes a key $k_t \in \mathbb{R}^d$, predicts its value, and then receives $v_t \in \mathbb{R}^m$. It incurs squared binding loss $\ell(W) = \frac{1}{2}\|Wk_t - v_t\|_2^2$ and updates its state. We compare its total loss with that of a fixed matrix U:

$$\ell_t(W) = \tfrac{1}{2}\|Wk_t - v_t\|_2^2, \qquad \mathrm{Reg}_T(U) = \sum_{t=1}^{T}[\ell_t(W_t) - \ell_t(U)]. \tag{1}$$

Thus regret measures the extra cost of learning the associations sequentially, rather than using the same comparison map throughout. A query $q_t \in \mathbb{R}^d$ reads the memory through $W_t q_t$. Query error is a different quantity from binding regret. All feature vectors use a common coordinate system. Unless we specify a noise model, our statements hold for each realized stream.

Memory updates. The single-step linear Delta rule (Schlag et al., 2021) and TTT-Linear-SGD abstraction (Sun et al., 2024) take a gradient step:

$$W_{t+1} = W_t - \eta_t(W_t k_t - v_t)k_t^\top. \tag{2}$$

Recursive least squares (RLS) instead exactly fits all past pairs, with a ridge penalty that favors smaller matrices. For $\lambda > 0$, define

$$A_t = \lambda I_d + \sum_{s\le t} k_s k_s^\top, \quad C_t = \sum_{s\le t} v_s k_s^\top, \quad W_t = C_{t-1}A_{t-1}^{-1}. \tag{3}$$

Here, $A_0 = \lambda I_d$ and $C_0 = 0$. This state minimizes $\lambda\|W\|_{\mathrm{F}}^2/2 + \sum_{s<t}\ell_s(W)$. Writing $e_t = W_t k_t - v_t$ and $h_t = k_t^\top A_{t-1}^{-1}k_t$, its recursive update is

$$W_{t+1} = W_t - e_t\left(\frac{A_{t-1}^{-1}k_t}{1+h_t}\right)^\top. \tag{4}$$

Delta uses one scalar step size. RLS changes the update direction using the accumulated key covariance. Appendix B gives the online-learning interpretation and proofs.

4 Which updates achieve optimal prediction regret?

4.1 The cost of learning with exact ridge memory

The key quantity is how much a new observation changes the ridge fit. The leverage h_t measures the size of the current key relative to the information in past keys. Define

$$L_T = \log\det\left(I_d + \lambda^{-1}\sum_{t=1}^{T} k_t k_t^\top\right) = \sum_{t=1}^{T}\log(1+h_t). \tag{5}$$

If $\|k_t\|_2 \le 1$ and the key covariance has rank at most $\kappa \ge 1$, then $L_T \le \kappa \log(1 + T/(\lambda\kappa))$. This quantity accounts for both the number of observed directions and the amount of information collected along them.

THEOREM 1 (Ridge stability). *For exact RLS, every finite stream, and every $U \in \mathbb{R}^{m\times d}$,*

$$\mathrm{Reg}_T(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + S_T, \qquad S_T = \frac{1}{2}\sum_{t=1}^{T}\|e_t\|_2^2\left(1 - \frac{1}{(1+h_t)^2}\right). \tag{6}$$

If the realized, unclipped residuals satisfy $\|e_t\|_2 \le 2B$ for every t, then $S_T \le 4B^2 L_T$. The full proof is in Appendix B.3.

The bound separates the size of the comparison map from the cost of updating the memory. In fact, the tth summand in S_T is exactly $\ell_t(W_t) - \ell_t(W_{t+1})$: the reduction in current-pair loss after incorporating that pair. RLS has low regret when these changes have a small cumulative cost. Bounded residuals convert this stability cost into a log-determinant bound.

Why it holds. The FTRL inequality bounds regret by the ridge penalty plus these loss decreases (McMahan, 2011). The RLS update changes the residual from e_t to $e_t/(1+h_t)$, givin Eq. 6. Finally, $h_t/(1+h_t) \le \log(1+h_t)$ links stability to L_T.

The data-dependent bound needs no boundedness assumption. The logarithmic rate does require B to stay bounded as T grows. Bounded labels alone do not supply that condition for ordinary RLS.

4.2 A RULE THAT ATTAINS THE BOUNDED-GAME OPTIMUM

A simple change to the prediction removes the residual condition. The Vovk–Azoury–Warmuth (VAW) forecaster (Vovk, 2001; Azoury and Warmuth, 2001) uses the current key to shrink the RLS prediction :

$$\widehat{v}_t = C_{t-1}A_t^{-1}k_t = \frac{W_t k_t}{1+h_t}. \tag{7}$$

This is causal: A_t uses the current key, but not its unseen value. The memory statistics remain those in Eq. 3. Only the played binding prediction changes. We write $\mathrm{Reg}_T(\widehat{v}; U)$ when the learner loss in Eq. 1 uses $\widehat{v}_t$.

THEOREM 2 (Bounded prediction and its minimax rate). *If $\|v_t\|_2 \le B_v$, VAW satisfies, for every U,*

$$\mathrm{Reg}_T(\widehat{v}; U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + \frac{B_v^2}{2}L_T. \tag{8}$$

For scalar values, let $T \ge 2\kappa$, $1 \le \kappa \le d$, $\|k_t\|_2 \le 1$, $|v_t| \le 1$, and let the key covariance have rank at most κ. Against the best fixed comparator with $\|u\|_2 \le \sqrt{\kappa}$, the smallest worst-case expected regret among all causal forecasters, denoted as $\mathfrak{R}_T^*$, obeys

$$\frac{\kappa}{48}\log(T/\kappa) \le \mathfrak{R}_T^* \le \frac{\kappa}{2} + \frac{\kappa}{2}\log(1 + T/\kappa). \tag{9}$$

VAW with $\lambda = 1$ attains the upper bound. The full proof is in Appendix B.4.

Interpretation. The rate $\Theta(\kappa \log(T/\kappa))$ is not merely an upper bound for one recurrence. No causal predictor can improve its worst-case order in this game. The lower bound uses κ separate basis-key problems. Each requires estimating an unknown Bernoulli mean from past observations, and their estimation costs add. The classical VAW guarantee supplies the matching upper bound.

The readout distinction matters. Eq. 8 applies to $W_t k_t/(1+h_t)$, not automatically to the ungated output $W_t k_t$. It also does not establish optimality for query retrieval. The comparator radius and bounded scalar-output game are part of the minimax statement.

4.3 ONE CONSTANT DELTA STEP CAN BE INSUFFICIENT

An upper bound for RLS does not single-handedly show that Delta performs worse. Therefore, we compare the updates on the same unit-key stream. The construction has one frequently excited direction and one weakly excited direction. A small scalar step learns the weak direction slowly, while a larger step admits more noise in the strong direction.

Proposition 1 (Separation for constant scalar steps). *Let $T \geq 16$, $p = T^{-1/2}$, and $U^\star = (0, 1)$. Set*

$$k_t = (\sqrt{1-p},\ s_t\sqrt{p})^\top, \qquad v_t = U^\star k_t + \xi_t, \tag{10}$$

where all s_t and ξ_t are independent uniform signs. Delta with $W_1 = 0$ and any constant $\eta \in [0, 1]$ satisfies

$$\mathbb{E}[\mathrm{Reg}_T^{\mathrm{Delta}}(U^\star)] \geq \frac{\sqrt{T}}{64}, \qquad \mathrm{Reg}_T^{\mathrm{VAW}}(U^\star) \leq \frac{1}{2} + 4\log(1 + T/2) \tag{11}$$

for VAW with $\lambda = 1$. The VAW bound holds on every realization. The full proof is in Appendix B.5.

Interpretation. One fixed scalar step cannot resolve the learning-speed/noise tradeoff in this example, even when chosen using the horizon and the key distribution. The key norms are all one, so normalizing the keys does not remove the separation. The result concerns constant effective steps; it is not a lower bound for every adaptive Delta schedule. Its labels satisfy $|v_t| \leq 2$, and its covariance condition number grows as $\sqrt{T} - 1$.

5 When does key learning support query retrieval?

Binding regret evaluates predictions on update keys. Retrieval can ask about different directions. For exact RLS, the output has the form

$$W_t q_t = \sum_{s<t} v_s\, k_s^\top A_{t-1}^{-1} q_t. \tag{12}$$

Thus, the same inverse covariance used for learning also weights the past values at retrieval time. Conditional on fixed keys and queries, this is a causal linear operator. Its strictly lower off-diagonal blocks have rank at most d, the semi-separable property proved in Appendix B.2. This does not make an entire neural layer linear when its features depend on the inputs.

5.1 Retrieval error is controlled by query leverage

A query is easier to answer when its directions have been well observed through keys. We measure this using

$$Q_T = \sum_{t=1}^{T} q_t^\top A_{t-1}^{-1} q_t, \qquad \mathcal{E}_T = \sum_{t=1}^{T} \|(W_t - U^\star)q_t\|_2^2, \tag{13}$$

where $U^\star$ is the fixed map generating the values. The following result turns query leverage into an error guarantee.

Theorem 3 (Retrieval from exact ridge memory). *Let W_t be exact RLS with $\lambda > 0$. If $v_t = U^\star k_t$ for every t, then, for any queries,*

$$\mathcal{E}_T \leq \lambda \|U^\star\|_{\mathrm{F}}^2 Q_T. \tag{14}$$

Alternatively, suppose $v_t = U^\star k_t + \xi_t$. Conditional on all keys and queries, assume $\mathbb{E}[\xi_t] = 0$, $\mathbb{E}[\xi_t \xi_t^\top] = \sigma^2 I_m$, and $\mathbb{E}[\xi_s \xi_t^\top] = 0$ for $s \neq t$. Then

$$\mathbb{E}[\mathcal{E}_T \mid k_{1:T}, q_{1:T}] \leq (\lambda \|U^\star\|_{\mathrm{op}}^2 + m\sigma^2) Q_T. \tag{15}$$

For random designs, the unconditional bound uses $\mathbb{E}[Q_T]$. The full proof is in Appendix C.

The memory need not see the exact query as an earlier key. It needs enough information in the directions used by that query. In the noiseless case, sublinear Q_T gives vanishing average retrieval error when $\lambda\|U^\star\|_{\mathrm{F}}^2$ is fixed. Under the stated noise model, the bound adds a noise contribution proportional to $m\sigma^2$. The operator norm measures the target map's largest gain along a unit direction.

The conditional moment assumptions hold, for example, for noise-independent keys and queries with independent mean-zero noise. They need not hold when future features depend on earlier noise. Appendix C.3 gives a separate high-probability result for predictable keys and conditionally sub-Gaussian noise, using the results of Abbasi-Yadkori et al. (2011).

5.2 WHAT QUERY COVERAGE MUST PROVIDE

The previous theorem leaves a concrete question: when is Q_T small? The answer requires a relation between keys and queries. The next result gives both a sufficient cumulative condition and an obstruction that applies to every learner.

Proposition 2 (Covered and unseen query directions)**.** *Suppose $\|q_t\|_2 \le 1$, the query span has dimension at most κ, and, for every prefix $n \le T$,*

$$\sum_{s\le n} q_s q_s^\top \preceq c \sum_{s\le n} k_s k_s^\top, \qquad c>0, \quad c\lambda \ge 1. \tag{16}$$

Then $Q_T \le 2c\kappa \log(1 + T/(c\lambda\kappa))$. Conversely, fix past keys and a query q, and let $q_\perp$ be its component orthogonal to those keys. In noiseless regression, every causal learner with output $\widehat{r}_t$ satisfies

$$\sup_{\|U\|_{\mathrm{F}}\le R} \mathbb{E}[\|\widehat{r}_t - Uq\|_2^2] \ge R^2 \|q_\perp\|_2^2. \tag{17}$$

The expectation is over learner randomization. The full proof is in Appendix C.4.

Condition Eq. 16 requires key observations to keep pace with query demand in every direction. Together with Theorem 3, it yields average retrieval error of order $\kappa \log(1 + T)/T$ when the other factors are fixed. By contrast, past pairs contain no information about an entirely unseen direction. For example, keys along the first coordinate cannot distinguish two maps that disagree only on the second coordinate. This is why low binding regret alone is not a retrieval guarantee. Both estimation error and query coverage matter. Eq. 17 is a one-query lower bound for the unseen component, not a matching lower bound for all of Q_T.

6 BEYOND EXACT MEMORY AND A FIXED TARGET

Approximation must preserve prediction, not only covariance. Reducing the cost of the inverse can introduce persistent prediction bias. Let $W_t^\star$ be exact RLS, $\widetilde{W}_t$ an approximate causal state, and $a_t = \ell_t(\widetilde{W}_t) - \ell_t(W_t^\star)$. If, for $\alpha, \varepsilon_T \ge 0$, the signed excess satisfies $\sum_t a_t \le \alpha S_T + \varepsilon_T$, then

$$\mathrm{Reg}_T(\widetilde{W}; U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + (1+\alpha)S_T + \varepsilon_T. \tag{18}$$

This sufficient certificate allows an approximation to outperform RLS on individual rounds. Uniform control of its cumulative excess transfers the exact-memory guarantee. A fixed relative inverse error alone is insufficient: Appendix D gives a scalar closed-form approximation with linear regret, as well as direct guarantees for frozen-chunk and specified sketched updates. These are different recurrences with different costs.

A changing target changes the comparison. When the target is U_t, define $\mathrm{DynReg}_T(U_{1:T}) = \sum_t [\ell_t(W_t) - \ell_t(U_t)]$. Remembering every observation equally can then hurt. For $T = 2n$, $k_t = 1$, and values that change from 0 to J after round n, we obtain

$$\mathrm{DynReg}_T^{\mathrm{RLS}} \ge \frac{J^2 T}{16}, \qquad \mathrm{DynReg}_T^{\mathrm{Delta}} \le \frac{J^2}{2\eta(2-\eta)} \tag{19}$$

for ordinary RLS with any $\lambda > 0$ and Delta with $W_1 = 0$, $0 < \eta \le 1$. Both are compared with $U_t = v_t$. Old observations slow RLS after the change; Delta's error decays geometrically.

Resets and discounting address this mismatch by limiting the influence of old pairs. Scalar decay can also introduce bias, so forgetting is not a universal improvement. Appendix E analyzes other implementations, including scalar decay, known-boundary resets, a causal multi-start VAW wrapper, and discounted RLS separately. The discounted result retains noise, ridge bias, target variation, and a quantitative condition requiring continued observation of the relevant key directions. Related dynamic-regret approaches use their own update and comparator assumptions (Jacobsen and Cutkosky, 2024).

7 EXPERIMENTS

7.1 RLS VERSUS DELTA ON WIKITEXT STREAMS

Here, the two experiments are to validate the quantities aligned with theorems. The fast-weight algorithms share the same fixed key and value representations inside each paired run, so comparisons are paired by seed, task, context length, and stream. The main metrics are cumulative KV loss, static regret against the offline ridge comparator, dynamic regret against the known drifting comparator in synthetic streams, and per-token mean-squared KV error. We also report win counts, which depends only on the direction of each paired comparison.

Experiment 1 uses token streams derived from WikiText (Merity et al., 2016). A vocabulary of 2048 frequent tokens is built from the training split, and validation plus test text supplies evaluation streams. Each token is mapped to a fixed random unit-normalized key in $\mathbb{R}^{32}$ and a fixed random unit-normalized value in $\mathbb{R}^{16}$. Two tasks are used. In the identity task, the value corresponds to the same token as the key, testing stable associative binding. In the next-token task, the key is the current token and the value is the next token, adding a noisier predictive binding problem. This gives 48 paired streams across 8 seeds, 2 tasks, and context lengths 1024, 2048, and 4096. The compared update rules are linear attention (Katharopoulos et al., 2020), Delta-SGD, gated Delta (Yang et al., 2025), and exact RLS, with the main comparison between Delta-SGD and exact RLS. The full experimental details are provided in Appendix F.

WikiText comparison	RLS mean	Delta mean	Paired diff.	RLS wins
All static regret	56.45	165.2	-108.8	48/48
All KV MSE	0.7711	0.8590	-0.0879	48/48
Identity static regret	53.93	100.1	-46.18	24/24
Next-token static regret	58.97	230.3	-171.4	24/24

Table 1: WikiText paired comparison between exact RLS and Delta-SGD.

Table 1 shows that exact RLS beats Delta on static regret and on KV MSE across all 48 WikiText streams. The effect is not concentrated in one task. On identity binding, RLS wins all 24 paired identity streams. RLS again wins all 24 paired streams on next-token binding. These results are consistent with the second-order classification: once the task is a stable KV regression problem, the second-order update utilized covariance and improves regret.

Experiment 2 constructs controlled drifting associative maps. Keys are random unit-normalized vectors in $\mathbb{R}^{32}$, values are generated as $v_t = U_t k_t + \xi_t$ with $U_t \in \mathbb{R}^{16\times 32}$, and the comparator sequence U_t is known. Each stream has length 2048. Abrupt drift streams change the active coordinates of U_t at block boundaries, with jump scales 0.7 and 1.2. Gradual drift streams evolve the active coordinates continuously, with drift rates 0.02 and 0.05. Both sparse-coordinate and full-coordinate drift are tested across eight seeds. This yields 64 controlled drifting streams. The compared algorithms include Delta-SGD, scalar gated Delta, oracle reset Delta, static RLS, discounted RLS, and oracle reset RLS.

A (adaptive) versus B (baseline)	A mean	B mean	Paired diff.	A wins
Abrupt reset Delta vs Delta	2024	3317	-1293	31/32
Abrupt reset RLS vs static RLS	1627	7797	-6170	32/32
Abrupt discounted RLS vs static RLS	3519	7797	-4277	32/32
Gradual discounted RLS vs static RLS	94.91	252.9	-158.0	32/32
Gradual scalar gated Delta vs Delta	95.54	79.59	15.95	0/32

Table 2: Controlled drift comparison between adaptive and baseline memories.

The abrupt-drift results (Table 2) show the value of matched memory erasure. Oracle-reset Delta improves dynamic regret and wins on 31 of 32 abrupt streams. Similarly, oracle-reset RLS improves dynamic regret for static RLS and wins on all 32 abrupt streams. Discounted RLS also improves

abrupt drift and gradual drift. We also include a negative control on gradual drift, where Delta outperforms its scalar gated version, winning all 32 paired streams. This is one important point of our findings: forgetting helps when the forgetting structure matches the drift, and it can hurt when it discards useful memory on the wrong time scale.

Moreover, we report the realized RLS stability $S_T = \sum_{t=1}^{T} \frac{1}{2}\|e_t\|_2^2 \left(1 - \frac{1}{(1+h_t)^2}\right)$, alongside the log-determinant potential L_T. We also report the effective residual scale $B_{\text{eff}}^2 = S_T/(4L_T)$ when $L_T > 0$. The results in Table 3 show that Experiment 1 exhibits a lower effective residual scale than Experiment 2 (0.064 vs. 0.114), indicating smaller stability cost relative to complexity. However, S_T remains moderate to L_T in both settings.

Table 3: Empirical RLS stability diagnostics.

Experiment	RLS runs	Mean S_T	Mean L_T	Mean B_{eff}^2
Experiment 1	18	12.74	42.59	0.064
Experiment 2	36	33.22	78.83	0.114

7.2 Attention reconstruction on LongBench-v2

We now test our theoretical findings with a Transformer-based language model.

Setup. We test memory updates on frozen pretrained features using LongBench-v2 (Bai et al., 2025) We extract normalized pre-RoPE queries and keys, and native-scale values, from layer 14 of Qwen2.5-7B-Instruct (Bai et al., 2023). Each memory learns from context tokens and is then queried at question and option positions, without further updates. Inputs are capped at 32,768 tokens, with the complete question and options preserved. The split contains 101 development and 402 held-out examples, including 368 test contexts. Identical contexts remain in the same partition. Each method family receives eight development configurations. The primary comparison is exact RLS against the better of constant step/data-dependent Delta.

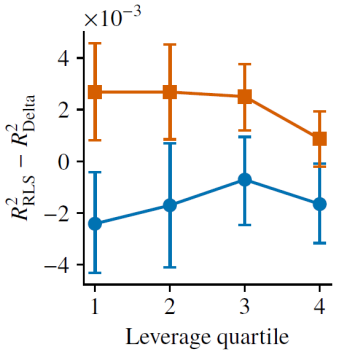


Figure 1: Query leverage and the RLS–Delta reconstruction gap.

Evaluation. We reconstruct two softmax targets: native attention, which also includes the causal question/option prefix and the current token, and context-restricted attention, which uses only the tokens stored in the memory. We report R^2_{mean}, defined as one minus the ratio of squared reconstruction error to the error of a running-mean value baseline. Higher scores are better, but this is not centered regression R^2. Scores are averaged across four KV groups within each example. Confidence intervals use 10,000 bootstrap resamples of context groups.

Table 4: Question/option attention reconstruction on 402 held-out examples. Means are R^2_{mean} scores. Paired differences are RLS minus the selected constant-step Delta, with intervals being 95% context-group bootstrap intervals. Both targets use the settings selected for native reconstruction.

Target	RLS	Delta	Paired difference [95% CI]
Native (primary)	0.209	0.211	-0.0016 $[-0.0026, -0.0006]$
Context-restricted	0.503	0.491	$+0.0116$ $[0.0092, 0.0144]$

Results. The ranking changes with the reconstruction target (Table 4). Selected Delta is slightly better on the primary native-attention target, whereas RLS is better on the context-restricted target. Both paired intervals exclude zero. Thus, the evaluation target matters when comparing these memories: RLS does not uniformly reconstruct native attention better than Delta. At a common ridge parameter, chunked RLS matches full-context RLS reconstruction because both reach the same final ridge state, although their online losses differ. Fig 1 shows that the comparison between RLS and

Delta changes with the query set. Across the displayed leverage quartiles, the mean difference favors Delta on own questions and RLS on foreign questions. Higher leverage does not produce a steadily larger RLS advantage. Thus, leverage alone does not identify the better reconstruction method. These are observational comparisons: foreign questions do not necessarily introduce unseen key directions, and the retrieval guarantees do not directly bound softmax reconstruction error.

Connection to the theory. The largest observed ratio of RLS regret to its exact stability bound is 0.998. Query leverage has a partial Spearman association of 0.184 with RLS reconstruction error after adjusting for context length, but does not by itself rank the methods. These findings separate a binding-regret guarantee from attention reconstruction: the former does not automatically guarantee the latter. The experiment evaluates the reconstruction of frozen features, not LongBench answer accuracy or trained TTT layers. Full details can be found in Appendix G.

8 Discussion and Limitations

Our main finding is that learning stored associations, answering queries, and tracking changing associations require different guarantees. Exact RLS uses the accumulated key covariance to support both prediction and retrieval analyses. However, low binding regret alone does not ensure accurate retrieval. Here, the observed keys must provide enough information about the query directions, and the error in the learned map must also be controlled. When associations change, retaining every old observation with equal weight can slow adaptation down. Resets and discounting may help, but scalar decay can also introduce persistent bias. Therefore, the forgetting mechanism needs an analysis of the particular update, rather than a general claim that it improves memory.

The LongBench-v2 results highlight the importance of the evaluation target. Selected Delta is slightly better on native-attention reconstruction, whereas RLS is better on context-restricted reconstruction. The native target includes question and option tokens that do not update the associative memory, while the context-restricted target uses only the stored context. This comparison does not directly test our fixed-map retrieval guarantees, but it shows why a binding-regret guarantee should not be understood as a guarantee of better attention reconstruction.

Several limitations remain. The minimax result concerns the stated bounded squared-loss game, and the Delta separation covers constant scalar steps, not arbitrary adaptive schedules. Retrieval rates require the stated coverage and assumptions about data generation. Approximation audits on finite streams do not establish guarantees for all streams, and the discounted RLS bound requires continued observation of relevant key directions. Empirically, we isolate memory updates using fixed features. The LongBench-v2 study only uses one frozen checkpoint with a $32,768$-token input cap. It does not aim to measure answer accuracy, performance across trained checkpoints, or the benefits of jointly training the representations and memory update.

9 Conclusion

We studied recurrent memory updates from the perspective of online associative prediction. We showed that exact RLS connects prediction and retrieval through a shared covariance structure, while the VAW readout attains the optimal prediction-regret rate in the stated bounded game. Retrieval and adaptation to changing targets require additional conditions and separate analyses. The mixed experimental results reinforce this distinction: strong binding guarantees do not automatically imply better attention reconstruction. Thus, memory updates should be assessed against the target they predict, the information they retain, and the assumptions their guarantees require.

AI use statement

In this paper, we use generative AI tools, namely ChatGPT, to help with polishing our writing and narration of our theoretical findings.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Aksenov, Y., Balagansky, N., Vaina, S. L. C., Shaposhnikov, B., Gorbatovski, A., and Gavrilov, D. (2024). Linear transformers with learnable kernel functions are better in-context models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9584–9597.

Azoury, K. S. and Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine learning*, 43(3):211–246.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, Y., Tu, S., Zhang, J., Peng, H., Wang, X., Lv, X., Cao, S., Xu, J., Hou, L., Dong, Y., et al. (2025). Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664.

Chen, Y., Wang, J., Yang, Z., Manivasagam, S., and Urtasun, R. (2024). G3r: Gradient guided generalizable reconstruction. In *European Conference on Computer Vision*, pages 305–323. Springer.

Gandelsman, Y., Sun, Y., Chen, X., and Efros, A. (2022). Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385.

Ghashami, M., Liberty, E., Phillips, J. M., and Woodruff, D. P. (2016). Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Han, D., Li, Y., Li, T., Cao, Z., Wang, Z., Song, J., Cheng, Y., Zheng, B., and Huang, G. (2025). Vit ˆ3: Unlocking test-time training in vision. *arXiv preprint arXiv:2512.01643*.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192.

Hazan, E. and Seshadhri, C. (2009). Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, pages 393–400.

Jacobsen, A. and Cutkosky, A. (2024). Online linear regression in dynamic environments via discounting. In *Proceedings of the 41st International Conference on Machine Learning*.

Kasai, J., Peng, H., Zhang, Y., Yogatama, D., Ilharco, G., Pappas, N., Mao, Y., Chen, W., and Smith, N. A. (2021). Finetuning pretrained transformers into rnns. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 10630–10643.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Liu, B., Wang, R., Wu, L., Feng, Y., Stone, P., and Liu, Q. (2025). Longhorn: State space models are amortized online learners. In *International Conference on Learning Representations*, volume 2025, pages 95419–95434.

Liu, J., Elflein, S., Litany, O., Gojcic, Z., and Li, R. (2026). Test-time training with kv binding is secretly linear attention. *arXiv preprint arXiv:2602.21204*.

Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., and Alahi, A. (2021). Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820.

Luo, H., Agarwal, A., Cesa-Bianchi, N., and Langford, J. (2016). Efficient second order online learning by sketching. *Advances in Neural Information Processing Systems*, 29.

McMahan, B. (2011). Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 525–533. JMLR Workshop and Conference Proceedings.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models.

Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N. A., and Kong, L. (2021). Random feature attention. *arXiv preprint arXiv:2103.02143*.

Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., and Zhong, Y. (2022). cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*.

Schlag, I., Irie, K., and Schmidhuber, J. (2021). Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pages 9355–9366. PMLR.

Sun, Y., Li, X., Dalal, K., Xu, J., Vikram, A., Zhang, G., Dubois, Y., Chen, X., Wang, X., Koyejo, S., et al. (2024). Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR.

Tumma, N., Loo, N., and Rus, D. (2026). Preconditioned deltanet: Curvature-aware sequence modeling for linear recurrences. *arXiv preprint arXiv:2604.21100*.

Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review*, 69(2):213–248.

Wang, K. A., Shi, J., and Fox, E. B. (2025). Test-time regression: a unifying framework for designing sequence models with associative memory. *arXiv preprint arXiv:2501.12352*.

Yang, S., Kautz, J., and Hatamizadeh, A. (2025). Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pages 29687–29707.

Yang, S., Wang, B., Zhang, Y., Shen, Y., and Kim, Y. (2024). Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522.

Yuan, J. and Lamperski, A. (2020). Trading-off static and dynamic regret in online least-squares and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6712–6719.

Yuan, Y., Shen, Q., Wang, S., Yang, X., and Wang, X. (2026). Test3r: Learning to reconstruct 3d at test time. *Advances in Neural Information Processing Systems*, 38:100445–100464.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936.

A Classification of Online Memories

Here, Table 5 provides a compact classification of the online memories in this paper.

Table 5: Classification of associative memory updates The guarantees apply to the stated recurrences and assumptions, not to all architectures sharing the same name. Here, $L_T = \log\det\Big(I + \lambda^{-1}\sum_{t=1}^{T} k_t k_t^\top\Big)$, S_T is the exact RLS stability cost, $P_T = \sum_{t=2}^{T} \|U_t - U_{t-1}\|_{\mathrm{F}}$, and κ is the design rank.

Memory / update	Learning rule	Main assumptions	Guarantee
Delta / TTT-SGD	First-order gradient step	Bounded domain and gradients for the generic OGD guarantee. The separation result considers a constant scalar step.	Generic OGD: $\mathrm{Reg}_T(U) \le \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_{\mathrm{F}}^2 = O(DG\sqrt{T})$ with the standard choice of η. Moreover, there exists a unit-key stream on which every constant $\eta \in [0,1]$ has $\mathbb{E}[\mathrm{Reg}_T] = \Omega(\sqrt{T})$, while VAW has $O(\log T)$ regret. See Proposition 7 and Proposition 1.
Exact RLS / exact preconditioned TTT	Quadratic FTRL / recursive least squares	No boundedness assumption for the data-dependent bound. For the log-determinant form, the realized residuals must satisfy $\|e_t\|_2 \le 2B$.	Always: $\mathrm{Reg}_T(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + S_T$. Under residual stability: $\mathrm{Reg}_T(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + 4B^2 L_T$. See Theorem 1.
VAW-gated RLS	Exact ridge state with a key-dependent output gate	Bounded values $\|v_t\|_2 \le B_v$. For the minimax statement, bounded keys and comparator norm, with design rank at most κ.	$\mathrm{Reg}_T(\widehat{v}; U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + \frac{B_v^2}{2} L_T$. For the stated bounded scalar-output game: $\mathfrak{R}_T^\star = \Theta\big(\kappa \log \frac{T}{\kappa}\big)$. See Theorem 2.
Scalar-gated Delta	Delta update with multiplicative state decay	$\|k_t\|_2 \le 1$, bounded values, $0 < \eta \le \gamma < 1$, bounded comparator path.	Dynamic regret: $\mathrm{DynReg}_T \le \frac{D^2 + 2DP_T}{2\eta} + \frac{\eta G_\mu^2 T}{2} + \frac{(1-\gamma)R^2 T}{2\eta}$. The final term is the decay bias. A scalar static example gives linear regret under persistent decay. See Proposition 8.
Reset / adaptive ridge	Restarted RLS or VAW, optionally a causal multi-start VAW wrapper	Comparator is piecewise constant with N segments; bounded residuals for ordinary RLS, or bounded values for VAW.	With known reset boundaries: $\mathrm{Reg}_T^{\mathrm{switch}} \le \frac{N\lambda R^2}{2} + c_\star N\kappa \log\big(1 + \frac{T}{N\lambda\kappa}\big)$, where $c_\star = 4B^2$ for RLS and $c_\star = B_v^2/2$ for VAW. The causal wrapper handles unknown boundaries with logarithmic overhead. See Proposition 9 and Proposition 10.
Discounted RLS	Exponentially discounted ridge statistics	Mean-zero noise, bounded comparator path, and the quantitative excitation condition of Proposition 24.	$\mathbb{E}[\mathrm{DynReg}_T(U_{1:T})] \le \mathcal{B}_T(\delta)$, where $\mathcal{B}_T(\delta)$ contains explicit noise, ridge-bias, discount, excitation, and path-variation terms. See Proposition 11 and Eq. E.17.
Chunked RLS	RLS statistics updated by chunks; predictions frozen inside each chunk	Chunk size at most b and chunk-start residual norm at most $2B_e$.	$\mathrm{Reg}_T^{\mathrm{chunk}}(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + 4bB_e^2 L_T$. The exact block-stability expression is tighter; the factor b captures worst-case staleness. See Proposition 4.
FD-sketched RLS	Frequent-Directions curvature with residual-feedback update	Specified FD inflation, bounded residuals, and a bounded projection domain.	$\mathrm{Reg}_T(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + 2B^2 \log\det(\lambda^{-1}\widetilde{A}_T) + \frac{D^2\rho_T}{2}$, with $\rho_T \le \frac{\Omega_r}{\ell - r}$. See Proposition 5.
Certified approximate preconditioning	Any causal approximation calibrated to exact RLS loss	Signed approximation error satisfies $\sum_t a_t \le \alpha \sum_t s_t + \varepsilon_T$. The log-det form also needs residual-stable exact RLS.	Always: $\mathrm{Reg}_T(\widetilde{W}; U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + (1+\alpha)S_T + \varepsilon_T$. Under residual stability: $\mathrm{Reg}_T(\widetilde{W}; U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + 4B^2(1+\alpha)L_T + \varepsilon_T$. See Eq. 18 and Appendix D.1. A fixed relative inverse error can still cause linear regret, see Proposition 6.

B PREDICTION: BACKGROUND AND PROOFS

The main text separates the scientific conclusions from the proof tools. This appendix gives the shared online-learning argument, the exact ridge identities, and the prediction lower bounds. Throughout, $\langle X, Y\rangle = \mathrm{Tr}(XY^\top)$ and $\|X\|_H^2 = \mathrm{Tr}(XHX^\top)$ for positive definite H.

B.1 GRADIENT UPDATES AND THE STANDARD REGRET ARGUMENT

The Delta update is an online gradient step on the squared binding loss. This connection explains which standard regret bounds apply. It is not a separate optimality claim for squared loss.

Let $\mathcal{W}$ be a nonempty closed convex domain. For a differentiable convex potential Ψ, define its Bregman divergence by

$$D_\Psi(X, Y) = \Psi(X) - \Psi(Y) - \langle \nabla\Psi(Y), X - Y\rangle. \tag{B.1}$$

Suppose Ψ_t is 1-strongly convex in a norm $\|\cdot\|_t$, $g_t = \nabla\ell_t(W_t)$, and the following minimizer exists:

$$W_{t+1} = \arg\min_{W\in\mathcal{W}}\{\eta_t\langle g_t, W\rangle + D_{\Psi_t}(W, W_t)\}, \qquad \eta_t > 0. \tag{B.2}$$

Write $\|\cdot\|_{t,*}$ for the dual norm. The standard mirror-descent argument gives

$$\ell_t(W_t) - \ell_t(U) \le \frac{D_{\Psi_t}(U, W_t) - D_{\Psi_t}(U, W_{t+1})}{\eta_t} + \frac{\eta_t}{2}\|g_t\|_{t,*}^2. \tag{B.3}$$

To see this, the optimality condition for Eq. B.2, followed by the three-point Bregman identity, gives

$$\eta_t\langle g_t, W_{t+1} - U\rangle \le D_{\Psi_t}(U, W_t) - D_{\Psi_t}(U, W_{t+1}) - D_{\Psi_t}(W_{t+1}, W_t), \tag{B.4}$$

$$\eta_t\langle g_t, W_t - W_{t+1}\rangle \le \tfrac{1}{2}\eta_t^2\|g_t\|_{t,*}^2 + D_{\Psi_t}(W_{t+1}, W_t). \tag{B.5}$$

Add the inequalities and use convexity of ℓ_t. For constant $\Psi_t = \Psi$ and $\eta_t = \eta$, the divergences telescope:

$$\mathrm{Reg}_T(U) \le \frac{D_\Psi(U, W_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_{\Psi,*}^2. \tag{B.6}$$

Euclidean $\Psi(W) = \|W\|_{\mathrm{F}}^2/2$ gives online gradient descent (Zinkevich, 2003). In particular, if $\|U - W_1\|_{\mathrm{F}} \le D$, $\|g_t\|_{\mathrm{F}} \le G$, and $D, G > 0$, then $\eta = D/(G\sqrt{T})$ gives $DG\sqrt{T}$. These distance and gradient bounds must be established for the actual recurrence. They do not follow merely from calling it Delta.

For changing potentials or step sizes, the sum instead contains

$$\begin{aligned}\mathrm{Reg}_T(U) \le \frac{D_{\Psi_1}(U, W_1)}{\eta_1} + \sum_{t=1}^{T-1}\left[\frac{D_{\Psi_{t+1}}(U, W_{t+1})}{\eta_{t+1}} - \frac{D_{\Psi_t}(U, W_{t+1})}{\eta_t}\right] \\ + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|_{t,*}^2.\end{aligned} \tag{B.7}$$

The middle term is the cost of changing the geometry. It cannot generally be discarded. For squared binding loss, the unconstrained mirror step with $\Psi_t(W) = \|W\|_{A_t}^2/2$ and $\eta_t = 1$ equals the RLS update. We analyze its regret using the full quadratic objective, rather than claiming a cancellation in Eq. B.7 without proof. The relation between these update views is standard (McMahan, 2011).

B.2 The ridge state and its causal operator

Define $F_{t-1}(W) = \lambda\|W\|_{\mathrm{F}}^2/2 + \sum_{s<t}\ell_s(W)$. Its normal equations are

$$W_t A_{t-1} = C_{t-1}, \tag{B.8}$$

so $W_t = C_{t-1}A_{t-1}^{-1}$. This proves the state formula and the causal readout Eq. 12.

For fixed keys and queries, let $K_{t,s} = k_s^\top A_{t-1}^{-1} q_t$ for $s < t$, and zero otherwise. For any cut r, the block with rows $t > r$ and columns $s \le r$ factors as

$$K_{\{t>r\},\{s\le r\}} = \begin{bmatrix}(A_r^{-1}q_{r+1})^\top \\ (A_{r+1}^{-1}q_{r+2})^\top \\ \vdots\end{bmatrix}\begin{bmatrix}k_1 & \cdots & k_r\end{bmatrix}. \tag{B.9}$$

Its rank is at most d. This is the semi-separable structure used in the paper. The statement is conditional on the features: when keys or queries depend on values, it does not imply that the overall input–output map is linear.

B.3 Proof of Theorem 1

Let $R(W) = \lambda\|W\|_{\mathrm{F}}^2/2$. Exact RLS is a follow-the-regularized-leader update. The regularized be-the-leader inequality gives

$$\sum_{t=1}^{T} \ell_t(W_{t+1}) + R(W_1) \le \sum_{t=1}^{T} \ell_t(U) + R(U). \tag{B.10}$$

Here $W_1 = 0$ and $R(W_1) = 0$. Therefore

$$\mathrm{Reg}_T(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + \sum_{t=1}^{T}[\ell_t(W_t) - \ell_t(W_{t+1})]. \tag{B.11}$$

Sherman–Morrison gives $A_t^{-1}k_t = A_{t-1}^{-1}k_t/(1+h_t)$. From the normal equations,

$$W_{t+1} = W_t - e_t(A_t^{-1}k_t)^\top, \qquad W_{t+1}k_t - v_t = \frac{e_t}{1+h_t}. \tag{B.12}$$

Consequently,

$$s_t := \ell_t(W_t) - \ell_t(W_{t+1}) = \frac{1}{2}\|e_t\|_2^2\left(1 - \frac{1}{(1+h_t)^2}\right). \tag{B.13}$$

Substitution into Eq. B.11 proves the unconditional bound.

If $\|e_t\|_2 \le 2B$, then

$$s_t \le 2B^2\frac{h_t(2+h_t)}{(1+h_t)^2} \le 4B^2\frac{h_t}{1+h_t}. \tag{B.14}$$

The determinant lemma gives $\log\det A_t - \log\det A_{t-1} = \log(1+h_t)$. Since $h/(1+h) \le \log(1+h)$ for $h \ge 0$, summing proves $S_T \le 4B^2 L_T$.

For the rank bound, let $\mu_1, \dots, \mu_\kappa \ge 0$ be the covariance eigenvalues, padded with zeros. Unit-bounded keys give $\sum_i \mu_i \le T$. Concavity of the logarithm yields

$$L_T = \sum_{i=1}^{\kappa} \log(1 + \mu_i/\lambda) \le \kappa\log(1 + T/(\lambda\kappa)). \tag{B.15}$$

If all keys are zero, $L_T = S_T = 0$. The residuals in this proof are those of the exact, unclipped RLS trajectory.

B.4 The VAW bound and the minimax lower bound

VAW upper bound. We include a proof of the classical prediction inequality (Vovk, 2001) to make the readout distinction explicit. Put $p_t = W_t k_t$ and $F_t^{\min} = \min_W F_t(W)$. Completing the square gives

$$F_t^{\min} - F_{t-1}^{\min} = \frac{\|p_t - v_t\|_2^2}{2(1+h_t)}. \tag{B.16}$$

Expanding the VAW loss gives the exact identity

$$\frac{1}{2}\|\frac{p_t}{1+h_t} - v_t\|_2^2 = F_t^{\min} - F_{t-1}^{\min} + \frac{h_t\|v_t\|_2^2}{2(1+h_t)} - \frac{h_t\|p_t\|_2^2}{2(1+h_t)^2}. \tag{B.17}$$

Drop the last term, sum, and use $F_0^{\min} = 0$ and $F_T^{\min} \le F_T(U)$. For every finite stream,

$$\mathrm{Reg}_T(\widehat{v}; U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + \frac{1}{2}\sum_{t=1}^{T}\|v_t\|_2^2\frac{h_t}{1+h_t}. \tag{B.18}$$

Bounded values and the determinant identity give Eq. 8. No bound on p_t or its residual was used.

Lower bound. Let $n = \lfloor T/\kappa \rfloor \ge 2$. Use each basis key e_i, $i \le \kappa$, for n rounds. Set the remaining keys and labels to zero. Independently for each coordinate, sample $p_i \sim \mathrm{Beta}(1,1)$ and then labels $y_{i,s} \mid p_i \sim \mathrm{Bernoulli}(p_i)$. The full stream is sampled before play.

The comparator $u = (p_1, \dots, p_\kappa, 0, \dots, 0)$ belongs to $[0,1]^\kappa \times \{0\}^{d-\kappa}$. At its sth observation of coordinate i, every learner has conditional expected excess loss over this comparator equal to $\frac{1}{2}\mathbb{E}[(\widehat{y}_{i,s} - p_i)^2]$. Other coordinates and the learner's independent random seed reveal nothing about p_i. Its posterior variance is therefore a lower bound on the squared estimation error.

After $s-1$ observations with j successes, the posterior is $\mathrm{Beta}(j+1, s-j)$. Under the uniform prior, j is uniform on $\{0, \dots, s-1\}$. Hence

$$\mathbb{E}\mathrm{Var}(p_i \mid \text{past}) = \frac{1}{s}\sum_{j=0}^{s-1} \frac{(j+1)(s-j)}{(s+1)^2(s+2)} = \frac{1}{6(s+1)}. \tag{B.19}$$

The expected excess loss at that observation is at least $1/[12(s+1)]$. Since $\sum_{s=1}^{n} 1/[12(s+1)] \geq (\log n)/24$ and $T/\kappa < n+1 \leq n^2$,

$$\mathbb{E}\,\mathrm{Reg}_T \geq \frac{\kappa}{24}\log n \geq \frac{\kappa}{48}\log(T/\kappa). \tag{B.20}$$

Regret against the best comparator is at least regret against the selected latent mean. The zero-padding rounds cannot reduce this lower bound. For each learner, its worst deterministic stream is at least as hard in expectation as the sampled stream distribution. The argument includes randomized learners.

The comparator cube lies in the radius-$\sqrt{\kappa}$ ball, so the lower bound applies to the game in Theorem 2. Conversely, VAW with $\lambda = B_v = 1$ and $\|u\|_2 \leq \sqrt{\kappa}$ gives

$$\mathrm{Reg}_T(\widehat{v}; u) \leq \frac{\kappa}{2} + \frac{\kappa}{2}\log(1 + T/\kappa). \tag{B.21}$$

For $T/\kappa \geq 2$, the two sides have the same order. This proves Theorem 2.

B.5 Proof of Proposition 1

Identify scalar-output states with column vectors and put $x_t = W_t^\top - (U^\star)^\top$, $a = 1-p$, and $\Sigma = \mathrm{diag}(a, p)$. The current key and noise are independent of x_t, so

$$\mathbb{E}[\ell_t(W_t) - \ell_t(U^\star)] = \tfrac{1}{2}\mathbb{E}[x_t^\top \Sigma x_t], \quad x_{t+1} = (I - \eta k_t k_t^\top)x_t + \eta \xi_t k_t. \tag{B.22}$$

Taking expectations yields $\mathbb{E}x_t = -(1-\eta p)^{t-1} e_2$. The bias alone gives

$$\mathbb{E}\,\mathrm{Reg}_T^{\mathrm{Delta}}(U^\star) \geq \frac{p}{2}\sum_{j=0}^{T-1}(1-\eta p)^{2j}. \tag{B.23}$$

If $\eta \leq T^{-1/2}$, then $\eta p \leq 1/T$. Since $(1-1/T)^{2(T-1)} \geq e^{-2} > 1/8$, the sum is at least $\sqrt{T}/16$.

Now let $\eta > T^{-1/2}$. Condition on the keys and expand the state into contributions from past independent noises. Cross terms from distinct noises vanish. Jensen's inequality and independence of future keys give

$$\mathbb{E}[x_{t,1}^2] \geq \eta^2 a \sum_{j=0}^{t-2}(1-\eta a)^{2j}. \tag{B.24}$$

Indeed, the expected product of future update matrices is a power of $I - \eta\Sigma$, and the first coordinate of every key is $\sqrt{a}$. For $n = t - 1 \geq \lfloor T/2 \rfloor$, the assumptions $T \geq 16$ and $a \geq 3/4$ give $\eta a n \geq 1$. The geometric sum is at least $(1-e^{-2})/(2\eta a)$. There are at least $T/2$ such rounds, and thus

$$\mathbb{E}\,\mathrm{Reg}_T^{\mathrm{Delta}}(U^\star) \geq \frac{(1-e^{-2})\eta a T}{8} \geq \frac{\sqrt{T}}{64}. \tag{B.25}$$

Finally, $\|U^\star\|_2 = 1$, $|v_t| \leq 2$, and the unit-key design has rank at most two. The VAW bound is therefore $1/2 + 2L_T \leq 1/2 + 4\log(1 + T/2)$ on every realization. The step η was fixed before the stream in both lower-bound cases.

C RETRIEVAL: PROOFS AND A HIGH-PROBABILITY EXTENSION

C.1 THE DETERMINISTIC TRANSFER AND NOISELESS CASE

Write $E_t = W_t - U^\star$ and $M_t = \|E_t\|^2_{A_{t-1}}$. Since $E_t q_t = E_t A_{t-1}^{1/2} A_{t-1}^{-1/2} q_t$, Cauchy–Schwarz gives

$$\|E_t q_t\|_2^2 \le M_t q_t^\top A_{t-1}^{-1} q_t. \tag{C.1}$$

This inequality alone does not bound the growth of M_t. The ridge objective gives the exact identity

$$F_{t-1}(U^\star) - F_{t-1}(W_t) = \tfrac{1}{2} M_t. \tag{C.2}$$

Because $F_{t-1}(W_t) \ge 0$, an always-valid bound is

$$M_t \le \lambda \|U^\star\|_{\mathrm{F}}^2 + \sum_{s<t} \|U^\star k_s - v_s\|_2^2. \tag{C.3}$$

It can grow with the cumulative target residual, which is why a stochastic noise model gives a different result from an arbitrary noisy stream.

If $v_s = U^\star k_s$, the normal equations simplify further:

$$C_{t-1} = U^\star (A_{t-1} - \lambda I), \qquad W_t - U^\star = -\lambda U^\star A_{t-1}^{-1}. \tag{C.4}$$

Hence

$$M_t = \lambda^2 \operatorname{Tr}(U^\star A_{t-1}^{-1} (U^\star)^\top) \le \lambda \|U^\star\|_{\mathrm{F}}^2. \tag{C.5}$$

Sum Eq. C.1 to prove Eq. 14. The target here is the generating map, not an arbitrary comparator with uncontrolled residuals.

For example, if $q_t = k_t$ and $\|k_t\|_2 \le 1$, then $h_t \le H = 1/\lambda$. Monotonicity of $h/\log(1+h)$ gives

$$Q_T = \sum_t h_t \le \frac{H}{\log(1+H)} L_T. \tag{C.6}$$

For fixed λ and bounded key rank, this is an explicit logarithmic bound on query leverage. The raw pre-update leverage sum is not equal to L_T.

C.2 EXPECTED RETRIEVAL ERROR UNDER MEAN-ZERO NOISE

Let $A = A_{t-1}$ and $Z_{t-1} = \sum_{s<t} \xi_s k_s^\top$. The state error is

$$W_t - U^\star = -\lambda U^\star A^{-1} + Z_{t-1} A^{-1}. \tag{C.7}$$

Under the conditional moment assumptions in Theorem 3, the bias–noise cross term and the cross terms between distinct noise times vanish. Thus

$$\begin{aligned} \mathbb{E}[\|(W_t - U^\star) q_t\|_2^2 \mid k_{1:T}, q_{1:T}] = \lambda^2 \|U^\star A^{-1} q_t\|_2^2 \\ + m\sigma^2 q_t^\top A^{-1} (A - \lambda I) A^{-1} q_t. \end{aligned} \tag{C.8}$$

The inequalities $\lambda A^{-2} \preceq A^{-1}$ and $A^{-1}(A - \lambda I)A^{-1} \preceq A^{-1}$ bound this expression by $(\lambda \|U^\star\|_{\mathrm{op}}^2 + m\sigma^2) q_t^\top A^{-1} q_t$. Summing proves Eq. 15, and averaging over random designs replaces Q_T by $\mathbb{E}Q_T$.

C.3 PREDICTABLE KEYS AND A HIGH-PROBABILITY GUARANTEE

The expectation result conditions on the full feature sequence. A different argument is needed when keys may depend on earlier noisy observations. The following result allows such dependence, provided the current key is determined before the current noise is observed.

Proposition 3 (Time-uniform control of the ridge error)**.** *Suppose $v_t = U^\star k_t + \xi_t$. Let $(\mathcal{F}_t)$ contain the data history, with k_t measurable with respect to $\mathcal{F}_{t-1}$ and ξ_t measurable with respect to $\mathcal{F}_t$. Assume, for each output coordinate i and every $a \in \mathbb{R}$,*

$$\mathbb{E}[\exp(a \xi_{t,i}) \mid \mathcal{F}_{t-1}] \le \exp(a^2 \sigma^2 / 2). \tag{C.9}$$

For $0 < \delta < 1$, with probability at least $1 - \delta$, simultaneously for all $t \ge 1$,

$$M_t \le 2\lambda \|U^\star\|_{\mathrm{F}}^2 + 2m\sigma^2 [L_{t-1} + 2\log(m/\delta)]. \tag{C.10}$$

On this event, for every query sequence,

$$\mathcal{E}_T \le Q_T \{2\lambda \|U^\star\|_{\mathrm{F}}^2 + 2m\sigma^2 [L_T + 2\log(m/\delta)]\}. \tag{C.11}$$

Proof. For coordinate i, set $z_{t-1,i} = \sum_{s<t} k_s \xi_{s,i}$. Apply the time-uniform self-normalized inequality of Abbasi-Yadkori et al. (2011, Theorem 1), with failure probability δ/m:

$$\|z_{t-1,i}\|^2_{A_{t-1}^{-1}} \le \sigma^2[L_{t-1} + 2\log(m/\delta)]. \tag{C.12}$$

A union bound over coordinates does not require independence between output coordinates. On this event, Eq. C.7 gives

$$M_t = \|(-\lambda U^\star + Z_{t-1})A_{t-1}^{-1/2}\|_{\mathrm{F}}^2 \tag{C.13}$$

$$\le 2\lambda^2 \operatorname{Tr}(U^\star A_{t-1}^{-1}(U^\star)^\top) + 2\sum_{i=1}^{m} \|z_{t-1,i}\|^2_{A_{t-1}^{-1}}. \tag{C.14}$$

Bound the first term by $2\lambda\|U^\star\|_{\mathrm{F}}^2$ and use Eq. C.1. Since $L_{t-1} \le L_T$, summation proves the retrieval bound. The event controls a matrix norm, so the final query inequality also allows data-dependent queries. □

The noise factor now contains L_T. It is not a horizon-independent constant. This result and the expected-risk bound use different assumptions and should not be interchanged.

C.4 Proof of Proposition 2

Sufficient coverage. Let $B_n = \lambda I + c^{-1}\sum_{s\le n} q_s q_s^\top$. The covariance condition gives $A_n \succeq B_n$. Define $x_t = c^{-1} q_t^\top B_{t-1}^{-1} q_t$. Since $\|q_t\|_2 \le 1$ and $c\lambda \ge 1$, we have $0 \le x_t \le 1$. Therefore

$$\begin{aligned} Q_T &\le c\sum_t x_t \le 2c\sum_t \log(1+x_t) \\ &= 2c\log\det\left(I + \frac{1}{c\lambda}\sum_t q_t q_t^\top\right) \le 2c\kappa\log\left(1 + \frac{T}{c\lambda\kappa}\right). \end{aligned} \tag{C.15}$$

The last inequality uses query rank at most κ and trace at most T. The assumption compares unregularized covariance sums, and replacing its right side by cA_n does not justify this same argument.

Unseen directions. If $q_\perp \neq 0$, choose a unit output vector $a \in \mathbb{R}^m$ and maps

$$U_\pm = \pm R a q_\perp^\top / \|q_\perp\|_2. \tag{C.16}$$

Both have Frobenius norm R and produce identical zero values on every past key. Their query answers are $\pm Ra\|q_\perp\|_2$. For any random learner output z, its average squared error on these two indistinguishable histories is

$$\frac{1}{2}\mathbb{E}[\|z - Ra\|q_\perp\|_2\|_2^2 + \|z + Ra\|q_\perp\|_2\|_2^2] = \mathbb{E}\|z\|_2^2 + R^2\|q_\perp\|_2^2. \tag{C.17}$$

At least one map has risk at least the second term. If $q_\perp = 0$, the claim follows from nonnegativity. This proves the one-query obstruction, without claiming a cumulative minimax lower bound for the whole regularized quantity Q_T.

D Approximate memories

Exact RLS supplies a reference trajectory. This appendix distinguishes a general sufficient loss certificate from direct analyses of particular cheaper recurrences. None of the statements says that an arbitrary approximate inverse inherits the exact-memory rate.

D.1 Signed excess loss and an exact subspace case

Let $W_t^\star$ be exact RLS and $\widetilde{W}_t$ any causal predictor on the same stream. Set $a_t = \ell_t(\widetilde{W}_t) - \ell_t(W_t^\star)$ and $s_t = \ell_t(W_t^\star) - \ell_t(W_{t+1}^\star)$, so $S_T = \sum_t s_t$. The exact decomposition

$$\mathrm{Reg}_T(\widetilde{W}; U) = \mathrm{Reg}_T(W^\star; U) + \sum_t a_t \tag{D.1}$$

proves Eq. 18 whenever $\sum_t a_t \le \alpha S_T + \varepsilon_T$. If the exact trajectory also satisfies $\|W_t^\star k_t - v_t\|_2 \le 2B$, then

$$\mathrm{Reg}_T(\widetilde{W}; U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + 4B^2(1+\alpha)L_T + \varepsilon_T. \tag{D.2}$$

To obtain a rate, B and α must be uniform in T, and ε_T must have the claimed growth. Checking finitely many streams certifies only those streams.

For a closed-form approximation $\widetilde{W}_t = C_{t-1}\widetilde{P}_{t-1}$, define $\Delta_t = (\widetilde{W}_t - W_t^\star)k_t$. Expanding the loss yields

$$a_t = \langle W_t^\star k_t - v_t, \Delta_t\rangle + \tfrac{1}{2}\|\Delta_t\|_2^2. \tag{D.3}$$

This explains why the certificate is signed. A bound on squared prediction discrepancies alone need not control the cross term at the same rate. For example, if $E_T = \sum_t \|\Delta_t\|_2^2$ and the exact residuals are bounded by $2B$, then only

$$\sum_t a_t \le 2B\sqrt{TE_T} + E_T/2 \tag{D.4}$$

follows directly. Even bounded E_T permits a square-root upper bound through this argument.

A lossless subspace implementation. Suppose a fixed subspace S containing every key is supplied in advance. If $\widetilde{P}_{t-1}$ agrees with A_{t-1}^{-1} on S, then $\Delta_t = 0$ and the certificate holds with $\alpha = \varepsilon_T = 0$. Maintaining exact ridge regression in an orthonormal basis of S gives such an implementation. Indeed, A_{t-1} is block diagonal on $S \oplus S^\perp$, and the two inverse actions agree on every update key. Arbitrary random sketches need not preserve this property.

D.2 RLS WITH PREDICTIONS FROZEN INSIDE A CHUNK

A chunked implementation may update its statistics only after several predictions. The following bound measures the resulting staleness directly.

Proposition 4 (Frozen-chunk stability)**.** *Partition time into consecutive chunks of lengths $b_j \le b$. Let $K_j \in \mathbb{R}^{b_j \times d}$ contain chunk keys as rows and $V_j \in \mathbb{R}^{m \times b_j}$ contain values as columns. Define chunk-level statistics*

$$A_j = \lambda I + \sum_{i \le j} K_i^\top K_i, \qquad C_j = \sum_{i \le j} V_i K_i, \qquad W_j = C_{j-1}A_{j-1}^{-1}. \tag{D.5}$$

Predict every pair in chunk j using W_j. Put $E_j = W_j K_j^\top - V_j$ and $H_j = K_j A_{j-1}^{-1} K_j^\top$. Then

$$\mathrm{Reg}_T^{\mathrm{chunk}}(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + \frac{1}{2}\sum_j \mathrm{Tr}\{E_j[I - (I+H_j)^{-2}]E_j^\top\}. \tag{D.6}$$

If every column of E_j has norm at most $2\widetilde{B}$, the stability term is at most $4b\widetilde{B}^2 L_T$. If each H_j is diagonal, the chunk predictions equal sequential RLS predictions.

Proof. For chunk loss $f_j(W) = \|WK_j^\top - V_j\|_{\mathrm{F}}^2/2$, the states are exact regularized leaders. Hence regret is at most the ridge penalty plus $\sum_j [f_j(W_j) - f_j(W_{j+1})]$. Woodbury gives

$$W_{j+1} = W_j - E_j K_j A_j^{-1}, \qquad W_{j+1}K_j^\top - V_j = E_j(I+H_j)^{-1}. \tag{D.7}$$

This proves the exact stability expression. For $x \ge 0$, $[1-(1+x)^{-2}]/2 \le \log(1+x)$. Diagonalizing H_j and using $\|E_j\|_{\mathrm{F}}^2 \le 4b\widetilde{B}^2$ gives a bound of $4b\widetilde{B}^2 \log\det(I+H_j)$. The determinant identity telescopes over chunks to L_T.

If H_j is diagonal, distinct chunk keys are orthogonal in the chunk-start inverse geometry. Updates from preceding keys in that chunk leave the next key's prediction unchanged. Induction gives equality with sequential predictions. □

The staleness cost can be substantial. For one chunk of b scalar pairs $k = v = 1$ with $\lambda = 1$, frozen predictions incur regret $b/2$ against $U = 1$, whereas sequential RLS incurs $\frac{1}{2}\sum_{t=1}^{b} t^{-2} \le \pi^2/12$. The factor b in the direct bound does not itself establish a signed-loss certificate with $\alpha = b - 1$ relative to sequential RLS.

D.3 A SPECIFIED FREQUENT DIRECTIONS FEEDBACK UPDATE

A sketch can reduce covariance storage, but discarded curvature must be accounted for. We use a conservatively inflated Frequent Directions (FD) construction, following the sketching approach of Luo et al. (2016) and the FD invariants of Ghashami et al. (2016).

Proposition 5 (FD curvature with residual feedback)**.** *Let $2 \le \ell \le d$. Initialize an FD sketch $B_0 = 0$. At each step suppose its discarded covariance D_t and threshold δ_t obey*

$$B_t^\top B_t = B_{t-1}^\top B_{t-1} + k_t k_t^\top - D_t, \quad 0 \preceq D_t \preceq \delta_t I, \quad \mathrm{Tr}(D_t) \ge \ell\delta_t. \tag{D.8}$$

Between compressions, $D_t = 0$ and $\delta_t = 0$. Set $\rho_t = \sum_{s\le t} \delta_s$ and $\widetilde{A}_t = (\lambda + \rho_t)I + B_t^\top B_t$. On a closed convex domain $\mathcal{W}$ containing zero and having Frobenius diameter D, initialize $W_1 = 0$ and use

$$W_{t+1} = \Pi_{\mathcal{W}}^{\widetilde{A}_t}\left(W_t - (W_t k_t - v_t)k_t^\top \widetilde{A}_t^{-1}\right), \tag{D.9}$$

where the projection uses $\|\cdot\|_{\widetilde{A}_t}$. If $\|W_t k_t - v_t\|_2 \le 2B$, then every $U \in \mathcal{W}$ satisfies

$$\mathrm{Reg}_T(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 + 2B^2 \log\det(\lambda^{-1}\widetilde{A}_T) + \frac{D^2\rho_T}{2}, \qquad \rho_T \le \frac{\Omega_r}{\ell - r}, \quad 0 \le r < \ell, \tag{D.10}$$

where Ω_r is the sum of covariance eigenvalues after the largest r of $\sum_t k_t k_t^\top$.

Proof. Write $X_t = W_t - U$ and $g_t = (W_t k_t - v_t)k_t^\top$. Metric projection implies

$$2\langle g_t, X_t\rangle \le \|X_t\|_{\widetilde{A}_t}^2 - \|X_{t+1}\|_{\widetilde{A}_t}^2 + \|g_t\|_{\widetilde{A}_t^{-1}}^2. \tag{D.11}$$

Square loss has the exact identity $\ell_t(W_t) - \ell_t(U) = \langle g_t, X_t\rangle - \|X_t k_t\|_2^2/2$. Summing and dropping the nonpositive terminal term gives

$$\begin{aligned}\mathrm{Reg}_T(U) \le \frac{\lambda}{2}\|U\|_{\mathrm{F}}^2 &+ \frac{1}{2}\sum_t \|W_t k_t - v_t\|_2^2 k_t^\top \widetilde{A}_t^{-1} k_t \\ &+ \frac{1}{2}\sum_t \|W_t - U\|_{\widetilde{A}_t - \widetilde{A}_{t-1} - k_t k_t^\top}^2.\end{aligned} \tag{D.12}$$

The last metric increment is $\delta_t I - D_t$, which lies between zero and $\delta_t I$. Thus the last sum is at most $D^2\rho_T/2$. Also $\widetilde{A}_t \succeq \widetilde{A}_{t-1} + k_t k_t^\top$. If $\widetilde{h}_t = k_t^\top \widetilde{A}_{t-1}^{-1} k_t$, then

$$k_t^\top \widetilde{A}_t^{-1} k_t \le \frac{\widetilde{h}_t}{1 + \widetilde{h}_t} \le \log(1 + \widetilde{h}_t) \le \log\det \widetilde{A}_t - \log\det \widetilde{A}_{t-1}. \tag{D.13}$$

The residual assumption proves the regret bound.

For the tail estimate, write $G = \sum_t k_t k_t^\top$ and $E = G - B_T^\top B_T = \sum_t D_t$. The invariants imply $0 \preceq E \preceq \rho_T I$ and $\mathrm{Tr}(E) \ge \ell\rho_T$. For the projector P_r onto the top r eigenvectors of G,

$$\mathrm{Tr}(B_T^\top B_T) \ge \mathrm{Tr}(P_r B_T^\top B_T) \ge \mathrm{Tr}(P_r G) - r\rho_T. \tag{D.14}$$

Therefore $\ell\rho_T \le \mathrm{Tr}(E) \le \Omega_r + r\rho_T$, proving Eq. D.10. Finally,

$$A_T \preceq \widetilde{A}_T = A_T + \rho_T I - E \preceq A_T + \rho_T I, \qquad \log\det(\lambda^{-1}\widetilde{A}_T) \le L_T + d\rho_T/\lambda. \tag{D.15}$$

□

Meaning and cost. When the design has exact rank $r < \ell$, $\Omega_r = 0$ and the sketch loses no curvature. Nonzero spectral tails contribute both the metric-inflation cost and the diameter cost, and no dimension-free tail guarantee is asserted. One implementation buffers at most 2ℓ rows, subtracts the squared ℓth singular value on compression, and retains positive singular values. Compression costs $O(d\ell^2)$ and occurs after at least ℓ additional keys, giving amortized $O(d\ell)$ sketch work. Inverse application costs $O(d\ell + \ell^2)$ and the memory update costs $O(md)$. Storage is $O(d\ell + md)$. These costs exclude the domain-dependent metric projection, which is part of the specified update.

D.4 WHY A FIXED RELATIVE INVERSE ERROR IS INSUFFICIENT

A small but persistent multiplicative error can prevent the prediction from reaching a constant target.

Proposition 6 (Linear regret from a closed-form approximation)**.** *For scalar $k_t = v_t = 1$, $\lambda > 0$, and fixed $\epsilon > 0$, let*

$$\widetilde{W}_t = C_{t-1}[(1+\epsilon)A_{t-1}]^{-1}. \tag{D.16}$$

Then $\mathrm{Reg}_T(\widetilde{W}; 1) \geq \frac{T}{2}[\epsilon/(1+\epsilon)]^2$. *With time-varying $0 \leq \epsilon_t \leq 1$, the lower bound becomes $\frac{1}{8}\sum_t \epsilon_t^2$.*

Proof. Here $C_{t-1} = t-1$ and $A_{t-1} = \lambda + t - 1$. Therefore

$$1 - \widetilde{W}_t = \frac{\lambda + \epsilon(\lambda + t - 1)}{(1+\epsilon)(\lambda + t - 1)} \geq \frac{\epsilon}{1+\epsilon}. \tag{D.17}$$

The comparator has zero loss. Squaring and summing proves the fixed-error claim. For $\epsilon_t \leq 1$, $\epsilon_t/(1+\epsilon_t) \geq \epsilon_t/2$, giving the second claim. □

Conversely, the same residual is at most $\lambda/(\lambda + t - 1) + \epsilon_t$, so

$$\mathrm{Reg}_T(\widetilde{W}; 1) \leq \sum_t \frac{\lambda^2}{(\lambda + t - 1)^2} + \sum_t \epsilon_t^2. \tag{D.18}$$

For fixed λ, choosing $\epsilon_t = t^{-1/2}$ gives logarithmic regret on this example. Thus the example imposes a cumulative squared-error requirement, not a universal pointwise decay rule. It concerns the displayed closed-form predictor, not the feedback update Eq. D.9.

E MEMORIES FOR CHANGING TARGET MAPS

For a moving comparator, write $P_T = \sum_{t=2}^{T} \|U_t - U_{t-1}\|_{\mathrm{F}}$. This quantity measures total target movement. We first state the standard moving-comparator gradient bound, then analyze actual decay, reset, and discount recurrences. These statements do not transfer automatically to every architecture with a gate.

E.1 A PROJECTED FIRST-ORDER BOUND

Proposition 7 (Dynamic regret for projected gradient descent)**.** *Let $\mathcal{W}$ be a nonempty closed convex domain of Frobenius diameter D. For convex losses, $W_1 \in \mathcal{W}$, and subgradients $g_t \in \partial \ell_t(W_t)$ with $\|g_t\|_{\mathrm{F}} \leq G$, use $W_{t+1} = \Pi_{\mathcal{W}}(W_t - \eta g_t)$ with $\eta > 0$. Every path $U_{1:T} \in \mathcal{W}^T$ satisfies*

$$\mathrm{DynReg}_T(U_{1:T}) \leq \frac{D^2 + 2DP_T}{2\eta} + \frac{\eta G^2 T}{2}. \tag{E.1}$$

For $D, G > 0$ and a known budget $P \geq P_T$, the choice $\eta = \sqrt{(D^2 + 2DP)/(G^2 T)}$ gives $G\sqrt{T(D^2 + 2DP)}$.

Proof. Projection and convexity give

$$\ell_t(W_t) - \ell_t(U_t) \leq \frac{\|W_t - U_t\|_{\mathrm{F}}^2 - \|W_{t+1} - U_t\|_{\mathrm{F}}^2}{2\eta} + \frac{\eta}{2}\|g_t\|_{\mathrm{F}}^2. \tag{E.2}$$

For $X, U, U' \in \mathcal{W}$, the change in squared distance is at most $2D\|U' - U\|_{\mathrm{F}}$. Applying this between consecutive comparators and summing proves Eq. E.1. Optimizing the two step-size terms gives the displayed choice. If $D = 0$, learner and comparator coincide. If $G = 0$, convexity makes every regret term nonpositive. □

This is the moving-comparator form of the first-order argument (Zinkevich, 2003). It requires the projected update and bounded domain. It is not, by itself, a guarantee for an arbitrary unprojected gated memory.

E.2 Scalar-gated Delta and its bias

Decay helps remove old information, but it also shrinks the target fit. We study the scalar recurrence

$$W_{t+1} = \gamma W_t - \eta(W_t k_t - v_t)k_t^\top. \tag{E.3}$$

The residual is evaluated before decay. This order is part of the model, and gates placed inside the residual define a different recurrence. This distinction matters when relating the abstraction to gated sequence architectures (Yang et al., 2025).

Proposition 8 (A dynamic bound with decay bias)**.** *Assume $\|k_t\|_2 \le 1$, $\|v_t\|_2 \le B_v$, $0 < \eta \le \gamma < 1$, and $W_1 = 0$. Define*

$$\mu = (1-\gamma)/\eta, \quad R_W = B_v/\mu, \quad D = R_W + R, \quad G_\mu = (1+\mu)R_W + B_v. \tag{E.4}$$

The iterates satisfy $\|W_t\|_{\mathrm{F}} \le R_W$. For every comparator path with $\|U_t\|_{\mathrm{F}} \le R$,

$$\mathrm{DynReg}_T(U_{1:T}) \le \frac{D^2 + 2DP_T}{2\eta} + \frac{\eta G_\mu^2 T}{2} + \frac{(1-\gamma)R^2 T}{2\eta}. \tag{E.5}$$

On the scalar static stream $k_t = v_t = 1$, the same update has $\mathrm{Reg}_T(1) \ge \frac{T}{2}[\mu/(1+\mu)]^2$.

Proof. The eigenvalues of $\gamma I - \eta k_t k_t^\top$ lie in $[0, \gamma]$, so

$$\|W_{t+1}\|_{\mathrm{F}} \le \gamma\|W_t\|_{\mathrm{F}} + \eta B_v \le \eta B_v \sum_{j=0}^{t-1} \gamma^j \le B_v/\mu. \tag{E.6}$$

The update is gradient descent on $f_t(W) = \ell_t(W) + \mu\|W\|_{\mathrm{F}}^2/2$. Its gradient is bounded by G_μ. Every learner–comparator distance is at most D, so the squared-distance proof from Appendix E.1 applies to f_t. Finally,

$$\ell_t(W_t) - \ell_t(U_t) = f_t(W_t) - f_t(U_t) + \tfrac{\mu}{2}(\|U_t\|_{\mathrm{F}}^2 - \|W_t\|_{\mathrm{F}}^2), \tag{E.7}$$

which gives the last, bias term.

For $k_t = v_t = 1$, the recursion is $W_{t+1} = (\gamma - \eta)W_t + \eta$. Starting at zero, it stays below its fixed point $1/(1+\mu)$. Each loss is therefore at least $[\mu/(1+\mu)]^2/2$, while $U = 1$ has zero loss. □

This lower bound shows a real static failure, not just a loose term in an upper bound. On the same constant stream, ungated Delta with $0 < \eta \le 1$ has total regret at most $1/[2\eta(2-\eta)]$.

E.3 Resetting ridge memory at changes

Proposition 9 (Known-boundary resets)**.** *Suppose keys are unit-bounded, the key design has rank at most κ, and the comparator is constant on N consecutive segments, with norm at most R. Reset the statistics to $(A, C) = (\lambda I, 0)$ at each boundary. Then*

$$\mathrm{Reg}_T^{\mathrm{switch}} \le \frac{N\lambda R^2}{2} + c_* N\kappa \log\left(1 + \frac{T}{N\lambda\kappa}\right). \tag{E.8}$$

For ordinary RLS, $c_* = 4B^2$ requires every reset trajectory to have residuals bounded by $2B$. For VAW, $c_* = B_v^2/2$ requires only values bounded by B_v.

Proof. Let the segment lengths be $n_1, \ldots, n_N$, summing to T. Apply the respective static bound to each independently initialized segment. The total is at most

$$\frac{N\lambda R^2}{2} + c_*\kappa \sum_{j=1}^{N} \log(1 + n_j/(\lambda\kappa)). \tag{E.9}$$

Concavity of the logarithm proves Eq. E.8. □

Each reset pays a fresh estimation cost, but no longer forces the memory to fit all earlier targets. The bound assumes the boundary information is supplied. The oracle-reset experiments evaluate this setting, not change-point detection.

E.4 A CAUSAL MULTI-START METHOD WITHOUT BOUNDARY INFORMATION

Multiple memories started at different times can avoid committing to a single reset time. The following is a specified history-expert construction, in the spirit of Hazan and Seshadhri (2009).

Assume $T \geq 2$ and $B_v > 0$. Consider all dyadic intervals $I = \{j2^r + 1, \ldots, (j+1)2^r\} \subseteq \{1, \ldots, T\}$. There are fewer than $2T$ such intervals and at most $H = 1 + \lceil \log_2 T \rceil$ active at any round. Start a fresh VAW copy at each interval's left endpoint and keep it active only on that interval. Clip its prediction $z_{t,I}$ to the radius-B_v ball. Clipping cannot increase its squared loss against a value in the same ball.

Give each interval prior weight $w_I = 1/(2T)$ and set $\alpha = 1/(4B_v^2)$. At round t, normalize active weights into $p_{t,I}$ and predict $\widehat{v}_t = \sum_{I \ni t} p_{t,I} z_{t,I}$. After observing v_t, update active weights by

$$w_I \leftarrow w_I \exp\left\{\alpha\left[\tfrac{1}{2}\|\widehat{v}_t - v_t\|_2^2 - \tfrac{1}{2}\|z_{t,I} - v_t\|_2^2\right]\right\}. \tag{E.10}$$

Unit intervals make the active set nonempty.

Proposition 10 (Switching regret without known boundaries)**.** *Under the bounded-value, bounded-key, rank, and comparator assumptions of Proposition 9, the preceding causal method satisfies*

$$\mathrm{Reg}_T^{\mathrm{switch}} \leq 2HN\left[\frac{\lambda R^2}{2} + \frac{B_v^2 \kappa}{2}\log\left(1 + \frac{T}{\lambda\kappa}\right) + 4B_v^2 \log(2T)\right]. \tag{E.11}$$

At most H independently initialized VAW copies are simultaneously active.

Proof. For $f(z) = \|z - v_t\|_2^2/2$ on the radius-B_v ball, $\nabla^2 f = I \succeq \alpha \nabla f \nabla f^\top$ since $\|z - v_t\|_2 \leq 2B_v$. Thus $\exp(-\alpha f)$ is concave. The weighted-average prediction makes the total active weight nonincreasing under Eq. E.10. Inactive weights are unchanged. The sum of all interval weights starts below one and stays at most one. Consequently, for every completed interval I,

$$\sum_{t \in I}[f_t(\widehat{v}_t) - f_t(z_{t,I})] \leq \alpha^{-1}\log(2T) = 4B_v^2 \log(2T). \tag{E.12}$$

Every contiguous comparator segment partitions into at most $2H$ dyadic intervals: its maximal contained dyadic intervals include at most two of each length. Apply the last inequality and the VAW bound on each interval against that segment's comparator. There are at most $2HN$ intervals, giving Eq. E.11. Dormant weights are fixed priors and expired weights need not be stored, so only active copies require storage. □

The result includes the logarithmic overhead of searching over start times. The supplied experiments do not evaluate this wrapper.

E.5 PROOF OF THE SINGLE-JUMP COMPARISON

For the stream in Eq. 19, the moving comparator has zero loss. Ordinary RLS predicts zero in the first segment. For $n < t \leq 2n$,

$$W_t = J\frac{t - n - 1}{\lambda + t - 1}, \qquad |W_t - J| = |J|\frac{\lambda + n}{\lambda + t - 1} \geq |J|/2. \tag{E.13}$$

There are $n = T/2$ such rounds, giving dynamic regret at least $J^2T/16$. Delta also starts the second segment at zero, but its residuals are $-J(1-\eta)^j$ for $0 \leq j < n$. Therefore

$$\mathrm{DynReg}_T^{\mathrm{Delta}} \leq \frac{J^2}{2}\sum_{j=0}^{\infty}(1-\eta)^{2j} = \frac{J^2}{2\eta(2-\eta)}. \tag{E.14}$$

The comparison demonstrates the difference between static prediction and tracking a changing target. It does not contradict a static-regret guarantee for ridge memory.

E.6 DISCOUNTED RLS WITH CONTINUED KEY EXCITATION

Discounting gives recent pairs more influence while retaining a ridge penalty. Its benefit depends on both target movement and how often relevant directions are observed. The result below keeps that directional condition explicit. Related discounted online-regression methods have separate analyses (Yuan and Lamperski, 2020; Jacobsen and Cutkosky, 2024).

Proposition 11 (Discounted RLS)**.** *Fix deterministic keys and a deterministic comparator path with $\|k_t\|_2 \le 1$ and $\|U_t\|_{\mathrm{F}} \le R$. Suppose $v_t = U_t k_t + \xi_t$, where the noises are independent, mean zero, and have covariance $\sigma^2 I_m$. Let $T \ge 2$, $\gamma = 1-\delta$, and $0 < \delta \le 1/2$. Use*

$$\begin{aligned} A_t^\gamma &= \gamma A_{t-1}^\gamma + k_t k_t^\top + (1-\gamma)\lambda I, \quad & A_0^\gamma &= \lambda I, \\ C_t^\gamma &= \gamma C_{t-1}^\gamma + v_t k_t^\top, & C_0^\gamma &= 0, \\ W_t^\gamma &= C_{t-1}^\gamma (A_{t-1}^\gamma)^{-1}. \end{aligned} \tag{E.15}$$

Let S contain all keys, with dimension κ, and set $N_n^\gamma = (1-\gamma^n)/(1-\gamma)$. Assume that, for some $\nu > 0$,

$$A_n^\gamma|_S \succeq \frac{\lambda + \nu N_n^\gamma}{2} I_S, \qquad 0 \le n < T. \tag{E.16}$$

Then $\mathbb{E}\,\mathrm{DynReg}_T(U_{1:T}) \le B_T(\delta)$, where

$$\begin{aligned} B_T(\delta) = 4R^2\left(1 + \frac{2\lambda}{\nu}\right) + \frac{m\sigma^2}{\lambda} + \frac{2m\sigma^2}{\nu}\log\left(1 + \frac{\nu}{2\lambda\delta}\right) \\ + \frac{2m\sigma^2\delta T}{\nu} + \frac{16\lambda^2 R^2 \delta^2 T}{\nu^2} + \frac{4RP_T}{\nu\delta}. \end{aligned} \tag{E.17}$$

Meaning. A shorter effective history increases the variance and persistent ridge-shrinkage terms, but reduces the influence of old target maps. The last term records the cost of target movement. The parameter ν records how well the discounted observations cover key directions, and it cannot be hidden in a rate claim if it deteriorates with T.

Proof. Before round t, unroll Eq. E.15 with weights $w_{t,s} = \gamma^{t-1-s}$:

$$A := A_{t-1}^\gamma = \lambda I + \sum_{s<t} w_{t,s} k_s k_s^\top, \qquad C_{t-1}^\gamma = \sum_{s<t} w_{t,s} v_s k_s^\top. \tag{E.18}$$

Put $a_t = \lambda + \nu N_{t-1}^\gamma$. Condition Eq. E.16 gives $\|A^{-1}|_S\|_{\mathrm{op}} \le 2/a_t$. The error relative to the current noiseless target splits into ridge bias, drift, and past noise:

$$\begin{aligned} (W_t^\gamma - U_t)k_t = &-\lambda U_t A^{-1} k_t + \sum_{s<t} w_{t,s}(U_s - U_t)k_s k_s^\top A^{-1} k_t \\ &+ \sum_{s<t} w_{t,s}\xi_s k_s^\top A^{-1} k_t =: b_t + d_t + z_t. \end{aligned} \tag{E.19}$$

The design and path are deterministic, so $\mathbb{E}z_t = 0$. Since $w_{t,s}^2 \le w_{t,s}$,

$$\mathbb{E}\|z_t\|_2^2 \le \frac{2m\sigma^2}{a_t}, \qquad \|b_t\|_2^2 \le \frac{4\lambda^2 R^2}{a_t^2}. \tag{E.20}$$

Weighted Cauchy–Schwarz bounds the drift contribution by

$$\|d_t\|_2^2 \le \frac{2}{a_t}\sum_{s<t} w_{t,s}\|U_s - U_t\|_{\mathrm{F}}^2. \tag{E.21}$$

Let $\Delta_r = \|U_r - U_{r-1}\|_{\mathrm{F}}$. Since $\|U_s - U_t\|_{\mathrm{F}} \le 2R$, its square is at most $2R\sum_{r=s+1}^t \Delta_r$. Also $\sum_{s<r} w_{t,s} = \gamma^{t-r} N_{r-1}^\gamma \le \gamma^{t-r} N_{t-1}^\gamma$. It follows that

$$\|d_t\|_2^2 \le \frac{4R}{\nu}\sum_{r=2}^{t} \gamma^{t-r}\Delta_r. \tag{E.22}$$

At $t = 1$ the sum is empty and the inequality still holds.

Independence of current noise makes expected dynamic excess loss equal to $\mathbb{E}\|(W_t^\gamma - U_t)k_t\|_2^2/2$. Use $\|b_t + d_t\|_2^2 \le 2\|b_t\|_2^2 + 2\|d_t\|_2^2$ and sum the geometric drift convolution to get

$$\mathbb{E}\,\mathrm{DynReg}_T \le 4\lambda^2 R^2 \sum_{t=1}^{T} a_t^{-2} + m\sigma^2 \sum_{t=1}^{T} a_t^{-1} + \frac{4RP_T}{\nu\delta}. \tag{E.23}$$

For $n \ge 0$, $N_n^\gamma \ge \frac{1}{2}\min\{n, \delta^{-1}\}$. Splitting the sums at $\lfloor \delta^{-1} \rfloor$ and comparing decreasing summands with integrals gives

$$\sum_{t=1}^{T} a_t^{-1} \le \frac{1}{\lambda} + \frac{2}{\nu}\log\left(1 + \frac{\nu}{2\lambda\delta}\right) + \frac{2\delta T}{\nu}, \tag{E.24}$$

$$\sum_{t=1}^{T} a_t^{-2} \le \frac{1}{\lambda^2} + \frac{2}{\lambda\nu} + \frac{4\delta^2 T}{\nu^2}. \tag{E.25}$$

Substitution proves Eq. E.17. □

An excitation example. Cycling periodically through κ orthonormal unit keys satisfies Eq. E.16 with $\nu = 1/\kappa$ whenever $\lambda \ge 1$ and $\gamma^{\kappa-1} \ge 1/2$. To verify this, put $q = \lfloor n/\kappa \rfloor$. Each coordinate has weighted covariance at least $\gamma^{\kappa-1}\sum_{j=0}^{q-1}\gamma^{j\kappa}$, while $N_n^\gamma \le \kappa(1 + \sum_{j=0}^{q-1}\gamma^{j\kappa})$. The smallest covariance entry is therefore at least $N_n^\gamma/(2\kappa) - 1/2$ (nonnegativity supplies this inequality also when $N_n^\gamma < \kappa$). Adding $\lambda \ge 1$ proves the condition.

Choosing a discount. For a known path budget $P \ge P_T$, minimize the explicit bound with P replacing P_T over discounts satisfying Eq. E.16 with the same ν. When $\sigma > 0$ and the ridge-bias term is negligible, balancing the linear variance and path terms suggests $\delta = \sqrt{2RP/(m\sigma^2 T)}$, clipped to an admissible interval. This is a conditional tuning approximation, not a universally optimal discount. The $\delta^2 T$ term remains even for a static noiseless target because a finite effective history retains ridge shrinkage.

F Full Details of the WikiText Experiment

Here, we provide the full experimental setup for the results reported in Section 7.1.

Implementation and compute. All online memory experiments are implemented in Python using NumPy. Unless otherwise noted, all generated keys, values, model states, and online updates use 32-bit floating-point arithmetic. Offline ridge solutions and log-determinant diagnostics are computed in 64-bit floating point for improved numerical stability.

For all methods, prediction is strictly causal: at round t, the algorithm first predicts from the current state W_t, incurs

$$\ell_t(W_t) = \frac{1}{2}\|W_t k_t - v_t\|_2^2,$$

and only then updates its state using (k_t, v_t). The initial fast-weight matrix is zero. Ridge-based methods use $\lambda = 1.0$ throughout the reported core experiments.

Compared online memories. The fixed-feature WikiText experiment compares four update rules:

1. *Linear attention*, which accumulates $B_t = B_{t-1} + v_t k_t^\top$ and normalizes the prediction by the accumulated scalar key norm;
2. *Delta-SGD*, with $W_{t+1} = W_t - \eta(W_t k_t - v_t)k_t^\top$, $\eta = 0.35$;
3. *Scalar gated Delta*, implemented as $W_{t+1} = \gamma W_t - \eta(W_t k_t - v_t)k_t^\top$, $\eta = 0.35$, $\gamma = 0.99$;
4. *Exact RLS*, initialized with $W_1 = 0$ and $P_0 = \lambda^{-1} I$, and updated using the Sherman–Morrison recursion $h_t = k_t^\top P_{t-1} k_t$, $z_t = \frac{P_{t-1}k_t}{1+h_t}$, $W_{t+1} = W_t - (W_t k_t - v_t)z_t^\top$, $P_t = P_{t-1} - \frac{P_{t-1}k_t k_t^\top P_{t-1}}{1+k_t^\top P_{t-1}k_t}$.

The controlled-drift experiment additionally includes oracle-reset Delta, coordinate-adaptive Delta variants, static RLS, discounted RLS, and oracle-reset RLS. Oracle-reset methods are supplied with the true abrupt change points and therefore serve as diagnostic upper controls rather than deployable baselines.

WikiText fixed-feature KV experiment. We use the WikiText train, validation, and test splits. Text is tokenized by lowercasing and separating alphabetic strings, numeric strings, and individual non-whitespace punctuation symbols. The vocabulary consists of the 2048 most frequent tokens in the training split. Every token outside this vocabulary is mapped to one additional unknown-token index. Evaluation streams are constructed from the concatenation of the validation and test splits.

For each random seed, every vocabulary item is assigned two independent fixed random Gaussian embeddings: a key embedding in $\mathbb{R}^{32}$ and a value embedding in $\mathbb{R}^{16}$. Each row is normalized according to $x \leftarrow \frac{x}{\max(1,\|x\|_2)+10^{-8}}$, so generated embeddings, whose Gaussian norms typically exceed one, are effectively unit-normalized. The same embedding tables are reused for all tasks and context lengths associated with a given seed.

We use eight random seeds, $\{11, 17, 23, 42, 73, 101, 137, 199\}$. For every seed, we evaluate two associative-prediction tasks and three stream lengths, $T \in \{1024, 2048, 4096\}$. This produces $8 \times 2 \times 3 = 48$ paired WikiText streams.

The two tasks are:

- *Identity binding:* the key and value are generated from the same token, $k_t = E_k[x_t],\ v_t = E_v[x_t]$. This gives a comparatively stable associative-recall problem.
- *Next-token binding:* the current token supplies the key and the following token supplies the value, $k_t = E_k[x_t],\ v_t = E_v[x_{t+1}]$. This introduces ambiguity because the same token may be followed by different next tokens.

For each seed–task–length configuration, a contiguous span is sampled uniformly from the available evaluation token stream, subject to containing the required $T+1$ tokens. The sampled start position is recorded so that each run can be reconstructed exactly. All compared algorithms within that run receive exactly the same key–value sequence.

Offline comparator and static regret. For every stream, the static comparator is the offline ridge-regression solution

$$U_{\mathrm{ridge}} = \arg\min_U \left\{ \frac{\lambda}{2}\|U\|_F^2 + \frac{1}{2}\sum_{t=1}^{T} \|Uk_t - v_t\|_2^2 \right\},$$

which is computed as $U_{\mathrm{ridge}} = C_T A_T^{-1}$, with $A_T = \lambda I + \sum_{t=1}^{T} k_t k_t^\top,\ C_T = \sum_{t=1}^{T} v_t k_t^\top$.

The reported cumulative static regret is

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \ell_t(W_t) - \sum_{t=1}^{T} \ell_t(U_{\mathrm{ridge}}).$$

We additionally report static regret per token, $\mathrm{Reg}_T\,/T$, and mean squared KV error,

$$\text{KV-MSE} = \frac{1}{T}\sum_{t=1}^{T} \|W_t k_t - v_t\|_2^2 = \frac{2}{T}\sum_{t=1}^{T} \ell_t(W_t).$$

The ridge log-determinant diagnostic is computed as

$$L_T = \log\det\left(I + \lambda^{-1}\sum_{t=1}^{T} k_t k_t^\top\right).$$

We also record the correlation between the running static-regret curve and both L_t and $\sqrt{t}$ as diagnostic quantities.

Controlled drifting associative maps. The synthetic drift experiment uses $T = 2048, d = 32, m = 16$, with the same eight seeds as the WikiText experiment. At each round, the key is sampled

independently from a standard Gaussian distribution in $\mathbb{R}^{32}$ and normalized to unit norm. Values are generated from the time-varying linear associative map $v_t = U_t k_t + \xi_t$, where $U_t \in \mathbb{R}^{16\times 32}$ is known to the experiment and $\xi_t \sim \mathcal{N}(0, 0.01^2 I)$. The resulting values are clipped coordinate-wise to $[-3, 3]$.

We test two drift regimes, two fractions of drifting coordinates, and two drift magnitudes. The drifting-coordinate fractions are $\rho \in \{0.25, 1.0\}$, corresponding to 8 active coordinates or all 32 coordinates, respectively. In the implementation, the active coordinates are the first $\max(1, \lfloor 32\rho \rfloor)$ input dimensions.

Abrupt drift. The stream is divided into eight equal blocks. The initial map is sampled with entries $(U_1)_{ij} \sim \mathcal{N}(0, 0.35^2)$. At each boundary between consecutive blocks, the coefficients associated with the active input coordinates are replaced by fresh Gaussian values. Two jump scales are evaluated, $\sigma_{\text{jump}} \in \{0.7, 1.2\}$. The remaining, inactive coordinates are left unchanged. This produces seven known change points over the 2048-round stream. These change points are supplied only to the oracle-reset baselines.

Gradual drift. The gradual map begins from the same type of Gaussian initialization. At every subsequent time step, $U_t = 0.998\, U_{t-1} + \Delta_t$, where Δ_t is zero outside the active coordinates. On each active coordinate, we have $(\Delta_t)_{ij} \sim \mathcal{N}\left(0, \frac{\delta^2}{d_{\text{active}}}\right)$. We test drift rates $\delta \in \{0.02, 0.05\}$. The normalization by $\sqrt{d_{\text{active}}}$ keeps the scale of the total perturbation more comparable across sparse- and full-coordinate drift.

Combining eight seeds, two drift regimes, two coordinate fractions, and two regime-specific drift levels produces $8 \times 2 \times 2 \times 2 = 64$, controlled drifting streams.

Hyperparameters for the drift experiment. Delta-SGD uses $\eta = 0.35$ and all ridge methods use $\lambda = 1.0$. Scalar gated Delta uses $\gamma = 0.995$. For discounted RLS, the discount factor is

$$\gamma_{\text{RLS}} = \begin{cases} 0.985, & \text{abrupt drift,} \\ 0.995, & \text{gradual drift.} \end{cases}$$

Discounted RLS maintains

$$A_t = \gamma_{\text{RLS}} A_{t-1} + k_t k_t^\top + (1 - \gamma_{\text{RLS}})\lambda I,$$

$$C_t = \gamma_{\text{RLS}} C_{t-1} + v_t k_t^\top, \qquad W_{t+1} = C_t A_t^{-1}.$$

For coordinate-adaptive Delta variants, the adaptive-gradient step size is 0.08. The coordinate-wise decay is 0.995 on inactive coordinates and is set to 0.96 for active coordinates under abrupt drift and 0.985 for active coordinates under gradual drift.

Dynamic-regret evaluation. Because the synthetic generator exposes the true map U_t, the dynamic comparator is exactly the generating sequence. Its cumulative loss is $\sum_{t=1}^{T} \frac{1}{2}\|U_t k_t - v_t\|_2^2$. Therefore, dynamic regret is

$$\text{DynReg}_T = \sum_{t=1}^{T} \ell_t(W_t) - \sum_{t=1}^{T} \ell_t(U_t).$$

We also record the comparator path variation $P_T = \sum_{t=2}^{T} \|U_t - U_{t-1}\|_F$, dynamic regret per token, and the ordinary online prediction MSE.

Pairing and reported comparisons. Every algorithm is evaluated on exactly the same generated stream within a run. Thus, we construct paired algorithm comparisons at the run level. For the WikiText experiment, a run is identified by seed, task, context length, and sampled start location. For the drift experiment, a run is identified by seed, drift regime, coordinate sparsity, and drift level. We report the mean paired difference and the number of runs won by each algorithm.

G DETAILS OF THE LONGBENCHV2 EXPERIMENT

The previous experiments study associative prediction on synthetic or simplified KV streams. Here, we ask whether the same differences appear on features taken from a pretrained Transformer. We use LongBench-v2 (Bai et al., 2025) to obtain long contexts together with questions and answer options. We freeze Qwen2.5-7B-Instruct (Bai et al., 2023) and extract its attention features. Then, we replace the attention computation by different online associative memories while keeping the extracted features fixed.

It is worth reminding that this experiment is a diagnostic of the memory update only. We do not train the backbone, and we do not evaluate LongBench answer accuracy. Therefore, these results should not be interpreted as an end-to-end comparison between Transformer and recurrent architectures.

Data and split. We examine 503 examples and 462 distinct contexts. We keep examples with the same context in the same partition to avoid sharing the same document between development and test sets. Using seed 17, we create a development split stratified by tokenizer length. This leads to 101 development examples and 402 held-out examples, corresponding to 94 and 368 distinct context groups, respectively. We limit the full model input to $32,768$ tokens. When an example is too long, we truncate the middle of the context while preserving the complete question and answer options. The final held-out set is therefore dominated by retained contexts between 16k and 32k tokens. Middle truncation affects 309 of the 402 held-out examples. The original context length distribution and the retained lengths are reported in Table 6.

Context tokens	Original dev	Original test	Retained dev	Retained test
$<$ 16k	6	23	6	23
16–32k	17	70	95	379
32–64k	14	54	0	0
64–128k	24	96	0	0
$\geq$ 128k	40	159	0	0

Table 6: Length counts under the 32,768-token total-input cap. There are 94 development and 368 test context groups. Middle truncation affects 78 development and 309 test examples; the complete own and foreign suffixes fit within the cap.

Native features. We conduct the experiment on a frozen `Qwen/Qwen2.5-7B-Instruct` checkpoint with 28 layers, 28 query heads, and four KV heads, with head dimension 128 (Bai et al., 2023). The $32,768$-token cap uses the checkpoint's native RoPE configuration without context extension. Layers are numbered from 1. The primary layer for the main experiment is 14, and ablations use 7 and 21 on a preselected subset. Let z_t denote the actual, normalized input to the attention projections. Including the checkpoint's projection biases, native features before RoPE are

$$\begin{aligned} k_{t,g} &= \frac{W_{K,g} z_t + b_{K,g}}{\|W_{K,g} z_t + b_{K,g}\|_2 + \epsilon}, \qquad v_{t,g} = W_{V,g} z_t + b_{V,g}, \\ q_{t,a} &= \frac{W_{Q,a} z_t + b_{Q,a}}{\|W_{Q,a} z_t + b_{Q,a}\|_2 + \epsilon}, \qquad \epsilon = 10^{-8}. \end{aligned} \tag{G.1}$$

All seven query heads associated with each KV head are evaluated. Values retain their native scale. Extraction uses 4096-token chunks, preserving causal KV caches and absolute positions across chunks. Computation stops after the highest measured layer, preserving features at all measured layers. Every retained context token updates the memory. The question and options are processed after the context but do not update the associative state. All methods share the cached features.

What does the memory reconstruct? This is important, as the memory and native attention do not always have access to the same information.

- **Native-attention target:** at each question or option position, we first compute the output of the model's original causal softmax attention. We use the post-RoPE attention output before output projection. The native target can attend to the context (stored by memory), earlier question/option tokens (not stored), and current token (not stored). On the contrary, our associative memory contains only the context. At a query position, it predicts $W_{T+1,g(a)} q_{t,a}$. Thus, the native-attention reconstruct experiment asks whether a context-

only recurrent memory can approximate the output of the original causal attention mechanism, even though native attention also sees the question/option prefix and the current token.

- **Context-restricted target:** we also construct a second target in which softmax attention is restricted to the stored context $o_{t,a}^{\mathrm{ctx}} = \mathrm{Attn}(q_{t,a}, K_{\mathrm{context}}, V_{\mathrm{context}})$. Now, softmax attention and the associative memory use the same information: context-restricted attention and associative memory both see only the context. Therefore, this is the cleaner diagnostic of retrieval from the learned memory. The native target instead measures approximation to the actual pretrained attention layer.

Reconstruction metric. We compare the memory preidction with the corresponding softmax attention target using $R^2_{\mathrm{mean}} = 1 - \frac{\sum_{t,a} \|\hat{o}_{t,a} - o_{t,a}\|_2^2}{\sum_{t,a} \|\bar{v}_{t,g(a)} - o_{t,a}\|_2^2}$, where $\hat{v}_{t,g}$ is the running mean of the values available to the memory. Thus, the denominator represents a simple running-mean baseline. Higher values are better, and the score can be negative if a method performs worse than this baseline. Importantly, this quantity is not the usual centered regression R^2. Native softmax itself has zero reconstruction error and is used only as a reference, not as a learned competitor. Scores are first aggregated over positions and query heads within each KV group and then averaged over the four KV groups for each example.

Methods and tuning. We compare the following online memories: scaled additive linear attention, constant-step Delta, data-dependent Delta, scalar-gated Delta, exact RLS, VAW-shrunk RLS, discounted RLS, chunked RLS with $b \in \{16, 64\}$, and residual-feedback FD with sketch size 64. Scaled linear attention uses $W_t = cC_{t-1}/t$, with development-tuned gain c. Data-dependent Delta uses the scalar step $\beta_t = \eta/(1 + k_t^\top D_{t-1}^{-1} k_t)$, where $D_{t-1} = \lambda I + \mathrm{diag}(\sum_{s<t} k_s k_s^\top)$. This additional diagonal state is charged in the cost report. VAW uses the key gate for KV regret and the query gate $1/(1 + q^\top A^{-1} q)$ for retrieval. We treat this as an experimental variant: the VAW KV-regret result does not imply a retrieval guarantee for that extension. Similarly, the executed FD recurrence is unprojected, so we do not claim that it directly satisfies the projected domain assumptions of the theoretical FD result.

Regarding hyperparameter selection, All hyperparameters are selected using the development set only. Each method family receives eight candidate configurations. The search was expanded from four to eight settings after development-only boundary checks and before any held-out result was computed. We select the setting with the highest mean native-attention reconstruction score on the development set. Deterministic tie-breaking follows the stored grid order. After selection, the chosen configuration is fixed for all test endpoints and layer analyses. We do not use held-out scores for tuning, and we do not clip residuals. The selected configurations are reported in Table 8.

Secondary recall. On a fixed length-stratified subset, a separate memory binds the native key of x_t to the backbone's input embedding of x_{t+1}, with value dimension 3584. Before observing that binding, it queries earlier bindings using the current key. Recall@1 uses cosine nearest-neighbor search over distinct next-token embeddings observed in the past, and absent targets count as misses. Repeated and previously unseen input-token identities are reported separately. These lexical groups do not certify geometric coverage. This diagnostic compares RLS with the development-selected Delta rule.

Coverage and controls. We record $q^\top A_T^{-1} q$ and summarize it by its per-example mean over unpooled queries. Quartiles of this mean are formed within retained-length bins. The partial Spearman correlation uses per-example squared error and leverage after regressing their ranks on $\log T$. Foreign-question queries use a fixed cyclic donor from a different context in the same split. Their leverage and method differences are measured: foreign queries need not be outside the key span. A separate constructed unseen-coordinate control appends a coordinate that is zero in every update key and nonzero in the query, with an explicit two-point target. Pooled query means are renormalized before leverage evaluation, and the primary endpoint uses unpooled queries.

Theory diagnostics. The comparison of unregularized KV regret uses a common $U_\lambda = C_T A_T^{-1}$ and L_T, with λ selected for exact RLS on development. We also report the exact-RLS bound ratio, effective stability, and each method's maximum played residual:

$$\frac{\mathrm{Reg}_T(U_\lambda)}{L_T}, \qquad \frac{\mathrm{Reg}_T^{\mathrm{RLS}}(U_\lambda)}{\lambda \|U_\lambda\|_F^2/2 + S_T}, \qquad B_{\mathrm{eff}}^2 = \frac{S_T}{4L_T}, \qquad \max_t \|\widehat{v}_t - v_t\|_2. \tag{G.2}$$

Here, $\widehat{v}_t$ is the played KV prediction, including the VAW gate when applicable. Zero denominators are flagged. A bound ratio exceeding one beyond the logged floating-point tolerance triggers an audit. At the same ridge parameter, chunked and FD approximations receive the signed audit

$$\widehat{\alpha} = \frac{\sum_t [\ell_t(W_t^{\mathrm{approx}}) - \ell_t(W_t^{\mathrm{RLS}})]}{S_T}. \tag{G.3}$$

Its nonnegative part is an empirical certificate parameter, not a uniform theorem. For chunks, let E_j and H_j be as in Proposition 4, and define the directly audited bound by

$$D_T^{\mathrm{chunk}} = \sum_j b_j \max_{i \le b_j} \|(E_j)_{:,i}\|_2^2 \log\det(I + H_j), \qquad \frac{\mathrm{Reg}_T^{\mathrm{chunk}}(U_\lambda)}{\lambda \|U_\lambda\|_F^2/2 + D_T^{\mathrm{chunk}}}. \tag{G.4}$$

The proof of Proposition 4 bounds its exact block-stability term by D_T^{chunk}. We also report the realized chunk/RLS regret ratio where the denominator is positive. The factor b does not bound that ratio of realized regrets. Stability and layer analyses are secondary outcomes.

Statistical protocol. As several examples can share the same underlying context, we do not treat all examples as independent observations. For each example, we first average across the four KV groups. We then estimate paired confidence intervals using $10{,}000$ bootstrap resamples of context groups. This preserves dependence between examples that share a context. We report both per-example and context-level sign tests, but use the context-level tests for inferential conclusions. The RLS-versus-Delta native-attention comparison is specified as the primary comparison. Secondary comparisons use Holm correction. Dropout is disabled, and all algorithms are requested to run deterministically. Seed 17 fixes both the data split and the bootstrap procedure. The reported intervals describe uncertainty within this fixed-feature benchmark. They do not represent variation across independently trained language models.

Component	Evaluation setting
Backbone	`Qwen/Qwen2.5-7B-Instruct`; cached revision recorded
Features	Native affine Q/K/V before RoPE; normalized Q/K; four KV groups, all 28 query heads
Layers	Primary 14; ablations 7 and 21 on a fixed subset
Seed	17; grouped example replication
Length policy	32,768 total tokens; middle-truncated context; suffix retained
Extraction / updates	4096-token chunks with persistent cache; update stride one
Split	Approximately 100 development examples; remaining examples held out; context grouping
Tuning	Eight configurations per family on development only
Clipping	None
Approximations	FD size 64; chunked RLS with $b = 16, 64$
Primary metric	Native attention reconstruction at question/option positions relative to running mean
Controls	Foreign questions; constructed unseen direction; length-stratified leverage
Uncertainty	Context-group bootstrap; sign tests; Holm-adjusted secondary comparisons

Table 7: Frozen pretrained attention-feature evaluation protocol.

Cost and reporting. Table 7 fixes the protocol before evaluation. We provide state size and leading per-token FLOPs separately from extraction, and measure synchronized throughput of cache features on 8192 tokens after a 1024-token warm-up, using the median of three repetitions. Chunked and FD rows place quality beside cost. Verified numerical results, bin counts, selected settings, and color figures are generated from saved artifacts below.

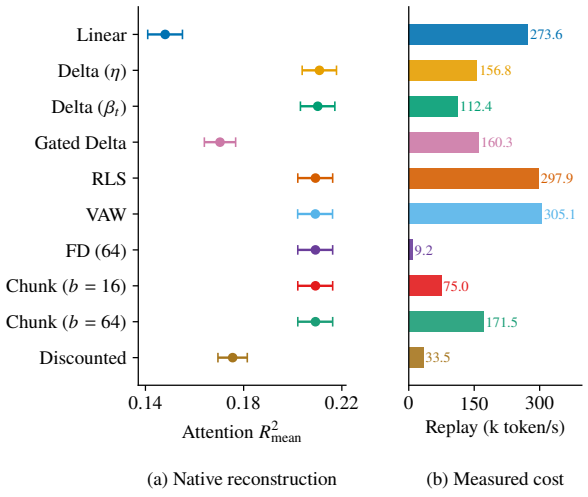
Held-out results. Exact RLS has mean native reconstruction score 0.2093, compared with 0.2109 for Delta (η). The pre-specified paired difference is -0.0016 (95% context-cluster bootstrap interval [-0.0026, -0.0006]); the two-sided context sign-test value is 1.87×10^{-6}. The interval supports a higher mean score for the selected Delta on this fixed-feature endpoint. RLS wins in 154 of 402 examples and 138 of 368 context groups, with 0 group ties. On the context-restricted target, the paired RLS–Delta difference is 0.0116 (95% interval [0.0092, 0.0144]; Holm $p = 2.09 \times 10^{-22}$). Therefore, the method ranking changes with the reconstruction target. Table 9 and Figure 2 place reconstruction beside measured replay cost.

Family	Selected setting
Linear attention	gain=3.0, λ=1.0
Delta (η)	η=0.001
Delta (β_t)	η=0.001, λ=1.0
Gated Delta	η=0.01, γ=0.9999
Exact RLS	λ=1000.0
VAW-shrunk RLS	λ=1000.0
FD ($\ell = 64$)	ℓ=64, λ=1000.0
Chunked RLS ($b = 16$)	b=16, λ=1000.0
Chunked RLS ($b = 64$)	b=64, λ=1000.0
Discounted RLS	γ=0.9999, λ=100.0

Table 8: Settings selected on 101 development examples, then fixed for all held-out layers and endpoints. Each family received eight configurations; each chunk size had its own eight-ridge grid.

Method	Native R^2 [95% CI]	Cosine	Matched R^2	k token/s	Core (KiB)
Linear attention	0.148 [0.141, 0.155]	0.276	0.428	273.55	512
Delta (η)	0.211 [0.204, 0.218]	0.377	0.491	156.83	512
Delta (β_t)	0.210 [0.203, 0.217]	0.377	0.492	112.41	516
Gated Delta	0.170 [0.164, 0.177]	0.338	0.325	160.31	512
Exact RLS	0.209 [0.202, 0.216]	0.386	0.503	297.92	1024
VAW-shrunk RLS	0.209 [0.202, 0.216]	0.386	0.503	305.11	1024
FD ($\ell = 64$)	0.209 [0.202, 0.216]	0.386	0.503	9.17	764
Chunked RLS ($b = 16$)	0.209 [0.202, 0.216]	0.386	0.503	75.00	1024
Chunked RLS ($b = 64$)	0.209 [0.202, 0.216]	0.386	0.503	171.51	1024
Discounted RLS	0.176 [0.170, 0.181]	0.348	0.304	33.49	512
Native softmax reference	1.000	1.000	1.000	–	$4 \cdot 2Td \cdot 2/1024$

Table 9: Held-out reconstruction and reference-replay cost on 402 examples. Scores average the four KV groups within each example. Intervals resample context groups. Core is the analytical post-update learner state across four KV groups, excluding buffers, while the softmax reference uses BF16 KV storage. All learner rows time GPU feedback, and FD compression uses the CPU LAPACK package. Throughput includes KV-loss and comparator-statistic accumulation, but excludes extraction and answer-query scoring. Native and context-restricted softmax targets differ.



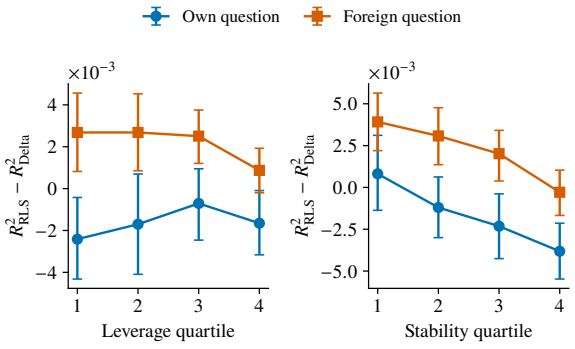
(a) Native reconstruction (b) Measured cost

Figure 2: Native attention reconstruction and synchronized cached-feature throughput. Horizontal intervals in (a) are pointwise 95% context-cluster bootstrap intervals. Panel (b) measures the reference implementation, including diagnostics. Every learner row uses GPU feedback, with CPU compression for FD, this is not a comparison of optimized production kernels.

Coverage and stability. Across examples, the partial Spearman correlation between RLS squared reconstruction error and query leverage, after removing a linear effect of $\log T$ from their ranks, is 0.184. Foreign suffixes have higher mean leverage in 93.3% of paired examples. For foreign queries, the paired RLS–Delta difference is 0.0022 (95% interval [0.0013, 0.0031]). These observations do not support using leverage alone to rank the methods on native reconstruction. These are observational diagnostics, and foreign queries do not enforce an unseen direction. The separate zero-coordinate control has squared error one for the indistinguishable two-point targets, matching the coverage obstruction. Figure 3 groups comparisons within retained-length bins. For exact RLS, median B_{eff}^2 is 12.334, the maximum residual norm is 20.459, and the largest exact-bound ratio is 0.998. Table 10 reports the signed approximation audits. These realized values do not establish a uniform residual bound or a projected-domain guarantee for unprojected FD. Where exact-RLS regret is positive, the realized chunk/RLS regret ratios are $b = 16$: median 1.012, maximum 1.091; $b = 64$: median 1.027, maximum 1.193. The factor b in the block stability bound is a conservative coefficient, not an upper bound on these realized ratios.

Method	Mean Reg/L_T	Max residual	Median $\widehat{\alpha}$	Max $\widehat{\alpha}$	Max bound ratio
Linear attention	2.73×10^4	20.57	–	–	–
Delta (η)	2710.07	20.43	–	–	–
Delta (β_t)	2881.15	20.45	–	–	–
Gated Delta	-1.00×10^4	19.84	–	–	–
Exact RLS	4896.45	20.46	–	–	0.998
VAW-shrunk RLS	4898.40	20.46	–	–	–
FD ($\ell = 64$)	4897.74	20.46	0.013	0.200	–
Chunked RLS ($b = 16$)	4949.32	20.47	1.078	1.925	0.930
Chunked RLS ($b = 64$)	5019.12	20.47	2.487	4.488	0.737
Discounted RLS	-6496.43	20.10	–	–	–

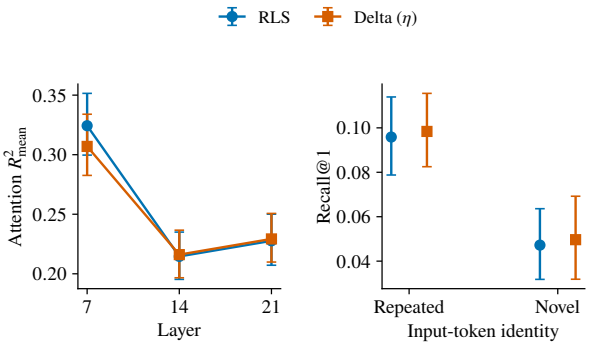
Table 10: Online KV diagnostics across test examples and KV groups. The regret column uses the same offline comparator and log-determinant denominator for every method, with the development-selected exact-RLS ridge. Negative regrets are permitted because these unregularized losses are compared with a ridge solution. The last column uses each method's own regularizer: exact stability for RLS and the direct block bound for chunks. Certificates compare with exact RLS at the approximation's ridge and remain signed.



(a) Coverage (b) Stability

Figure 3: Paired reconstruction differences between RLS and Delta, grouped by leverage and effective stability. Quartiles are formed within retained-length bins; intervals resample context groups. Own and foreign questions use the same target context.

Layer and recall diagnostics. The preselected subset contains 48 examples. Fig 4 evaluates layers 7, 14, and 21 with settings selected at layer 14, and reports causal next-token embedding recall separately for repeated and novel token identities. For repeated inputs, RLS recall@1 is 0.096 versus 0.098, and for novel inputs it is 0.047 versus 0.050. Both paired recall intervals include zero. Targets absent from the past candidate set count as misses. These lexical categories do not certify key-span coverage.



(a) Layer ablation (b) Causal recall

Figure 4: Fixed-setting layer ablation and causal embedding recall on the preselected test subset. Points average examples, with context-cluster bootstrap intervals. Recall compares the native key with previous next-token bindings, a different query construction from the primary attention-reconstruction endpoint.

Cost and scope. Table 11 gives analytical costs, while Table 9 gives measured throughput. RLS, VAW and chunk replay batch exact causal prefix statistics, with the prefix arrays charged as workspace. The reference implementation uses double precision for the memory audits and single precision for fixed-ridge discounted solves, and cached features use half precision. The main timing rows use GPU feedback for every method, with CPU compression for FD, and the execution report additionally gives entirely CPU FD timing. Resident audit tensors and peak GPU workspace are recorded separately in the execution report; CPU scratch peak is not measured, so the compact-state formulas should not be read as measured peak allocation. Chunked RLS has exactly the same final full-context ridge state as RLS at a common regularizer. Therefore, its answer-position reconstruction matches, while its online losses can differ. The sampled context queries also fall on common chunk boundaries, so those samples do not measure retrieval between chunk updates. An additional explicit Wq retrieval costs $2md$ leading FLOPs, and the VAW query gate adds an $O(d^2)$ quadratic form. The 32k cap leaves longer-context behavior untested. These results measure frozen-feature update rules and do not establish LongBench task performance, robustness across trained checkpoints, or the noiseless assumptions of the retrieval propositions.

Rule	State (scalars per KV group)	Leading FLOPs per update/query
Linear / Delta	md	$4md$
Data-dependent Delta	$md + d$	$4md + O(d)$
Gated Delta	md	$5md$
RLS / VAW	$md + d^2$	$4md + 4d^2 + O(d + m)$
Fixed-ridge discounted RLS	$md + d^2$	$O(d^3 + md)$
Chunked RLS	$md + d^2 + b(d + m)$	$O(md + d^2 + bd + b^2)$
FD feedback	$md + (2\ell - 1)d + 1$	$O(md + d\ell + \ell^2)$ amortized
Softmax, one query	$2Td$	$4Td + O(T)$

Table 11: Analytical learner costs, excluding diagnostic accumulators. The first seven rows include one pre-update KV prediction; softmax costs are per query head. Fixed-ridge discounting replenishes λI, so a rank-one inverse update alone does not implement it. FD costs describe a compact low-rank implementation. The measured audit backend uses dense block solves and CPU compression. This table includes chunk and FD key buffers, while Table 9 lists the post-update core state.

Table 12 demonstrates variation across the four KV groups (it does not treat heads as independent seeds), while Table 13 reports the remaining paired comparisons. Table 13's pointwise confidence intervals are not simultaneous; Holm correction applies to the context sign-test values. The intervals concern mean differences, whereas the sign tests assess the balance of wins and losses, so they test different null hypotheses.

KV group	RLS R^2	Delta R^2	Difference [95% CI]
1	0.191	0.190	0.0012 [0.0005, 0.0020]
2	0.252	0.248	0.0034 [0.0025, 0.0042]
3	0.162	0.172	-0.0102 [-0.0123, -0.0081]
4	0.233	0.233	-0.0009 [-0.0022, 0.0006]

Table 12: Native reconstruction by KV group at the primary layer. Each group aggregates its seven query heads. Intervals resample the same context groups; these four rows are dependent measurements, and the primary test averages them within example.

Comparison / endpoint	Paired difference [95% CI]	Holm p
Linear attention	0.0612 [0.0590, 0.0634]	1.84×10^{-107}
Delta (β_t)	-0.0009 [-0.0019, 0.0001]	0.020
Gated Delta	0.0389 [0.0333, 0.0448]	1.36×10^{-44}
VAW-shrunk RLS	0.0000557 [0.0000516, 0.0000595]	3.93×10^{-81}
FD ($\ell = 64$)	0.00000363 [0.00000023, 0.00000706]	0.783
Chunked RLS ($b = 16$)	0.0000 [0.0000, 0.0000]	1.000
Chunked RLS ($b = 64$)	0.0000 [0.0000, 0.0000]	1.000
Discounted RLS	0.0337 [0.0282, 0.0395]	1.89×10^{-35}
Context-restricted retrieval	0.0116 [0.0092, 0.0144]	2.09×10^{-22}
Context-token retrieval	-0.0291 [-0.0302, -0.0280]	5.32×10^{-110}
Foreign-question retrieval	0.0022 [0.0013, 0.0031]	7.71×10^{-10}
Own minus foreign gain	-0.0038 [-0.0048, -0.0029]	4.94×10^{-12}
Layer 7 retrieval	0.0174 [0.0141, 0.0208]	1.44×10^{-10}
Layer 21 retrieval	-0.0016 [-0.0043, 0.0011]	1.000
Recall: repeated input	-0.0025 [-0.0071, 0.0020]	1.000
Recall: novel input	-0.0024 [-0.0112, 0.0053]	1.000

Table 13: Secondary contrasts. Method-name rows compare RLS with that method on the primary endpoint. Endpoint, layer, and recall rows compare RLS with the selected Delta. The own-minus-foreign row compares their paired gains. All secondary context sign tests shown here share one Holm correction family.