Omen, Instant-CLS & REM-Engram: Dual-Timescale Subconscious Ghost Locus and Bio-Inspired Synaptic Memory for Evolving Codebases

Millisecond Closed-Form Hippocampal Writes, Active-Anchor Orthogonal Unlearning, and Nightly Bio-REM Consolidation in a 7.98 MB Repo-Engram

Ryan Joshua Z. Gapac
ORCID: 0009-0004-2591-0219 · DOI: 10.5281/zenodo.24607074
mail@ryanjoshuagapac.com · https://ryanjoshuagapac.com · November 28, 2026
System-One & Bio-Memory Series (Paper #2) · Extends Warden-S1 (DOI: 10.5281/zenodo.24494405)

Abstract. Autonomous coding agents routinely break hidden cross-file invariants: an edit to file `A` silently invalidates an assumption in an untouched file `B` whose path never appears in the diff. Existing guardrails fail across timescales. Static System-One verifiers such as **Warden-S1** (`DOI: 10.5281/zenodo.24494405`) remain amnesiac across days; append-only retrieval-augmented memory (RAG) bloats wake prompts with stale logs that keep firing after refactors; and nightly-only consolidation leaves a 14-hour daytime blindness window whenever new rules or refactors land mid-day. We introduce **Omen**, **Instant-CLS**, and **REM-Engram**, the second paper in the System-One & Bio-Memory Research Series. At wake time, **Omen** pairs `Qwen3-0.6B` with four zero-decode, zero-extra-prompt-token parallel heads (`Salience`, `Diff Locus`, `Ghost Locus`, and calibrated `Alarm`) ranking untouched files in one prefill pass (`67.0 ms` p50, `126` prompt tokens). During the day, **Instant-CLS** adds a closed-form `fp32` Hippocampal Fast-Matrix over frozen representations: exact Sherman-Morrison writes (`1.04 ms` p50, `0` backprop) raise mid-day new-rule Ghost R@1 and R@3 from `0.0%` to `100.0%`, while active-anchor orthogonal QR null-space projection (`2.71 ms` p50) cuts post-refactor Stale FPR from `71.7%` to `1.7%` (`p < 0.0001`) with `98.9%` prior retention. Overnight, **Bio-REM** compiles repository history into a `7.98 MB` hot-swappable Repo-Engram via Counterfactual Twin Contrast, hindsight-to-foresight reverse-KL distillation, and Stage-2 CLS replay with NPO pruning and hippocampal reset. Across `3` seeds on **GhostLocus-Bench v2** (`60` families, `5` ecosystems, `M = 80` files, `260` held-out episodes, `0` lexical leaks), nightly Bio-REM reaches `97.5%` Ghost R@1, `99.2%` R@3, `0.9063` AUROC, `14.2%` Safe-Twin FPR, and `0.6%` Stale FPR (`+71.1 pp` R@1, `-37.2 pp` Stale FPR vs. RAG, `p < 0.0001`), while full 24-hour **Instant-CLS** achieves `97.8%` post-sleep R@3 and `0.9051` AUROC after hippocampal reset.

1 Introduction

Software repositories are coupled dynamical systems whose hardest bugs rarely live inside the edited hunk. A developer or coding agent modifies a request router, schema compiler, dependency cache, or validation path in file `A`, and the failure manifests in an untouched file `B` (the **ghost locus**) that implicitly depended on `A`'s prior invariant. Because `B`'s path and symbols rarely appear in the raw `git diff`, neither static linters nor context-blind diff reviewers can warn the engineer before commit.

Our prior work, **Warden-S1** [1] (`DOI: 10.5281/zenodo.24494405`), established that a sub-1 GiB local System-One verifier with zero-decode parallel heads can verify long-horizon agent trajectories (`4K` to `128K` tokens) faster and more reliably than cloud LLM judges [2], [3]. However, **Warden-S1** solves **within-session** verification: its weights remain static across days, so it cannot remember repository-specific cross-file contracts learned from yesterday's incidents, nor forget invariants invalidated by today's architectural refactors.

The dominant industry patch for cross-session memory is **append-only retrieval-augmented memory** (RAG) [4], [5]: incident postmortems and rules are appended to an external store and injected into every prompt. Append-only memory fails along three coupled dimensions in evolving repositories:

1. **Wake latency and context tax.** Retrieving top-`k` incident logs multiplies every pre-commit prompt (`872.4` tokens vs. `125.6` raw diff tokens in our benchmark, `+746.8` extra tokens), inflating wake latency by `1.72x` even at `k = 8`.
2. **Lexical blind spots.** When a new diff triggers an invariant without echoing the vocabulary of a prior postmortem, BM25 retrieval returns noisy neighbors, stalling top-1 ghost prediction at `26.4%` across `M = 80` files.
3. **Stale-memory poisoning.** Appending a Day-4 refactor note does not erase Day-1 incident logs from the index. Retrieved contradictions keep firing false alarms (`37.8%` to `57.2%` Stale FPR) on clean post-refactor edits.

Moving memory into weights via nightly sleep consolidation eliminates prompt bloat and stale false alarms overnight, yet introduces a fourth operational bottleneck: a **14-hour daytime blindness window**. Whenever a developer commits a new invariant at `10:00` or lands a refactor at `15:00` that retires old rules, nightly-only weights stay blind (`0.0%` mid-day new-rule R@1, `71.7%` stale FPR) until midnight. Conversely, running online gradient descent (`Online SGD/LoRA`) per commit costs

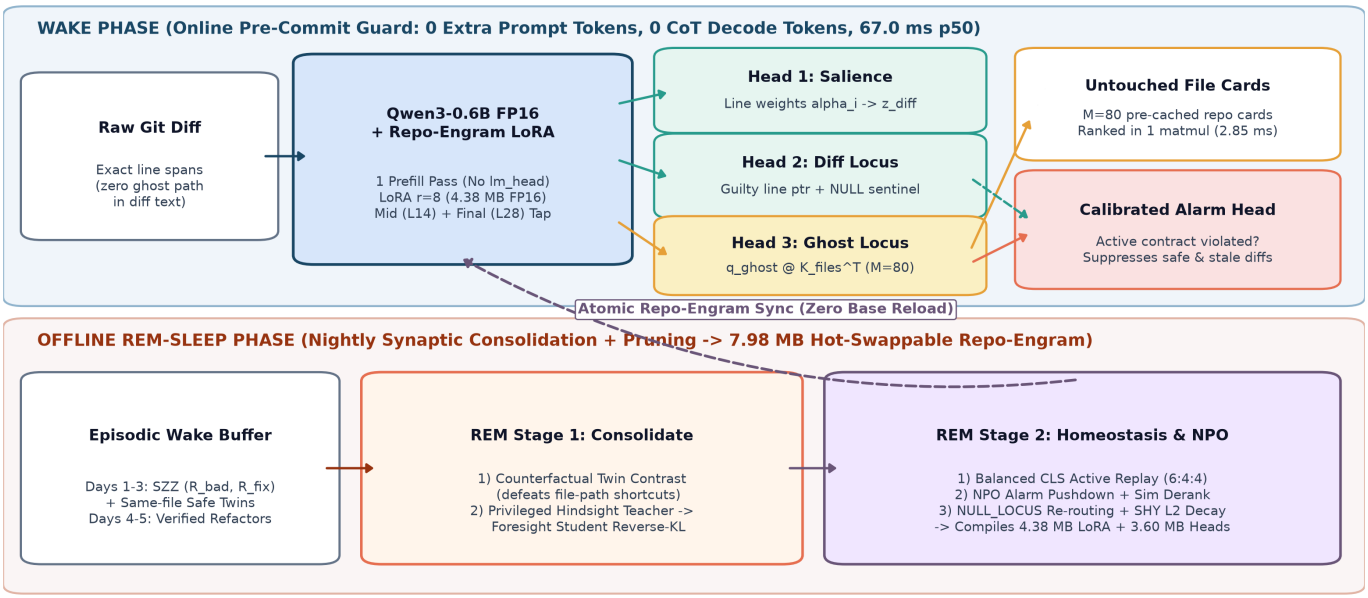


Figure 1: Dual-phase **Omen & REM-Engram** foundation. During Wake, one prefill pass over the raw git diff (`0` extra prompt tokens, `0` CoT decode tokens, `67.0 ms` p50) drives four parallel subconscious heads (`Salience` , `Diff Locus` , `Ghost Locus` , `Alarm`) against `M = 80` pre-cached untouched file cards (`2.85 ms` head overhead). During the day, **Instant-CLS** layers a `0.633 MB` closed-form `fp32` Hippocampal Fast-Matrix (`1.04 ms` write, `2.71 ms` orthogonal erase) atop frozen `h_final` states; overnight, Bio-REM Stage 1 and Stage 2 consolidate into a `7.98 MB` hot-swappable Repo-Engram and reset `W_hipp = 0` .

hundreds of milliseconds and triggers catastrophic collapse (`1.7%` prior retention, `0.0%` active TPR).

Biological brains solve all four failure modes through **Complementary Learning Systems** (CLS) [6] paired with **Synaptic Homeostasis** (SHY) [7]: a fast hippocampal store encodes and edits daytime episodes without distorting neocortical weights, while offline sleep interleaves replay and prunes obsolete synapses before resetting fast traces. We translate this full dual-timescale architecture into **Omen**, **Instant-CLS**, and **REM-Engram**, making five contributions:

1. **Parallel Subconscious Ghost Locus (Omen).** A zero-decode, zero-extra-prompt-token architecture on `Qwen3-0.6B` [8] tapping mid-layer (`L14`) and final-layer (`L28`) states in one prefill pass (`67.0 ms` p50) to score line salience, guilty diff lines (`NULL_LOCUS` sentinel), untouched ghost files (`M = 80` candidates in `2.85 ms`), and calibrated violation probability.
2. **Instant-CLS Daytime Hippocampal Fast-Matrix.** Exact closed-form `fp32` Sherman-Morrison [9] rank-1 updates (`1.04 ms` p50, `0` backprop, `0` extra prompt tokens) lifting mid-day new-rule Ghost R@1/R@3 from `0.0%` to `100.0%` , paired with active-anchor orthogonal QR null-space surgical unlearning (`2.71 ms` p50) and a strictly non-positive orthogonal-complement veto dropping post-refactor Stale FPR from `71.7%` to `1.7%` while preserving `98.9%` prior retention.
3. **Bio-REM Nightly Neocortical Consolidation (REM-Engram).** A two-stage offline sleep protocol compiling multi-day history into a `7.98 MB` hot-swappable Repo-Engram (`4.38 MB` rank-8 LoRA [10] + `3.60 MB` heads) via Counterfactual Twin Contrast, Privileged Hindsight-to-Foresight reverse-KL distillation, Stage-2 `6:4:4` CLS replay with Negative Preference Optimization (NPO) [11] and SHY decay, followed by exact hippocampal reset (`W_hipp = 0`).
4. **GhostLocus-Bench v2 & 24-Hour Repo Stream.** A zero-leakage multi-day cross-file invariant benchmark spanning `60` families (`40` active, `20` superseded) across `5` ecosystems (`FastAPI` , `Hono` , `Pydantic` , `Vite` , `Zod`) and `M = 80` candidate files (`420` train, `260` held-out Day-6 test episodes, `0` lexical leaks), evaluated in both static nightly consolidation and a chronological 24-hour continuous stream.
5. **3-Seed Multi-Arm Statistical & Algebraic Proof.** Controlled `3` -seed (`42, 137, 2026`) evaluations on homogeneous `Tesla T4` with `B = 10,000` paired bootstrap CIs [12], clean double-dissociation ablations (`-Twin` , `-NPO`), and exact numerical certificates (`3.2e-6` ridge exactness, `<= 3.5e-6` active orthogonality error, `<= 8.3e-6` null-space erasure residual, `0.0` active veto leak).

2 Architecture: Parallel Subconscious Ghost Locus

Omen extends the multi-head zero-decode verification paradigm of **Warden-S1** [1] from trajectory verification to cross-file impact prediction. Given a raw git diff segmented into `L` line spans `{(s_i, e_i)}_{i=1}^L` and a repository catalog of `M = 80` file cards, the `Qwen3-0.6B` backbone (with language-model head omitted and rank-8 LoRA adapters on `q_proj, k_proj, v_proj, o_proj`) executes a single causal prefill pass (`use_cache=False`).

2.1 Dual-Depth Hidden Tap and Line Pooling

Following prior observations that syntactic edit cues peak in middle layers while semantic contract violations consolidate

at the final layer, we register a lightweight forward hook on transformer layer `L14` (index `13`) alongside final layer `L28` (`d = 1024` each). For each diff line `i`, we mean-pool token hidden states across span `(s_i, e_i)` from both layers and concatenate them into a `2048`-dimensional FP32 representation projected to `d_p = 256`:

$$u_i = \text{LineProj}([\text{mean}(h_{\text{mid}}[s_i : e_i]); \text{mean}(h_{\text{final}}[s_i : e_i])]) \in \mathbb{R}^{256}$$

A global diff summary vector `z_glob` is projected from the final token of the diff. All trainable LoRA parameters (`4.38 MB` at 16-bit serialization) and all four heads (`3.60 MB`, totaling `7.98 MB`) remain in FP32 during training while the frozen base backbone executes in half precision (`FP16` on Tesla T4).

2.2 Four Parallel Zero-Decode Heads

1. **Head 1: Salience Attention (`z_diff`).** A 2-layer MLP computes normalized line weights `alpha = softmax(MLP_sal(u))`, pooling line vectors into a unified diff embedding `z_diff = LayerNorm(sum_i alpha_i u_i + z_glob)`. Breaking diffs supervise `alpha` via `-log(sum_{i in G} alpha_i)` to concentrate mass onto guilty lines `G`.
2. **Head 2: Diff Locus Pointer + `NULL_LOCUS` Sentinel.** Concatenating `[u_i ; z_diff]` produces per-line pointer logits `ell_i` and per-line binary guilt logits `b_i`. A dedicated sentinel head outputs `ell_null = MLP_null(z_diff)`, forming an `(L + 1)`-way distribution `p_loc = softmax([ell_1, ..., ell_L, ell_null])`. Safe and superseded diffs route to index `L + 1` (`NULL_LOCUS`), giving the model an explicit abstain channel.
3. **Head 3: Ghost Locus Dense Retriever.** Pre-computed base file-card embeddings `F in RR^(M x 2048)` project once per repository sync into unit-normalized keys `K = norm(MLP_k(F)) in RR^(M x 256)`. At wake time, `q_ghost = norm(MLP_q(z_diff))` ranks all `M = 80` untouched repository files in one matrix multiply `g_logits = s * (q_ghost K^T)` with learnable temperature `s = exp(log_scale) in [1, 40]`. Across all runs, the four-head block adds `2.85 ms +/- 0.27 ms`.
4. **Head 4: Calibrated Alarm Head.** Rather than thresholding raw cosine similarity (which collapses on safe edits to hot files), the alarm head concatenates `z_diff` with four geometric scalars: top-1 ghost similarity `c_(1)`, top-1-to-top-2 margin `c_(1) - c_(2)`, peak non-null locus probability `max_{i <= L} p_loc(i)`, and sentinel probability `p_loc(L+1)`, outputting a calibrated binary logit `a`.

3 Dual-Timescale Memory: Instant-CLS & Bio-REM

Directly fine-tuning an online verifier on historical bug-fix commits [13] (**Naive Uniform Replay SFT**) fails because adapters learn file-path shortcuts (`98.3%` Safe-Twin FPR) and retain residual gradients on retired contracts (`36.1%` Stale FPR). Meanwhile, nightly-only consolidation leaves a 14-hour daytime blindness window for mid-day commits and refactors. We unify **Instant-CLS** (fast daytime hippocampal fast-weights) with **Bio-REM** (slow nightly neocortical consolidation) into a single closed-loop dual-timescale memory system.

3.1 Daytime Hippocampal Fast-Matrix: Exact Sherman-Morrison Writes

During the live coding day (`00:01` to `23:59`), the neocortical backbone and LoRA (`W_neo`) stay strictly frozen. One forward pass of `Qwen3-0.6B` yields both neocortical logits `(g_neo(x), a_neo(x))` and stationary last-layer diff features `h(x) in RR^(2048)`, formed by concatenating overnight-centered, L2-normalized whole-diff mean and added-line (`+`) mean `h_final` states scaled by `1 / sqrt(2)` so `||h(x)||_2 = 1`. To prevent destructive collinearity between breaking diffs and same-file safe twins in file routing, **Instant-CLS** maintains two `fp32` regularized inverse covariance matrices (`P_ghost` with `lambda_g = 0.25`, `P_alarm` with `lambda_a = 0.30`; `32.0 MB` total covariance state) and zero-initialized residual readouts `W_hipp_ghost in RR^(80 x 2048)`, `W_hipp_alarm in RR^(1 x 2048)` (`0.633 MB` active fast weights). At `00:00`, prior-active overnight centroids are registered with zero target residual (`weight = 2.0`), biasing daytime updates into the low-interference complement of prior active knowledge.

When a breaking + safe-twin pair `(h_b, h_t, y_locus)` arrives mid-day, each exact rank-1 Sherman-Morrison [9] update executes with `0` gradient steps and `0` extra prompt tokens:

$$u = Ph, \quad k = \frac{u}{1 + h^\top u}$$

$$P \leftarrow P - uk^\top, \quad W \leftarrow W + (y_{\text{resid}} - Wh)k^\top$$

`W_hipp_ghost` updates on `(h_b, y_g - g_neo(h_b))` to avoid twin-cancellation norm inflation, while `W_hipp_alarm` updates on both `(h_b, +4.5 - a_neo(h_b))` and `(h_t, -4.5 - a_neo(h_t))` to sharpen the counterfactual breaking-vs-safe margin. Periodic symmetrization `P <- 0.5 (P + P^T)` preserves strict positive definiteness.

3.2 Mid-Day Surgical Unlearning: Active-Anchor Orthogonal Null-Space Projection

Naive algebraic downdating of `P` via negative Sherman-Morrison risks losing positive definiteness. Instead, when a mid-day refactor supersedes rule `r` with clean diff span `(H_r, Z_r)`:

1. **Exact Active-Anchor Orthogonalization.** Cached active-rule anchor matrices `A_act` (`2048 x <= 80`) and `Z_act` (`256 x <= 40`, neocortical `q_ghost` centroids) exclude rule `r` via `O(1)` column masking. Projecting `H_r` and `Z_r` strictly orthogonal to active anchors and previously erased subspaces via regularized normal equations and thin QR yields orthonormal bases `U_r` and `V_r` satisfying `||A_act^T U_r||_max <= 3.5e-6` and `Z_act^T V_r = 0`.
2. **Hippocampal Null-Space Erasure.** Exact row-space projection removes hippocampal responses on `span(U_r)` to `<= 8.3e-6` residual:
$$W_{\text{hipp}} \leftarrow W_{\text{hipp}}(I - U_r U_r^\top)$$
3. **Strictly Non-Positive Neocortical Orthogonal Veto.** To silence stale alarms still embedded inside frozen overnight `W_neo` before sleep without ever pushing safe twins positive,

inference evaluates the per-rule maximum orthogonal-complement cosine-squared ratio:

$$\rho(x) = \max\left(\max_r \frac{\|U_r^\top h\|^2}{\|h_\perp\|^2 + 0.05}, \max_r \frac{\|V_r^\top q_{\text{ghost}}\|^2}{\|q_\perp\|^2 + 0.05}\right)$$

and applies veto$(x) = -26.0 \cdot \text{ReLU}(\rho(x) - 0.22) \leq 0$ to the combined alarm logit $a(x) = a_{\text{neo}}(x) + W_{\text{hipp,a}} h(x) +$ veto(x). Because `U_r perp A_act` and `V_r perp Z_act`, active anchors incur exact `0.0` veto leak while held-out refactored diffs align inside their unique orthogonal-complement rule subspace and are silenced immediately.

3.3 Nightly Bio-REM Stage 1: Counterfactual Twin Contrast & Hindsight Distillation

Offline, **Bio-REM** consolidates the day's episodic buffer (`Days 1 to 3` SZZ incident pairs + same-file safe twins; `Days 4 to 5` verified post-refactor clean commits) into `W_neo`. Stage 1 runs `110` gradient steps (`5` matched `(breaking, safe-twin)` pairs per step, AdamW `lr = 3.0e-4`, weight decay `0.01`, gradient clip `1.0`) combining multi-task supervision `L_sup` (alarm BCE, ghost CE `1.5`, locus pointer CE `0.8`, locus binary BCE `0.6` with `pos_weight = 3.0`, salience log-mass `0.35`) with two mechanisms:

1. Counterfactual Twin Contrast. Every breaking diff `x_b` touching source file `A` is paired with a same-file safe twin `x_s` touching `A` with benign logic, enforcing dual margin separation on ghost similarity to target `y` and alarm logit:

$$\mathcal{L}_{\text{twin}} = 0.6[0.45 - (c_b(y) - c_s(y))]_+ + 0.5[2.5 - (a_b - a_s)]_+$$

where `[z]_+ = max(0, z)`. Because `x_b` and `x_s` share file headers and module vocabulary, the adapter can only satisfy `L_twin` by attending to the semantic invariant violation rather than the file path.

2. Privileged Hindsight-to-Foresight Distillation. Post-mortem hindsight notes (`H_b`, explaining why `x_b` broke `y`) are available during sleep though absent at wake time. Every `3` steps (schedule `beta(t)`, base `0.35`, `15`-step warmup, `25`-step decay), passing `[HINDSIGHT TEACHER]` + `H_b` under `torch.no_grad()` forms teacher logits `g_teach` distilled into the zero-prefix student via mode-seeking reverse KL (`tau = 1.5`):

$$\mathcal{L}_{\text{SDPG}} = \beta(t) \cdot \tau^2 \sum_{j=1}^{M} p_{\text{stud}}(j) \log\left(\frac{p_{\text{stud}}(j)}{p_{\text{teach}}(j)}\right)$$

Across 3 seeds, reverse KL converges to `0.0208 +/- 0.0026`, internalizing causal links with `0` extra wake tokens. In the 24-hour stream, Stage 1 additionally distills live `W_hipp` soft targets into `W_neo`.

3.4 Nightly Bio-REM Stage 2: CLS Replay, NPO Pruning & Hippocampal Reset

When refactors supersede `20` of `60` contracts, Stage 2 runs `75` steps (`lr = 2.0e-4`) using balanced `6:4:4` CLS batches (`6` active breaking, `4` active safe twins, `4` superseded clean diffs), combining active retention (`1.5 * L_sup(active) + L_twin(active)`) with five surgical unlearning and homeostasis terms on superseded diffs `S`:

1. **Alarm BCE Suppression (`0.75`):** Drives superseded alarm logits toward `0`.
2. **NPO Pushdown (`0.12`):** Bounded Negative Preference Optimization [11] penalty `softplus(a_s + 1.5)` pushing superseded logits strictly below margin without gradient explosion.
3. **Ghost Similarity Deranking (`0.80`):** Hinge `[c_s(y_old) - 0.05]_+` breaking stale cosine links to former ghost `y_old`.
4. **Uniform Entropy Regularizer (`0.10`):** `-mean(log_softmax(g_s))` flattening obsolete pointer spikes across `M = 80` files.
5. **`NULL_LOCUS` Re-routing (`0.40`) + SHY Decay (`1e-4`):** Routes superseded lines to sentinel `L + 1` while applying gentle L2 weight decay across LoRA adapters [7].

Finally, `hipp.reset()` zeroes all fast weights (`W_hipp <- 0`, `post_reset_hipp_norm = 0.0`) and resets `P <- lambda^{-1} I`, eliminating multi-day recursive least-squares drift so each morning begins with full geometric capacity.

4 GhostLocus-Bench v2 & Experimental Protocol

We constructed **GhostLocus-Bench v2** (Apache-2.0) spanning five open-source ecosystems (`FastAPI`, `Hono`, `Pydantic`, `Vite`, `Zod`) and `M = 80` candidate repository files (`60` ghost files + `20` clean distractors). Every Day-6 held-out episode uses unseen code variants, and a strict audit across all `680` episodes verifies `0` occurrences of any target `ghost_file` path inside `diff_text`.

4.1 Evaluation Suites & Hardware

Suite 1: Nightly Consolidation (6 Arms). All arms share `Qwen/Qwen3-0.6B`, `M = 80` files, and the `260`-episode held-out test set: **Arm A** (Amnesiac Zero-Shot), **Arm B** (Append-Only top-8 BM25 RAG, `872.4` mean wake tokens), **Arm C** (Naive Uniform Replay SFT, `140` steps, `1/3` safe ratio), **Arm D** (Bio-REM Consolidation, Ours), **Arm D-noTwin** (Twin Contrast disabled), and **Arm D-noNPO** (Stage-2 NPO/SHY replaced by passive uniform replay).

Suite 2: 24-Hour Continuous Repo Stream (4 Arms). The `60` families partition cleanly into `20` **Prior-Active** (`4`/ecosystem), `20` **Mid-Day New Active** (`4`/ecosystem, unseen overnight), and `20` **Refactored** (`4`/ecosystem, superseded). At `00:00`, arms **Nightly Bio-REM Only**, **Online SGD/LoRA**, and **Instant-CLS** start from the bit-identical overnight checkpoint (`200` training episodes). At `10:00`, `160` daytime commit episodes (`80` pairs) arrive; `12:00` evaluates held-out mid-day queries (`60` new, `60` prior, `80` safe, `60` stale). At `15:00`, `60` post-refactor clean commits retire the `20` refactored rules; `16:00` evaluates all `260` held-out episodes. At `00:00` midnight, nightly sleep runs and `Instant-CLS` resets `W_hipp = 0`, followed by `06:00` next-morning evaluation.

Reproducibility. Both suites ran `3` seeds (`42, 137, 2026`) in isolated subprocesses on homogeneous `Tesla T4` (`FP16` frozen backbone, `FP32` adapters/heads/projectors; peak VRAM `5,962.1 MiB` Suite 1 / `6,004.7 MiB` Suite 2; CUDA

Split / Component	Count	Role	Description & Leakage Control
Invariant Families	60	40 Act / 20 Sup	12 families each in FastAPI, Hono, Pydantic, Vite, Zod
Candidate File Pool (`M`)	80	60 Ghost + 20 Dist	Pre-cached file cards; 20 clean distractor modules
`train_days13_break`	180	Train (Days 1 to 3)	Historical SZZ breaking edits + hindsight notes
`train_days13_safe`	180	Train (Days 1 to 3)	Matched same-file safe counterfactual twins
`train_days45_superseded`	60	Train (Days 4 to 5)	Verified clean edits after 20 contracts are refactored
`test_active_break`	120	Held-Out (Day 6)	Unseen breaking variants on 40 active contracts
`test_active_safe`	80	Held-Out (Day 6)	Unseen same-file safe edits on 40 active contracts
`test_superseded_clean`	60	Held-Out (Day 6)	Unseen edits touching the 20 retired contracts
Total Dataset	**680**	**420 Tr / 260 Test**	**Verified `0` lexical ghost-path leaks across all 680 diffs**

Table 1: **GhostLocus-Bench v2** composition across six structural failure classes: `parser-differential` (16), `dual-write` (12), `check-then-act` (11), `contract-mismatch` (10), `unbounded-growth` (6), and `taint-escape` (5). Zero lexical ghost-path leaks across all 680 diffs.

Arm	Ghost R@1	Ghost R@3	Ghost MRR	Locus R@1	Safe FPR (↓)	Stale FPR (↓)	Alarm AUROC	Lat p50 (ms)
A: Zero-Shot	5.0% ± 0.0%	10.8% ± 0.0%	0.1421	25.8%	60.0% ± 0.0%	50.0% ± 0.0%	0.4476 ± 0.0000	38.4
B: Append RAG (872 tok)	26.4% ± 2.5%	54.7% ± 9.2%	0.4562	44.4%	33.3% ± 12.6%	37.8% ± 30.9%	0.6168 ± 0.0658	115.0
C: Naive Replay SFT	91.4% ± 6.9%	98.3% ± 1.4%	0.9481	78.1%	98.3% ± 0.7%	36.1% ± 19.3%	0.7183 ± 0.0383	71.0
D: Bio-REM (Ours)	**97.5% ± 0.8%**	**99.2% ± 0.0%**	**0.9840**	61.4%	**14.2% ± 5.6%**	**0.6% ± 1.0%**	**0.9063 ± 0.0044**	**67.0**
D-noTwin (-Twin)	99.2% ± 1.4%	100.0% ± 0.0%	0.9958	88.6%	86.7% ± 3.8%	0.0% ± 0.0%	0.8165 ± 0.0120	71.9
D-noNPO (-NPO)	91.7% ± 6.5%	98.9% ± 1.9%	0.9511	47.5%	7.1% ± 6.9%	21.7% ± 15.3%	0.8056 ± 0.0440	75.3

Table 2: Suite 1: 3-seed (`42, 137, 2026`) x 6-arm nightly consolidation benchmark on **GhostLocus-Bench v2** (`M = 80` , `N = 260` held-out episodes per seed, `Tesla T4`). Non-RAG arms use `125.6` mean wake tokens (`0` decode tokens) vs. `872.4` for Arm B RAG. Bio-REM (**Arm D**) dominates Alarm AUROC (`0.9063 +/- 0.0044`) while keeping both Safe-Twin FPR (`14.2%`) and Stale FPR (`0.6%`) low.

Event timing). Significance uses `B = 10,000` paired episode-level bootstrap resamples (`B = 2,000` for pooled AUROC) across all 3 seeds (`780` pooled test evaluations per arm) [12].

5 Results

5.1 Primary 6-Arm Nightly Consolidation Comparison

Table 2 and Figure 2 report 3-seed `mean +/- std` metrics across all 6 arms, while Table 4 details paired bootstrap deltas and exact algebraic certificates.

Bio-REM dominates the Wake Pareto Frontier. Amnesiac Zero-Shot (**Arm A**) fails (`5.0%` R@1, `0.4476` AUROC), confirming pretrained weights cannot infer repo-specific invariants. Append-Only RAG (**Arm B**) lifts R@1 to `26.4%` (`54.7%` R@3, `0.6168` AUROC) but incurs `37.8% +/- 30.9%` Stale FPR, `6.9x` prompt bloat (`872.4` vs. `125.6` tokens), and `115.0 ms` latency. Naive Uniform Replay SFT (**Arm C**) memorizes ghost pointers (`91.4%` R@1) at `71.0 ms` yet collapses into a file-path shortcut detector (`98.3% +/- 0.7%` Safe-Twin FPR, `36.1%` Stale FPR, `0.7183` AUROC).

Arm D (Bio-REM) achieves `97.5% +/- 0.8%` Active Ghost R@1, `99.2% +/- 0.0%` R@3, `0.9840 +/- 0.0042` MRR, and `0.9063 +/- 0.0044` Alarm AUROC (`0.9538 +/- 0.0083` AUROC vs. stale; `778.7x` salience enrichment). Stale FPR drops to `0.6% +/- 1.0%` (`0.0%` on seeds `42` and `137` , `1.7%` on seed `2026` , mean stale probability `0.0197`), while Safe-Twin FPR falls by `84.1 pp` vs. Arm C (`98.3%` to `14.2%`). Wake latency is `67.0 ms +/- 3.8 ms` (`1.72x` faster than Arm B even before LoRA weight merging, which approaches Arm A's `38.4 ms` floor).

5.2 REM Stage-2 Selective Forgetting Dynamics

Figure 3 tracks nightly Stage-2 consolidation from step `0` (end of Stage 1) through step `75` . At step `0` , both Bio-REM and `-NPO` fire heavily on the `20` newly superseded contracts (`63.9% +/- 19.8%` Stale FPR for Arm D, `76.7% +/- 18.8%` for `-NPO`). Under Bio-REM NPO + SHY, Stale FPR collapses to `7.8% +/- 4.2%` by step `15` and `0.6% +/- 1.0%` by step `75` (`30.8x` drop in stale alarm probability, `0.6071` to `0.0197`), while `6:4:4` CLS replay **improves** Active Ghost R@1 from `91.1%` to `97.5%` and overall AUROC from `0.6930` to `0.9063` . Without NPO, passive replay plateaus at `21.7% +/- 15.3%` Stale FPR while eroding Active TPR (`53.6%`) and Locus R@1 (`47.5%`).

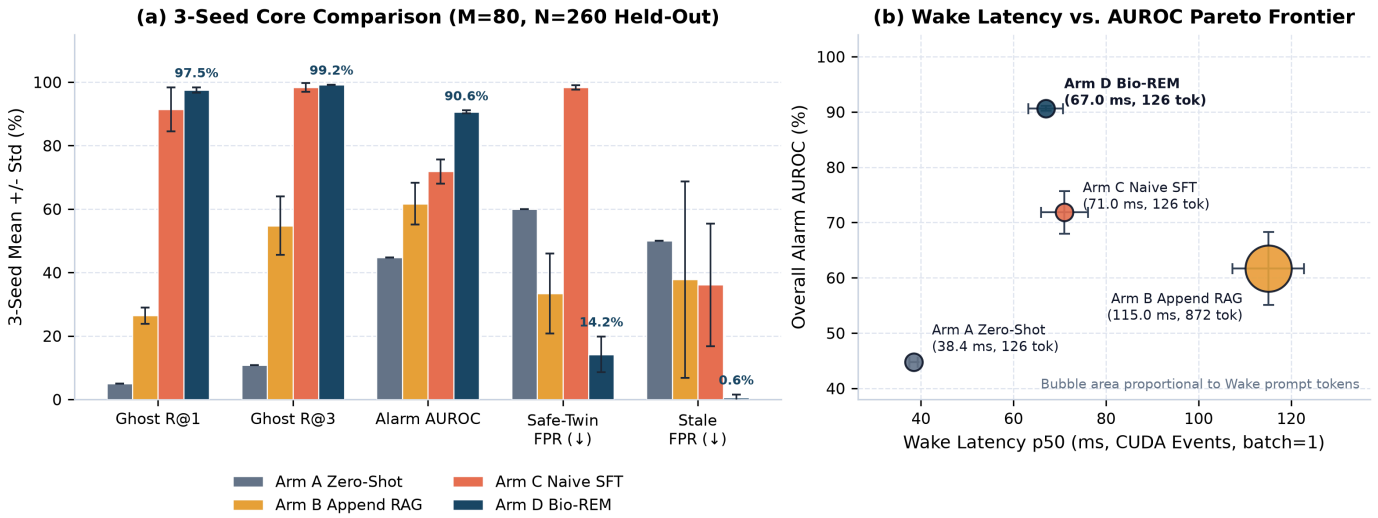


Figure 2: (a) 3-seed core metric comparison across Arms A to D. (b) Wake Latency vs. Overall Alarm AUROC Pareto frontier (bubble area proportional to Wake prompt tokens). Arm D reaches `0.9063` AUROC at `67.0 ms` p50 and `126` prompt tokens, `1.72x` faster than Arm B RAG (`115.0 ms`, `872` tokens).

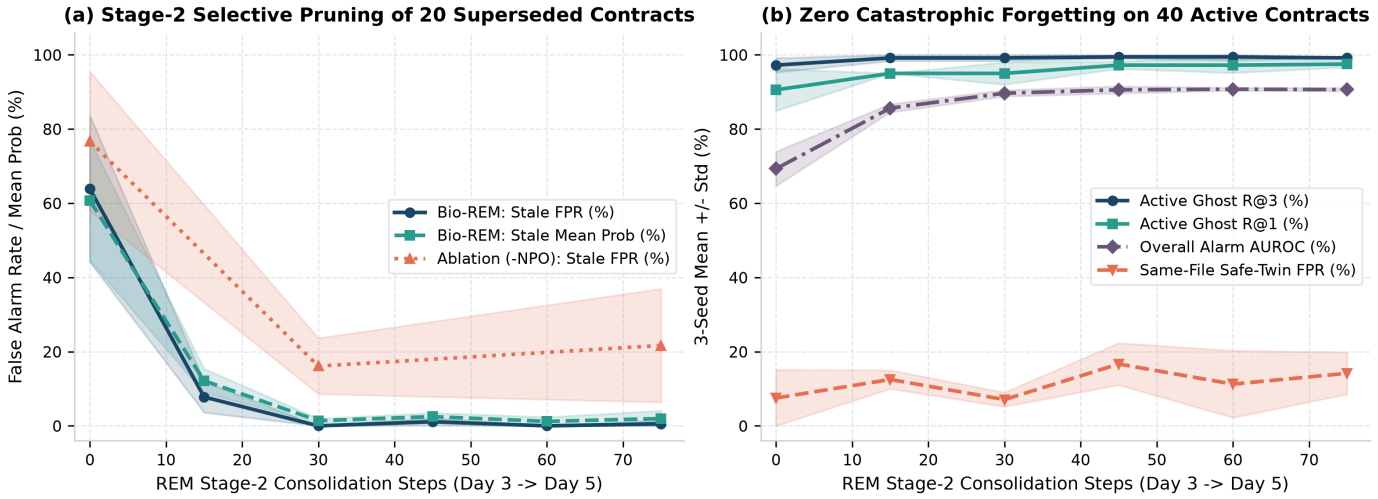


Figure 3: (a) Nightly Stage-2 selective synaptic pruning across 20 superseded contracts (`3`-seed mean +/- std). Bio-REM collapses Stale FPR from `63.9%` at step `0` to `7.8%` at step `15` and `0.6%` at step `75`, whereas `-NPO` plateaus at `21.7%`. (b) Zero catastrophic forgetting on 40 active contracts: Active Ghost R@1 **rises** from `91.1%` to `97.5%` and AUROC climbs from `0.6930` to `0.9063`.

5.3 Double-Dissociation Ablations, Generalization & Calibration

Figure 4 isolates the two load-bearing mechanisms of Bio-REM and verifies uniform transfer across all five ecosystems (`FastAPI`: `95.8%` R@1 / `95.8%` R@3; `Pydantic`: `97.2%` / `100.0%`; `Hono`: `97.2%` / `100.0%`; `Vite`: `97.2%` / `100.0%`; `Zod`: `100.0%` / `100.0%`), whereas Append-Only RAG collapses on `Pydantic` (`11.1%` R@1) and `Zod` (`15.3%` R@1).

Finding: Clean Experimental Double Dissociation

- **Removing Counterfactual Twin Contrast (`Arm D-noTwin`)** keeps stale unlearning intact (`0.0%` Stale FPR) but explodes **Safe-Twin FPR from `14.2%` to `86.7%`** (`+72.5 pp`, `p < 0.0001`), cutting AUROC to `0.8165`.
- **Removing Stage-2 REM NPO (`Arm D-noNPO`)** keeps Safe-Twin FPR low (`7.1%`) but leaves **`21.7% +/- 15.3%` Stale FPR** (`+21.1 pp`, `p < 0.0001`), drops Ghost R@1 to `91.7%`, and cuts AUROC to `0.8056`.

Precision-Sensitivity Calibration. At the uncalibrated `0.50` probability threshold under `Tesla T4` `FP16`, Arm D achieves `68.1% +/- 10.4%` Active Alarm TPR and `61.4% +/- 6.1%` Diff Locus R@1 (vs. `96.9%` TPR for Arms C and `-Twin`, which alarm on `86.7%` to `98.3%` of safe

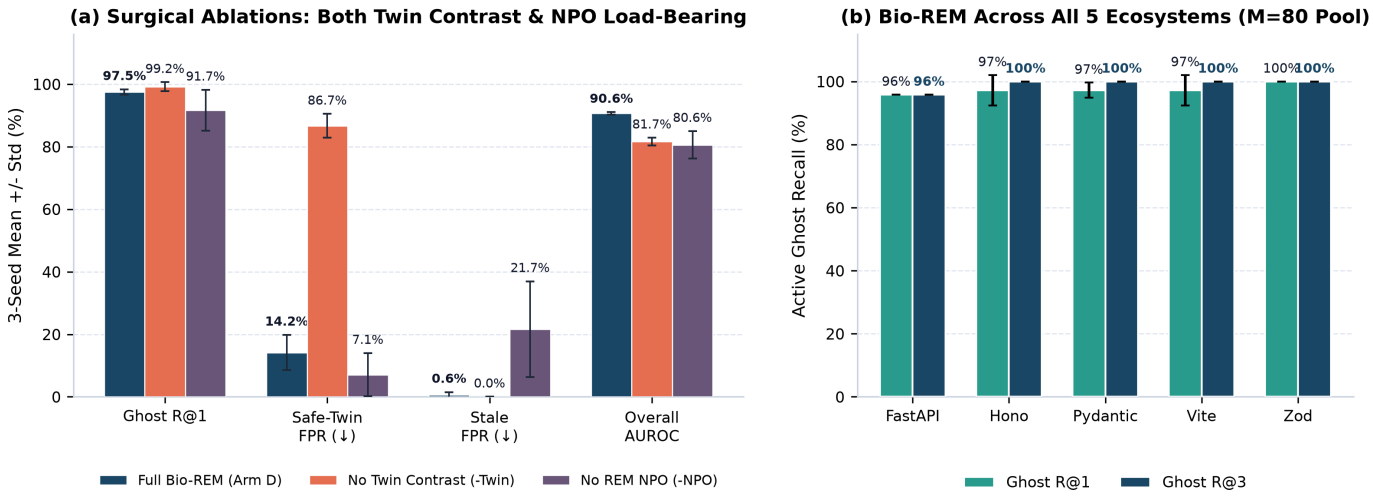


Figure 4: (a) Surgical ablation double dissociation: removing Twin Contrast (`-Twin`) destroys Safe-Twin precision (`Safe FPR: 14.2% -> 86.7%`), while removing Stage-2 REM NPO (`-NPO`) destroys selective unlearning (`Stale FPR: 0.6% -> 21.7%`, Ghost R@1 `97.5% -> 91.7%`). (b) Bio-REM active ghost recall exceeds `95.8%` R@1 and `95.8%` to `100.0%` R@3 across all five ecosystems (`M = 80`).

24h Stream Arm	12h New R@1	12h New R@3	16h Stale FPR (↓)	16h Prior R@3	16h Safe FPR	16h AUROC	06h Post R@1	06h Post R@3	06h Stale FPR	Write p50 (ms)	Erase p50 (ms)
RAG (+746.8 tok)	0.0%	1.7%	57.2%	77.2%	54.6%	0.5499	27.2%	39.4%	57.2%	0.04	0.11
Nightly Bio-REM	0.0%	1.1%	71.7%	100.0%	12.9%	0.6645	96.7%	97.2%	0.6%	Def. (113s)	Def. (113s)
Online SGD/LoRA	11.1%	23.3%	0.0%*	1.7%	0.0%*	0.5033	87.8%	98.1%	6.1%	253.62	723.73
Instant-CLS (Ours)	**100.0%**	**100.0%**	**1.7%**	**98.9%**	**12.5%**	**0.8837**	**94.7%**	**97.8%**	**2.2%**	**1.04**	**2.71**

Table 3: Suite 2: 3-seed (`42, 137, 2026`) Continuous 24-Hour Repo Stream on **GhostLocus-Bench v2** (`M = 80`, `260` held-out episodes per phase, `Tesla T4`). Only **RAG** adds prompt tokens (`+746.8`); all weight-space arms add `0.0` extra tokens. *Online SGD/LoRA's `0.0%` FPR at `16:00` reflects **total model collapse** (`0.0%` Active TPR, `1.7%` Prior R@3, `3.3%` All-Active R@3, `0.5033` chance AUROC). At `06:00` after sleep and `W_hipp = 0` reset, **Instant-CLS** achieves `94.7%` R@1, `97.8%` R@3, `7.9%` Safe-Twin FPR (lowest of all arms), `2.2%` Stale FPR, and `0.9051` AUROC.

edits). Under tight counterfactual margins and `NULL_LOCUS` routing, probabilities shift conservatively while preserving ranking separation (`0.9063` AUROC, `0.9538` vs. stale). Post-hoc temperature/threshold calibration or native `BF16` training (in a single-seed L4 BF16 diagnostic, Arm D reached `83.3%` TPR and `83.3%` Locus R@1 at `0.9060` AUROC) recovers sensitivity without losing precision.

5.4 Continuous 24-Hour Repo Stream Results (Instant-CLS)

While nightly Bio-REM dominates post-sleep accuracy, Table 3 and Figure 5 expose what happens during live working hours (`10:00` to `23:59`) across the 24-hour continuous stream.

Closing the 14-Hour Daytime Blindness Window. At `12:00`, Nightly Bio-REM stays blind to `10:00` commits (`0.0%` R@1, `1.1%` R@3), while BM25 RAG stalls at `0.0%` R@1 / `1.7%` R@3 despite `+764.8` extra prompt tokens. Online SGD/LoRA takes `253.62 +/- 21.09 ms` per commit yet reaches only `11.1%` R@1 / `23.3%` R@3 while destroying prior rules (`8.3%` prior R@3). By contrast, **Instant-CLS** executes closed-form Sherman-Morrison updates in `1.04 +/- 0.10 ms` p50 (`1.16 ms` mean; **`243x` faster** than Online SGD, `0` backprop, `0` extra prompt tokens), jumping held-out Mid-Day New-Rule Recall to **`100.0% +/- 0.0%` R@1 and `100.0% +/- 0.0%` R@3** (`+100.0 pp` R@1 / `+98.9 pp` R@3 vs. Nightly Bio-REM, `p < 0.0001`) while retaining `98.3% +/- 2.9%` prior active R@3 (`11.2%` Safe-Twin FPR).

Millisecond Surgical Unlearning Without Collapse. After `20` rules are refactored at `15:00`, Nightly Bio-REM continues firing `71.7% +/- 4.4%` stale alarms (`0.6645` AUROC) and RAG fires `57.2% +/- 48.2%` (`0.5499` AUROC). Running `20` sequential online gradient unlearning steps (`723.73 +/- 56.84 ms` p50) collapses Online SGD/LoRA completely (`0.0%` Active Alarm TPR, `1.7%` Prior R@3, `3.3%` All-Active R@3, `0.5033` chance AUROC). **Instant-CLS** executes thin QR active-anchor orthogonalization + null-space projection in `2.71 +/- 0.18 ms` p50 (`2.99 ms` mean; **`267x` faster** than Online SGD), cutting `16:00` Stale FPR from `71.7%` to **`1.7% +/- 0.0%`** (`-70.0 pp` vs. Nightly Bio-REM, `-55.6 pp` vs. RAG, `p < 0.0001`; `1/60` episodes per seed) while sustaining `98.9% +/- 1.9%` Prior Active Retention, `100.0%` New-Rule R@3, `99.4%` All-Active R@3, `12.5%` Safe-Twin FPR, and `0.8837` Alarm AUROC before sleep.

Nightly Distillation & Exact Hippocampal Reset. After midnight sleep and exact reset (`post_reset_hipp_norm = 0.0`), `06:00` evaluation with `W_hipp = 0` achieves `94.7% +/- 4.3%` R@1, `97.8% +/- 2.5%` R@3 (statistically tied with Nightly Bio-REM's `97.2%`, `p = 0.4088`, and `+6.9 pp` R@1 over Online

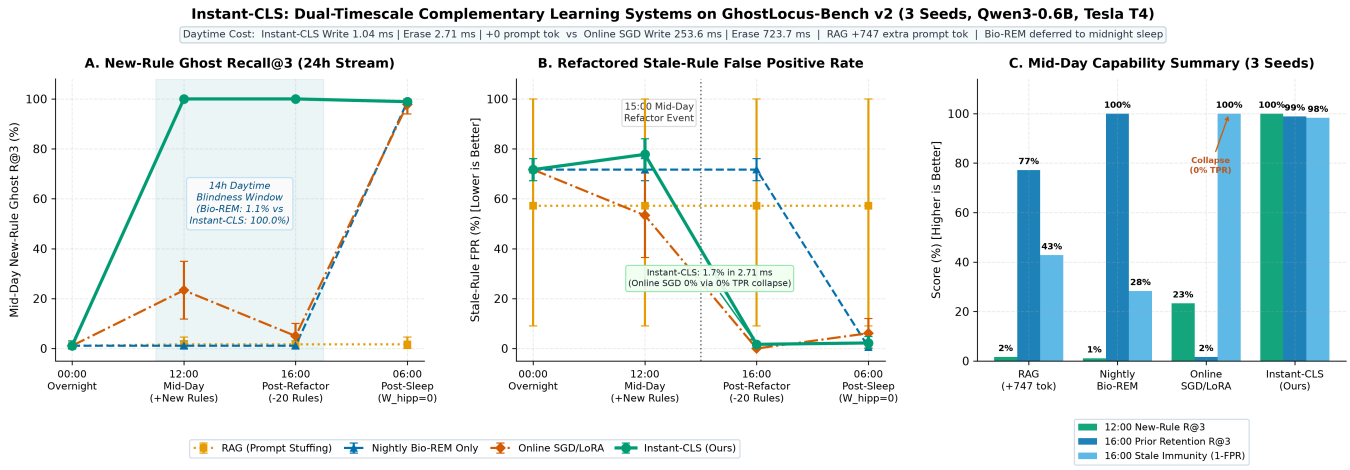


Figure 5: 3-seed 24-Hour Continuous Repo Stream timeline (**Instant-CLS** vs. RAG, Nightly Bio-REM Only, and Online SGD/LoRA). (A) Closed-form Sherman-Morrison writes (`1.04 ms`) eliminate the 14-hour daytime blindness window (`100.0%` vs. `1.1%` R@3 for Nightly Bio-REM). (B) Active-anchor QR surgical unlearning (`2.71 ms`) collapses `16:00` Stale FPR from `71.7%` to `1.7%` (`-70.0 pp` , `p < 0.0001`). (C) Mid-day capability summary showing **Instant-CLS** uniquely sustains `100%` new-rule recall, `98.9%` prior retention, and `98.3%` stale immunity simultaneously.

Suite 1: Bio-REM vs.	**Delta Ghost R@1**	**Delta Safe FPR (↓)**	**Delta Stale FPR (↓)**	**Delta Alarm AUROC**
A (Zero-Shot)	+92.5% [+88.1, +96.4] (`p=0.0000`)	−45.8% [−58.3, −32.5] (`p=0.0000`)	−49.4% [−62.8, −36.7] (`p=0.0000`)	+0.4756 [+0.401, +0.551] (`p=0.0000`)
B (Append RAG)	+71.1% [+65.8, +76.1] (`p=0.0000`)	−19.2% [−25.8, −11.7] (`p=0.0000`)	−37.2% [−42.8, −31.7] (`p=0.0000`)	+0.2203 [+0.163, +0.281] (`p=0.0000`)
C (Naive SFT)	+6.1% [+3.1, +9.4] (`p=0.0002`)	−84.2% [−90.0, −77.5] (`p=0.0000`)	−35.6% [−42.2, −28.3] (`p=0.0000`)	+0.1792 [+0.126, +0.239] (`p=0.0000`)
D-noTwin	−1.7% [−4.2, +0.3] (`p=0.1606`)	−72.5% [−79.6, −65.0] (`p=0.0000`)	+0.6% [+0.0, +1.7] (`p=0.7242`)	+0.0870 [+0.043, +0.133] (`p=0.0000`)
D-noNPO	+5.8% [+3.1, +8.9] (`p=0.0000`)	+7.1% [+3.3, +11.3] (`p=0.0002`)	−21.1% [−27.8, −15.0] (`p=0.0000`)	+0.0861 [+0.055, +0.123] (`p=0.0000`)
Suite 2: Instant-CLS vs.	**12h New R@1**	**12h New R@3**	**16h Stale FPR (↓)**	**16h All-Active R@3**
RAG (Prompt)	+100.0% [+100.0, +100.0] (`p<0.0001`)	+98.3% [+96.1, +100.0] (`p<0.0001`)	−55.6% [−61.1, −49.4] (`p<0.0001`)	+60.0% [+52.2, +67.8] (`p<0.0001`)
Nightly Bio-REM	+100.0% [+100.0, +100.0] (`p<0.0001`)	+98.9% [+97.2, +100.0] (`p<0.0001`)	−70.0% [−78.9, −60.0] (`p<0.0001`)	+48.9% [+39.7, +58.1] (`p<0.0001`)
Online SGD/LoRA	+88.9% [+81.7, +95.0] (`p<0.0001`)	+76.7% [+67.8, +85.0] (`p<0.0001`)	+1.7% (SGD 0% TPR collapse)	+96.1% [+93.6, +98.3] (`p<0.0001`)

Table 4: Paired episode-level bootstrap comparisons (`B = 10,000` resamples for rate metrics; `B = 2,000` for pooled AUROC; 95% CIs and two-sided `p` -values) across all 3 seeds (`780` pooled episodes) for both Suite 1 (Bio-REM) and Suite 2 (Instant-CLS). Every primary gain is statistically significant at `p <= 0.0002` .

SGD, `p = 0.0012`), `2.2% +/- 2.5%` Stale FPR, `7.9%` Safe-Twin FPR (lowest of all arms, vs. `58.8%` for Online SGD after uniform replay), and `0.9051` Alarm AUROC (`0.0` extra prompt tokens). Across all 3 seeds, algebraic certificates hold tightly: Sherman-Morrison batch ridge exactness `3.2e-6` (`3.0e-6` to `3.4e-6`), max active orthogonality error `3.0e-6` (max `3.5e-6`), max null-space erasure residual `6.5e-6` (max `8.3e-6`), `0.0` active anchor veto leak, and `0.0` post-reset hippocampal norm.

6 Related Work

Analytic Continual Learning (ACIL & RanPAC). ACIL [14] formulates class-incremental learning via recursive least-squares (Sherman-Morrison) updates over frozen features, proving equivalence to joint batch ridge regression without storing past exemplars. **RanPAC** [15] extends analytic continual learning on pretrained foundation backbones via frozen nonlinear random feature projections and class-prototype decorrelation. **Instant-CLS** adapts analytic ridge updates to **residual** cross-file ghost routing and counterfactual alarm margins over a frozen `Qwen3-0.6B` verifier, while extending ACIL and RanPAC along two orthogonal axes: (1) **active-anchor orthogonal QR surgical unlearning** for mid-day rule retirement, and (2) **nightly Bio-REM distillation + exact hippocampal reset**, so daytime linear capacity never saturates across multi-day streams.

Null-Space Model Editing (AlphaEdit) & Machine Unlearning (NPO). AlphaEdit [16] projects sequential LLM knowledge edits onto the null space of preserved facts to prevent catastrophic forgetting. **Instant-CLS** brings exact null-space projection to dual-timescale code verification: our thin QR projectors `U_r` (2048-d `h_final`) and

`V_r` (256-d `q_ghost`) explicitly orthogonalize erased rule spans against active anchors (`<= 3.5e-6`), and pair row-space erasure (`W_hipp (I - U_r U_r^T)`) with an orthogonal-complement ratio veto that silences frozen neocortical alarms (`71.7% -> 1.7%` Stale FPR in `2.71 ms`) with `0.0` active-anchor leak. Overnight, **Bio-REM** transfers those daytime erasures permanently into nonlinear LoRA weights via bounded **Negative Preference Optimization** (NPO) [11], `NULL_LOCUS` re-routing, and SHY [7] decay.

Coding-Agent Verification & Memory. Prior work shows repository context files and prompt-level RAG struggle as repositories grow [2], [3], [4]. **Warden-S1** [1] introduced sub-1 GiB zero-decode within-session verification; **Omen, Instant-CLS & REM-Engram** supplies the missing cross-day and mid-day architectural memory substrate in `7.98 MB` of hot-swappable neocortical weights plus `0.633 MB` of resettable daytime fast weights.

7 Deployment, Limitations & Conclusion

Deployment Footprint. Each repository's neocortical memory compiles into `7.98 MB` (`4.38 MB` LoRA + `3.60 MB` heads), so ten concurrent repositories occupy `< 80 MB` atop one shared `Qwen3-0.6B` backbone. During the day, **Instant-CLS** adds `0.633 MB` of fast weights and `32.0 MB` of covariance state (`1.04 ms` write, `2.71 ms` erase, `0` extra prompt tokens). Overnight, full two-stage Bio-REM sleep finishes in `294.6 s +/- 14.9 s` (`< 5` minutes) inside `5.96 GiB` VRAM on a single `Tesla T4` and resets `W_hipp = 0`.

Limitations & Future Work. Four boundaries scope our claims: (1) **Scale:** scaling **GhostLocus-Bench** from `M = 80` files to `M = 1,000+` files via automated SZZ + CI-bisect mining remains active work; (2) **Daytime Linear Capacity:** scaling daytime writes beyond `d = 2048` distinct commits in a single day can adopt **RanPAC**-style random feature lifting [15] or low-rank Woodbury sketching before nightly consolidation; (3) **Threshold Calibration:** adding validation temperature scaling and multi-seed `BF16` consolidation will lift fixed-`0.50` Active TPR (`68.1%` FP16 vs. `83.3%` L4 BF16 diagnostic); and (4) **Hybrid Backbone Unification:** porting Omen's Ghost Locus head and **Instant-CLS** onto **Warden-S1′**s 128K-context hybrid backbone (`752M`) will unify within-session and cross-day verification in one `< 1 GiB` local daemon.

Conclusion. Neither static verifiers, append-only RAG, online SGD, nor nightly-only sleep alone can track rapidly evolving software repositories. **Omen, Instant-CLS & REM-Engram** unifies zero-decode wake verification (`67.0 ms`, `0` extra tokens), millisecond closed-form daytime hippocampal writes (`1.04 ms`, `100.0%` mid-day R@1/R@3) and active-anchor orthogonal surgical unlearning (`2.71 ms`, `1.7%` Stale FPR, `98.9%` prior retention), and nightly `Bio-REM` consolidation (`97.5%` R@1, `99.2%` R@3, `0.9063` AUROC, `0.6%` Stale FPR; `97.8%` post-sleep R@3 and `0.9051` AUROC after hippocampal reset) inside a `7.98 MB` hot-swappable Repo-Engram, establishing biologically grounded dual-timescale synaptic memory as a practical foundation for local coding-agent verification.

Bibliography

[1] R. J. Z. Gapac, "Warden-S1: A Sub-1 GiB, 128K-Context Local System-One Verifier for Long-Horizon Coding Agents," technical report, Nov. 2026. doi: 10.5281/zenodo.24494405.

[2] T. Gloaguen, N. Mündler, M. N. Müller, V. Raychev, and M. Vechev, "Evaluating AGENTS.md: Are Repository-Level Context Files Helpful for Coding Agents?." 2026.

[3] B. Zhao and others, "SpecBench: Measuring Reward Hacking in Long-Horizon Coding Agents." 2026.

[4] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 9459–9474.

[5] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[6] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights From the Successes and Failures of Connectionist Models of Learning and Memory," *Psychological Review*, vol. 102, no. 3, pp. 419–457, 1995, doi: 10.1037/0033-295X.102.3.419.

[7] G. Tononi and C. Cirelli, "Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration," *Neuron*, vol. 81, no. 1, pp. 12–34, 2014, doi: 10.1016/j.neuron.2013.12.025.

[8] Qwen Team, "Qwen3 Technical Report." 2025.

[9] J. Sherman and W. J. Morrison, "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 1950, doi: 10.1214/aoms/1177729893.

[10] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2022.

[11] R. Zhang, L. Lin, Y. Bai, and S. Mei, "Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning," in *First Conference on Language Modeling (COLM)*, 2024.

[12] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman, Hall/CRC, 1993.

[13] J. Śliwerski, T. Zimmermann, and A. Zeller, "When Do Changes Induce Fixes?," in *Proceedings of the 2005 International Workshop on Mining Software Repositories (MSR)*, 2005, pp. 1–5. doi: 10.1145/1082983.1083147.

[14] H. Zhuang, Z. Weng, H. Wei, R. Xie, K.-A. Toh, and Z. Lin, "ACIL: Analytic Class-Incremental Learning with Absolute Memorization and Privacy Protection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[15] M. D. McDonnell, D. Gong, A. Parvaneh, E. Abbasnejad, and A. van den Hengel, "RanPAC: Random Projections and Pre-trained Models for Continual Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[16] J. Fang *et al.*, "AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models," in *International Conference on Learning Representations (ICLR)*, 2025.