# SKILLAGE: LIFECYCLE-AWARE SKILL AGING AND SAFE PRUNING FOR SELF-EVOLVING AGENTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Self-evolving Large Language Model (LLM) agents excel at accumulating procedural skills for complex tasks. Yet, unchecked skill accumulation inevitably triggers severe repository bloat, where redundant or obsolete routines inflate retrieval overhead and disrupt execution fidelity. To tackle this challenge, we propose SkillAge[1], an online lifecycle-aware framework for verifiable skill governance and safe pruning. SkillAge models procedural knowledge as dynamic assets transitioning across controllable operational states. From trajectory feedback, it extracts multidimensional evidence, including exposure-weighted performance evidence, semantic redundancy, and context cost, to parameterize an evidence-grounded retention score. This score guides a constrained Markov decision process (CMDP) reinforcement learning policy to optimize repository compactness without future-task annotations. Crucially, irreversible deletion is excluded from online actions; instead, candidates are staged in quarantine where a dual-gate replay protocol certifies permanent removal by bounding both cumulative and worst-case capability degradation. Evaluations across continuous lifelong benchmarks demonstrate that SkillAge achieves up to 89.5% repository compression and 92.0% token exposure reduction while reliably preserving critical future capabilities.

## 1 INTRODUCTION

Large Language Model (LLM) agents increasingly rely on external skill libraries to execute complex tasks requiring tool use, long-horizon planning, and situational adaptation (Yao et al., 2022; Wang et al., 2023; Zheng et al., 2025b). A skill repository stores systematically organized procedural knowledge—such as API call sequences, Python scripts, and modular sub-routines—enabling agents to leverage prior experience rather than solving every problem from scratch. Recent frameworks demonstrate that skill accumulation is central to lifelong learning, allowing agents to iteratively expand their capabilities across diverse domains (Zhao et al., 2023; Zhang et al., 2026d; Li et al., 2026). However, continuous operation inevitably transforms these static libraries into highly dynamic, unbounded assets. This evolution dictates that the skill repository is not merely a passive cache, but a vital persistent ecosystem that demands rigorous lifecycle governance to maintain execution fidelity over time (Huang et al., 2026; Li, 2026; Wang et al., 2024).

Current self-evolving agents focus predominantly on the acquisition and reuse of skills, implicitly treating knowledge accumulation as a monotonically beneficial process (Fan et al., 2026; Park et al., 2023; Packer et al., 2023). Yet, this unchecked, append-only accumulation introduces a critical structural bottleneck: skill bloat. As the repository expands indefinitely, redundant, obsolete, or noisy skills begin to crowd out genuinely useful ones. This overpopulation heavily inflates the retrieval context overhead and disrupts the agent's downstream decision-making process by introducing conflicting procedural guidance (Taheri, 2026; Hu et al., 2025). Without explicit lifecycle constraints, an expanding repository transforms from a helpful knowledge resource into a primary source of execution interference. Ultimately, unbounded skill accumulation paradoxically degrades long-term agent performance and task generalization across lifelong environments rather than enhancing it (Gao et al., 2026; Sinha et al., 2025; Zhang et al., 2026a; Zheng et al., 2025a).

A standard countermeasure to repository growth is pruning inactive or underperforming skills using frequency-based cache eviction heuristics (Du et al., 2025; Zhang et al., 2026c; Shen et al., 2026).

[1] https://anonymous.4open.science/r/SkillAge-BAAF.



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

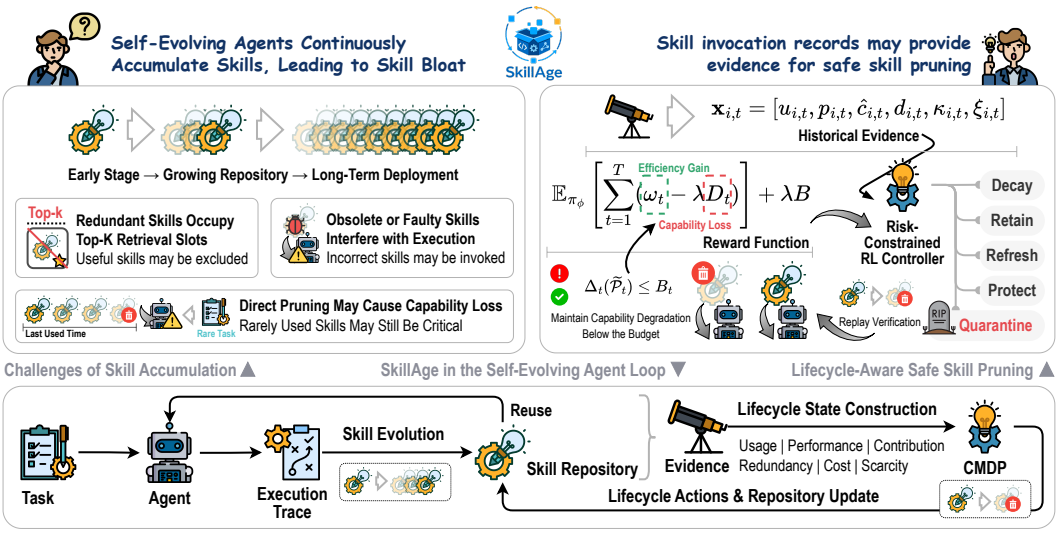


Figure 1: Motivation and overview of SkillAge. Append-only skill accumulation leads to repository bloat, where redundant or obsolete skills degrade long-term execution, while pruning risks capability loss. SkillAge addresses this via lifecycle-aware, risk-constrained quarantine replay verification.

However, direct pruning induces a severe secondary failure mode: catastrophic capability collapse. Unlike passive text entries, procedural skills operate as executable artifacts encoding strict environment assumptions and implicit execution guards (Dritsas & Trigka, 2026; Maragheh & Deldjoo, 2025). Consequently, rarely invoked skills often hold highly specialized utility for edge-case tasks, and purging them irreparably destroys the agent's long-tail problem-solving capacity. As illustrated in Figure 1, agents remain trapped in an irreconcilable conflict between execution-degrading repository bloat and capability-destroying heuristic pruning. Recognizing these risks (Ouyang et al., 2026; Pu et al., 2026), we argue that sustainable autonomy requires a paradigm shift from ad-hoc deletion to mathematically verifiable, risk-constrained skill governance.

To tackle this challenge, we introduce SkillAge, an online framework for verifiable skill governance and safe pruning in self-evolving agents without requiring future task annotations or risking irreversible capability collapse. As illustrated in Figure 1, SkillAge treats procedural knowledge not as static executable scripts, but as evolving assets governed by explicit operational states. During long-term interaction, the agent monitors multidimensional execution feedback–including invocation frequency, semantic redundancy, and context footprint–to track the operational viability of each tool. Instead of executing irrevocable deletions directly during task solving, the agent dynamically adjusts accessibility through reversible actions such as decay and protection. Suspicious or underperforming tools are staged in an isolated quarantine holding area, preventing execution interference while preserving functional integrity until rigorous historical replay certifies their retirement.

Our primary contributions are summarized as follows. 1) We formulate skill governance as a constrained Markov decision process (CMDP) that decouples reversible lifecycle maintenance from physical retirement. 2) We develop an exposure-weighted performance estimator whose regularized shrinkage structurally protects infrequently called specialized capabilities from premature eviction. 3) We introduce a dual-gate quarantine verification protocol that provably bounds both cumulative degradation and worst-case single-task capability collapse on historical benchmarks. 4) We design a factorized reinforcement learning policy optimized via primal-dual updates to navigate the compactness-utility trade-off without future-task annotations. 5) We evaluate SkillAge across continuous lifelong benchmarks on three LLM backbones, demonstrating substantial repository compaction and retrieval efficiency alongside transparent accounting of amortized verification compute.

## 2 Related Work

**Lifelong Agent Learning and Skill Accumulation.** Modern LLM agents consolidate traces into libraries for planning (Wang et al., 2023; Zhao et al., 2023; Wang et al., 2024). Yet, prevail-

ing paradigms assume monotonic capability growth, causing severe repository bloat (Zhang et al., 2026d; Li et al., 2026). Under restricted top-$K$ contexts, redundant routines dilute bandwidth and introduce conflicting guidance (Taheri, 2026; Gao et al., 2026). Sustainable lifelong agency demands structural governance over expanding libraries rather than unconstrained accumulation.

**Skill Repository Governance and Lifecycle Maintenance.** Recent approaches mitigate bloat via evolving ecosystems or reinforcement-learned memory maintenance (Pu et al., 2026; Packer et al., 2023). However, these schemes govern passive textual records (Li et al., 2025). In contrast, procedural skills are executable routines governed by runtime dependencies. Pruning them constitutes active intervention rather than passive eviction, demanding risk-constrained formulations that provably bound capability loss across non-linear interactions (Lin et al., 2026; Zhang et al., 2026b).

**Continual Retention, Replay, and Verifiable Skill Retirement.** Preserving competence during repository compaction parallels preventing catastrophic forgetting (Feng et al., 2026; Cho et al., 2025). Traditional heuristics fail because rare routines protect specialized cases and redundancies provide fallbacks (Dritsas & Trigka, 2026). Evaluating removals in isolation overlooks non-linear synergies between tools. Existing architectures lack verifiable retirement mechanisms that certify candidate evictions before committing irreversible deletion (Ouyang et al., 2026; Lin et al., 2026).

In summary, prior work exposes three central tensions: (i) open-ended skill acquisition versus retrieval dilution; (ii) passive textual memory versus stateful executable routines; and (iii) unverified pruning under combinatorial dependencies. Resolving this trilemma requires shifting from ad-hoc cache eviction to verifiable, lifecycle-aware governance that decouples reversible maintenance from permanent eviction, isolates collaborative tool credit, and mathematically certifies retirement against capability collapse. These fundamental requirements directly motivate the design of SkillAge.

## 3 PRELIMINARIES

We formalize lifelong agent interaction over an online task stream $\mathcal{T} = \{\tau_t\}_{t=1}^{T}$ (Wang et al., 2023), where each task $\tau_t = (q_t, \mathcal{E}_t, \nu_t)$ comprises instruction $q_t$, environment $\mathcal{E}_t$, and verifier $\nu_t$. At round $t$, the agent maintains an executable skill repository $\mathcal{R}_t = \{s_i\}_{i=1}^{N_t}$. Given context budget $K$, a retriever $G[\cdot]$ extracts an active view $\mathcal{V}_t = G[q_t, \mathcal{R}_t, K] \subseteq \mathcal{R}_t$ (Packer et al., 2023). Over $M_t$ execution steps, the agent generates trajectory $\zeta_t$ using invoked skills $\mathcal{I}_t \subseteq \mathcal{V}_t$, yielding trajectory fidelity score $y_t = \nu_t[\zeta_t] \in [0, 1]$ from the verifier (Yao et al., 2022). To mitigate retrieval dilution while preventing premature eviction, SkillAge profiles each skill $s_i \in \mathcal{R}_t$ via a normalized six-dimensional lifecycle state vector over continuous lifelong agent operation:

$$\mathbf{x}_{i,t} = [u_{i,t},\ p_{i,t},\ \hat{c}_{i,t},\ d_{i,t},\ \kappa_{i,t},\ \xi_{i,t}]^\top \in [0,1]^2 \times [-1,1] \times [0,1]^2 \times (0,1], \tag{1}$$

which captures frequency $u_{i,t}$, success $p_{i,t}$, contribution $\hat{c}_{i,t}$, redundancy $d_{i,t}$ (Reimers & Gurevych, 2019), context cost $\kappa_{i,t}$, and scarcity $\xi_{i,t}$. The controller observes context $\mathbf{s}_t = (\{(\mathbf{x}_{i,t}, \rho_{i,t}, m_{i,t})\}_{i=1}^{N_t}, \mathbf{g}_t, \beta_t)$ with retention score $\rho_{i,t}$, operational mode $m_{i,t} \in \{m^a, m^d, m^p\}$, repository statistics $\mathbf{g}_t$, and normalized remaining budget $\beta_t$.

A factorized policy $\pi_\phi(\mathbf{a}_t \mid \mathbf{s}_t)$ outputs lifecycle actions across $\{a^r, a^d, a^f, a^p, a^q\}$. Excluding physical deletion from policy choices, quarantined routines $\mathcal{Q}_t$ undergo historical replay certification (Rolnick et al., 2019) on $\mathcal{B}_t^\circ = \bigcup_{s_i \in \mathcal{Q}_t} \mathcal{H}_{i,t}$. Deletions $\mathcal{P}_t \subseteq \mathcal{Q}_t$ are accepted only under dual bounds $D_t \le B_t$ and $D_t^{\max} \le \epsilon_{\max}$, while rejected candidates are restored under mode $m^p$. Evaluated against reference library $\bar{\mathcal{R}}_t$ and view $\bar{\mathcal{V}}_t$, step overhead $O_t$ is formulated as:

$$O_t = \frac{1}{3}\left[\frac{|\mathcal{R}_t|}{N_0} + \min\left\{1,\ \frac{\ell[\mathcal{V}_t]}{\max\{1,\ \ell[\bar{\mathcal{V}}_t]\}}\right\} + \frac{1}{N_t}\sum_{i=1}^{N_t} d_{i,t}\right]. \tag{2}$$

Given newly acquired skills $\mathcal{N}_t$, the repository updates as $\mathcal{R}_{t+1} = (\mathcal{R}_t \setminus \mathcal{P}_t) \cup \mathcal{N}_t$, and the remaining budget decrements by $B_{t+1} = B_t - D_t$. Ultimately, SkillAge formalizes repository governance as a finite-horizon CMDP (Altman, 2021) to minimize cumulative step overhead while guaranteeing provable dual-gate safety bounds over dynamic task execution environments:

$$\pi^* = \arg\min_\pi \mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=1}^{T} O_t\right] \quad \text{s.t.} \quad \mathbb{E}_\pi\left[\sum_{t=1}^{T} D_t\right] \le B, \quad D_t^{\max} \le \epsilon_{\max}, \quad \forall t \in [T]. \tag{3}$$

This formulation reconciles library compactness, context economy, and redundancy mitigation into a unified objective, while safety constraints prevent capability degradation and collapse.

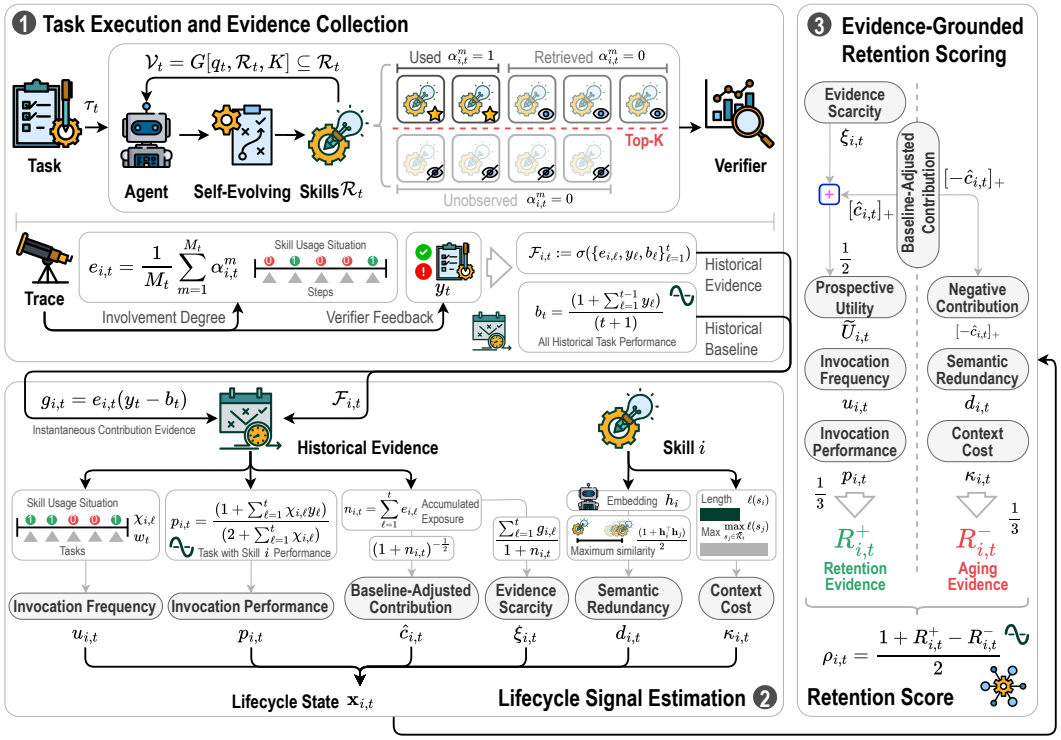


Figure 2: The SkillAge lifecycle profiling pipeline. Trajectory feedback informs the extraction of state vector $\mathbf{x}_{i,t}$ across usage dynamics, contribution, redundancy, and context overhead. Positive retention evidence $R^+_{i,t}$ and negative aging pressure $R^-_{i,t}$ are balanced to derive retention score $\rho_{i,t}$.

## 4 METHOD

Figure 2 illustrates SkillAge's three-stage profiling pipeline: execution, signal estimation, and retention scoring. This parameterizes $\rho_{i,t}$ to guide risk-constrained policy decisions and gate pruning via historical replay verification, ensuring safe long-horizon lifelong agent operation.

### 4.1 TASK EXECUTION AND EVIDENCE COLLECTION

From trajectory $\zeta_t$, step invocation indicator $\alpha^m_{i,t} = \mathbb{I}[s_i \in \mathcal{I}_{t,m}]$ yields task exposure $e_{i,t} = M_t^{-1} \sum_{m=1}^{M_t} \alpha^m_{i,t} \in [0,1]$ and activation flag $\chi_{i,t} = \mathbb{I}[e_{i,t} > 0]$. To isolate individual skill credit from holistic outcome $y_t \in [0,1]$, we measure instantaneous performance evidence $g_{i,t}$ as the exposure-weighted residual against running Laplace baseline $b_t = (1 + \sum_{\ell=1}^{t-1} y_\ell)/(t+1)$:

$$g_{i,t} = e_{i,t}(y_t - b_t) = \left(\frac{1}{M_t}\sum_{m=1}^{M_t} \mathbb{I}[s_i \in \mathcal{I}_{t,m}]\right)\left(y_t - \frac{1 + \sum_{\ell=1}^{t-1} y_\ell}{t+1}\right). \tag{4}$$

Consequently, a positive $g_{i,t}$ indicates that invoking $s_i$ correlates with above-baseline execution success across diverse interactive tasks, whereas a negative $g_{i,t}$ mathematically flags potential procedural interference or execution noise throughout continuous lifelong autonomous agent deployment.

### 4.2 LIFECYCLE SIGNAL ESTIMATION

SkillAge parameterizes accumulated feedback into six signals forming state vector $\mathbf{x}_{i,t}$. Given sliding window $W$ with horizon $w_t = \min[W, t]$, invocation frequency is $u_{i,t} = w_t^{-1} \sum_{\ell=t-w_t+1}^{t} \chi_{i,\ell}$. To prevent unobserved routines from unjustified penalties, invocation-conditioned success adopts Laplace smoothing: $p_{i,t} = (1 + \sum_{\ell=1}^{t} \chi_{i,\ell} y_\ell)/(2 + \sum_{\ell=1}^{t} \chi_{i,\ell})$. Cumulative step exposure is

recorded as $n_{i,t} = \sum_{\ell=1}^{t} e_{i,\ell}$, yielding exposure-weighted contribution evidence:

$$\hat{c}_{i,t} = \frac{\sum_{\ell=1}^{t} g_{i,\ell}}{1 + n_{i,t}} = \frac{\sum_{\ell=1}^{t} e_{i,\ell}(y_\ell - b_\ell)}{1 + n_{i,t}} \in [-1, 1]. \tag{5}$$

Observation scarcity is $\xi_{i,t} = (1 + n_{i,t})^{-1/2}$ as buffer, redundancy is $d_{i,t} = \max_{j \neq i}(1 + \mathbf{h}_i^\top \mathbf{h}_j)/2$ via normalized embeddings $\mathbf{h}$ (zero if isolated), and context cost is $\kappa_{i,t} = \ell[s_i] / \max_{s_j \in \bar{\mathcal{R}}_t} \ell[s_j]$ normalized by library $\bar{\mathcal{R}}_t$. To analyze $\hat{c}_{i,t}$ under non-stationary environments, we formulate the evidence sequence under conditional filtration (Hoeffding, 1963):

**Assumption 1** (Piecewise Local Evidence Regularity)**.** For skill $s_i$, conditional on realized exposure sequence $\mathbf{e}_{i,1:t}$, instantaneous evidence variables $\{g_{i,\ell}\}_{\ell=1}^{t}$ are conditionally independent with local expectation $\mathbb{E}[g_{i,\ell} \mid \mathbf{e}_{i,1:t}] = e_{i,\ell}\mu_c$, where latent utility $\mu_c \in [-1, 1]$. Their conditional variances are bounded by squared exposures: $\mathbb{E}[(g_{i,\ell} - e_{i,\ell}\mu_c)^2 \mid \mathbf{e}_{i,1:t}] \leq \sigma^2 e_{i,\ell}^2$ with $\sigma^2 \leq 1$.

**Lemma 1** (Shrinkage, Boundedness, and Estimation Consistency)**.** *Under Assumption 1, $\hat{c}_{i,t}$ is strictly bounded in $(-1, 1)$, exhibits regularized shrinkage toward zero with conditional bias $\left|\mathbb{E}[\hat{c}_{i,t} \mid \mathbf{e}_{i,1:t}] - \mu_c\right| = \frac{|\mu_c|}{1+n_{i,t}}$, and achieves vanishing mean squared error bounded by:*

$$\mathbb{E}\left[(\hat{c}_{i,t} - \mu_c)^2 \;\middle|\; \mathbf{e}_{i,1:t}\right] \leq \frac{\mu_c^2}{(1 + n_{i,t})^2} + \frac{\sigma^2}{1 + n_{i,t}} = \mathcal{O}\left(\frac{1}{n_{i,t}}\right). \tag{6}$$

Appendix A.2 provides the formal proof. Lemma 1 ensures that for low-frequency specialized skills, $\hat{c}_{i,t}$ shrinks toward zero rather than hallucinating negative penalties, safeguarding rare tools from premature obsolescence.

### 4.3 Evidence-Grounded Retention Scoring

To prevent premature eviction, SkillAge decouples positive operational evidence from architectural aging pressures. We first define prospective utility $\widetilde{U}_{i,t} = ([\hat{c}_{i,t}]_+ + \xi_{i,t})/2 \in [0, 1]$, balancing empirical efficacy $[\hat{c}_{i,t}]_+ = \max(\hat{c}_{i,t}, 0)$ against an epistemic buffer $\xi_{i,t}$ that temporarily shields rare routines. Positive retention evidence and negative aging pressure aggregate opposing factors via $R_{i,t}^+ = (\widetilde{U}_{i,t} + u_{i,t} + p_{i,t})/3$ and $R_{i,t}^- = (d_{i,t} + \kappa_{i,t} + [-\hat{c}_{i,t}]_+)/3$, respectively. As shown in Figure 3, the boundary $R_{i,t}^+ = R_{i,t}^-$ demarcates operational regimes, isolating underperforming candidates while safeguarding unobserved tools. The evidence-grounded retention score $\rho_{i,t}$ balances these signals linearly:

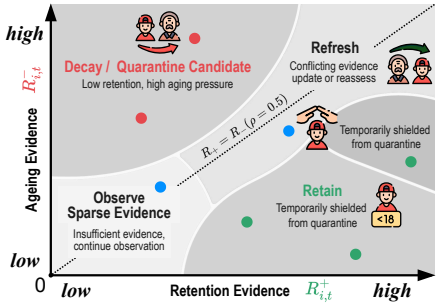


Figure 3: Decision landscape of the retention scoring mechanism. Skills are partitioned into operational regimes by $R_{i,t}^+$ and $R_{i,t}^-$, demarcated by the diagonal balance $R_{i,t}^+ = R_{i,t}^-$.

$$\rho_{i,t} = \frac{1 + R_{i,t}^+ - R_{i,t}^-}{2} = \frac{1}{2} + \frac{1}{6}\left(\widetilde{U}_{i,t} + u_{i,t} + p_{i,t} - d_{i,t} - \kappa_{i,t} - [-\hat{c}_{i,t}]_+\right) \in [0, 1]. \tag{7}$$

**Proposition 1** (Boundedness and Coordinate-wise Monotonicity)**.** *For normalized inputs in $[0, 1]$, $\rho_{i,t} \in [0, 1]$ holds universally. Furthermore, $\rho_{i,t}$ is strictly increasing in invocation frequency $u_{i,t}$ and performance $p_{i,t}$ (rate $+1/6$), strictly decreasing in redundancy $d_{i,t}$ and context cost $\kappa_{i,t}$ (rate $-1/6$), and globally non-decreasing in evidence $\hat{c}_{i,t}$.*

Derivations and Clarke subdifferential analyses are provided in Appendix B.1. Crucially, $\rho_{i,t}$ establishes an admissibility barrier rather than directly triggering deletion: quarantine action $a^q$ is admissible for $s_i$ if and only if prior exposure exists ($n_{i,t} > 0$), aging pressure dominates ($R_{i,t}^- > R_{i,t}^+$), and protection is absent ($m_{i,t} \neq m^p$). Violating any condition masks $a^q$ during sampling, analytically protecting edge-case routines:

**Corollary 1** (Epistemic Immunity of Specialized Skills)**.** *Let $s_i$ be an infrequent skill with $\hat{c}_{i,t} \geq 0$, and define net aging pressure $\Omega_{i,t} := 2(d_{i,t} + \kappa_{i,t}) - 2(u_{i,t} + p_{i,t}) - \hat{c}_{i,t}$. If $\Omega_{i,t} \leq 0$, quarantine action $a^q$ is unconditionally inadmissible. If $\Omega_{i,t} \in (0, 1)$, $a^q$ remains provably inadmissible whenever*

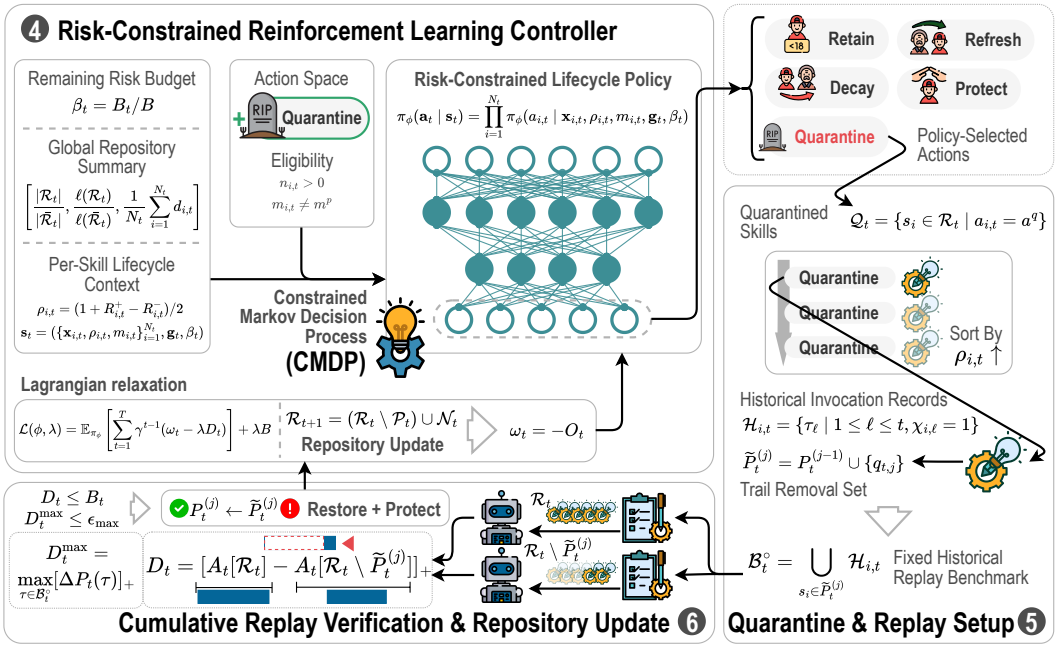


Figure 4: Overview of the risk-constrained RL controller and closed-loop replay verification in SkillAge. Quarantined skills from the factorized CMDP policy undergo sequential dual-gate replay screening on historical benchmarks before committing irreversible physical deletion.

*cumulative exposure satisfies:*

$$n_{i,t} < \Omega_{i,t}^{-2} - 1 = \left[2(d_{i,t} + \kappa_{i,t}) - 2(u_{i,t} + p_{i,t}) - \hat{c}_{i,t}\right]^{-2} - 1. \tag{8}$$

Corollary 1 guarantees that scarcity buffer $\xi_{i,t} = (1 + n_{i,t})^{-1/2}$ structurally prevents premature eviction until adequate interaction evidence accumulates.

### 4.4 Risk-Constrained Reinforcement Learning Controller

As depicted in Figure 4, we formalize repository governance as a finite-horizon CMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^{N_t}, \mathcal{P}, \omega, \mathcal{D}, \gamma \rangle$ (Altman, 2021) to balance individual skill utility against global repository overhead while gating physical pruning through closed-loop replay certification.

**Factorized State-Action Formulation and Policy Parameterization:** At step $t$, the controller observes context $\mathbf{s}_t = (\{(\mathbf{x}_{i,t}, \rho_{i,t}, m_{i,t})\}_{i=1}^{N_t}, \mathbf{g}_t, \beta_t)$. To ensure permutation-invariant decentralized decision-making, each skill $s_i$ is mapped to a unified feature representation $\tilde{\mathbf{s}}_{i,t} = [\mathbf{x}_{i,t}^\top, \rho_{i,t}, \mathbf{e}(m_{i,t})^\top, \mathbf{g}_t^\top, \beta_t]^\top \in \mathbb{R}^{d_s}$, where $\mathbf{e}(m_{i,t})$ is the one-hot embedding of mode $m_{i,t} \in \{m^a, m^d, m^p\}$. The action vector $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N_t,t})$ factorizes over primitives $\mathcal{A} = \{a^r, a^d, a^f, a^p, a^q\}$. To guarantee safe exploration, the policy network $h_\phi(\tilde{\mathbf{s}}_{i,t}, \cdot)$ enforces retention barriers via action masking, restricting candidates to admissible support $\mathcal{A}_{\text{adm}}(s_i) \subseteq \mathcal{A}$:

$$\pi_\phi(\mathbf{a}_t \mid \mathbf{s}_t) = \prod_{i=1}^{N_t} \pi_\phi(a_{i,t} \mid \tilde{\mathbf{s}}_{i,t}) = \prod_{i=1}^{N_t} \frac{\exp\left(h_\phi(\tilde{\mathbf{s}}_{i,t}, a_{i,t})\right) \cdot \mathbb{I}[a_{i,t} \in \mathcal{A}_{\text{adm}}(s_i)]}{\sum_{a' \in \mathcal{A}_{\text{adm}}(s_i)} \exp\left(h_\phi(\tilde{\mathbf{s}}_{i,t}, a')\right)}. \tag{9}$$

If barrier conditions ($n_{i,t} > 0$, $R_{i,t}^- > R_{i,t}^+$, $m_{i,t} \neq m^p$) fail, $\mathbb{I}[a^q \in \mathcal{A}_{\text{adm}}(s_i)] = 0$, rendering premature quarantine actions structurally inadmissible.

**Primal-Dual Objective and Two-Timescale Optimization:** The agent maximizes cumulative compaction reward $\omega_t = -O_t$ subject to cumulative historical degradation budget $B$. Denoting discounted performance return as $J_\omega(\pi_\phi) = \mathbb{E}_{\pi_\phi}[\sum_{t=1}^{T} \gamma^{t-1} \omega_t]$ and degradation cost as

$J_D(\pi_\phi) = \mathbb{E}_{\pi_\phi}[\sum_{t=1}^{T} \gamma^{t-1} D_t]$, we cast the constrained problem into an unconstrained saddle-point formulation via the Lagrangian functional over finite-horizon lifelong agent interaction steps:

$$\min_{\lambda \geq 0} \max_{\phi} \; \mathcal{L}(\phi, \lambda) \coloneqq J_\omega(\pi_\phi) - \lambda\big(J_D(\pi_\phi) - B\big), \tag{10}$$

where $\lambda \in \mathbb{R}_+$ adaptively penalizes safety violations. Under the Constrained Policy Gradient Theorem (Achiam et al., 2017), the primal policy gradient evaluates to:

$$\nabla_\phi \mathcal{L}(\phi, \lambda) = \mathbb{E}_{\pi_\phi}\left[\sum_{t=1}^{T} \gamma^{t-1} \left(\sum_{i=1}^{N_t} \nabla_\phi \log \pi_\phi(a_{i,t} \mid \tilde{\mathbf{s}}_{i,t})\right) A_{\mathcal{L}}^{\pi_\phi}(\mathbf{s}_t, \mathbf{a}_t; \lambda)\right], \tag{11}$$

where $A_{\mathcal{L}}^{\pi_\phi} \coloneqq A_\omega^{\pi_\phi} - \lambda A_D^{\pi_\phi}$ is estimated via Generalized Advantage Estimation (Schulman et al., 2015). The saddle point is resolved via alternating two-timescale stochastic approximation:

$$\phi_{k+1} = \phi_k + \eta_\phi^{(k)} \widehat{\nabla}_\phi \mathcal{L}(\phi_k, \lambda_k), \qquad \lambda_{k+1} = \big[\lambda_k + \eta_\lambda^{(k)} (\widehat{J}_D(\pi_{\phi_k}) - B)\big]_+, \tag{12}$$

where $[\cdot]_+ = \max[\cdot, 0]$ projects the multiplier onto the non-negative orthant.

**Proposition 2** (Asymptotic Convergence and Constraint Feasibility)**.** *Under Robbins-Monro conditions* $\sum_k \eta_\phi^{(k)} = \sum_k \eta_\lambda^{(k)} = \infty$, $\sum_k[(\eta_\phi^{(k)})^2 + (\eta_\lambda^{(k)})^2] < \infty$ *and timescale separation* $\lim_{k\to\infty} \eta_\lambda^{(k)}/\eta_\phi^{(k)} = 0$*, the iterates* $(\phi_k, \lambda_k)$ *from Eq.* (12) *converge almost surely to a local saddle point* $(\phi^*, \lambda^*)$ *of* $\mathcal{L}(\phi, \lambda)$*, and* $\pi_{\phi^*}$ *satisfies constraint* $J_D(\pi_{\phi^*}) \leq B$*.*

The ODE-based convergence proof is provided in Appendix C.5 for interested and thorough readers.

### 4.5 Quarantine Replay and Repository Update

To prevent capability collapse, quarantined skills $\mathcal{Q}_t = \{s_i \in \mathcal{R}_t \mid a_{i,t} = a^q\}$ are staged and sorted ascendingly by retention score $\rho_{i,t}$. Candidates are certified on their joint historical benchmark $\mathcal{B}_t^\circ = \bigcup_{s_j \in \mathcal{Q}_t} \mathcal{H}_{i,t}$, where $\mathcal{H}_{i,t} = \{\tau_\ell \mid 1 \leq \ell \leq t, \chi_{i,\ell} = 1\}$. To capture non-linear synergies and fallback dependencies, we greedily construct the deletion set $\mathcal{P}_t$ (initialized to $\varnothing$). A candidate $q_{t,j}$ is accepted into $\widetilde{\mathcal{P}}_t = \mathcal{P}_t \cup \{q_{t,j}\}$ if and only if deterministic replay on $\mathcal{B}_t^\circ$ satisfies dual safety gates:

$$\Delta_t^{\text{avg}}[\widetilde{\mathcal{P}}_t] = \big[A_t[\mathcal{R}_t] - A_t[\mathcal{R}_t \setminus \widetilde{\mathcal{P}}_t]\big]_+ \leq B_t, \tag{13}$$

$$\Delta_t^{\max}[\widetilde{\mathcal{P}}_t] = \max_{\tau \in \mathcal{B}_t^\circ} \big[P[\tau; \mathcal{R}_t] - P[\tau; \mathcal{R}_t \setminus \widetilde{\mathcal{P}}_t]\big]_+ \leq \epsilon_{\max}, \tag{14}$$

where $A_t[\mathcal{R}] = |\mathcal{B}_t^\circ|^{-1} \sum_{\tau \in \mathcal{B}_t^\circ} P[\tau; \mathcal{R}]$ is the mean task fidelity. Unverified candidates $\mathcal{Q}_t \setminus \mathcal{P}_t$ are restored to $\mathcal{R}_t$ in protected mode ($m^p$). Given verified degradation $D_t = \Delta_t^{\text{avg}}[\mathcal{P}_t]$ and newly added skills $\mathcal{N}_t$, the repository and budget update as $\mathcal{R}_{t+1} = (\mathcal{R}_t \setminus \mathcal{P}_t) \cup \mathcal{N}_t$ and $B_{t+1} = B_t - D_t$.

**Theorem 1** (Dual-Gate Removal Guarantee)**.** *With* $B_1 = B$ *and* $B_{t+1} = B_t - D_t$*, the sequential quarantine protocol provably ensures cumulative historical replay degradation is bounded by* $\sum_{t=1}^{H} D_t \leq B - B_{H+1} \leq B$*, while uniformly bounding worst-case single-task capability collapse:* $\max_{t \in [H]} \Delta_t^{\max}[\mathcal{P}_t] \leq \epsilon_{\max}$ *across any arbitrary deployment horizon* $H \leq T$*.*

See Appendix C.4 for the proof. Grounded in experience replay (Rolnick et al., 2019), Theorem 1 guarantees historical competence preservation in stationary settings, while out-of-distribution robustness is sustained via continual exploration.

## 5 Experiments

**Setup and Baselines.** We evaluate SkillAge on SkillFlow (Zhang et al., 2026d) across Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 ($N_0 = 158, K = 10$) under continuous lifelong agent interaction streams. We benchmark against frequency/recency cache eviction (Random, LRU, LFU), heuristic lifecycle decay (ASAP), external memory baselines (Combined Deletion (Xiong et al., 2026), AgeMem (Yu et al., 2026)), and an unconstrained Oracle. Metrics evaluate task utility, token reduction, compression, and certified degradation. See Appendix D for details.

Table 1: Lifecycle comparison across Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 Codex ($K = 10$, dynamic exposure). Comp: compression (%); †lookahead Oracle bound.

| Method | Qwen 3.7-plus | | | | DeepSeek-V4-Flash | | | | GPT-5.5 (Codex) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Util. ↑ | Token ↑ | Comp. ↑ | Loss ↓ | Util. ↑ | Token ↑ | Comp. ↑ | Loss ↓ | Util. ↑ | Token ↑ | Comp. ↑ | Loss ↓ |
| Random | 0.142 | 0.412 | 12.5 | 0.858 | 0.155 | 0.428 | 13.1 | 0.845 | 0.171 | 0.445 | 14.2 | 0.829 |
| LRU | 0.169 | 0.627 | 48.2 | 0.831 | 0.165 | 0.618 | 46.5 | 0.835 | 0.173 | 0.635 | 49.8 | 0.827 |
| LFU | 0.181 | 0.650 | 51.4 | 0.819 | 0.174 | 0.642 | 50.1 | 0.826 | 0.188 | 0.661 | 53.0 | 0.812 |
| ASAP | 0.210 | 0.712 | 58.6 | 0.790 | 0.202 | 0.701 | 57.2 | 0.798 | 0.218 | 0.725 | 60.1 | 0.782 |
| Combined Del. | 0.325 | 0.824 | 78.5 | 0.675 | 0.318 | 0.815 | 76.8 | 0.682 | 0.332 | 0.831 | 79.4 | 0.668 |
| AgeMem | 0.602 | 0.865 | 82.1 | 0.398 | 0.592 | 0.856 | 80.5 | 0.408 | 0.615 | 0.872 | 83.4 | 0.385 |
| **SkillAge (Ours)** | **0.681** | **0.912** | **88.4** | **0.319** | **0.672** | **0.901** | **87.2** | **0.328** | **0.695** | **0.920** | **89.5** | **0.305** |
| Oracle† (Upper Bound) | 0.986 | 0.931 | 93.7 | 0.014 | 0.982 | 0.925 | 93.0 | 0.018 | 0.990 | 0.938 | 94.2 | 0.010 |

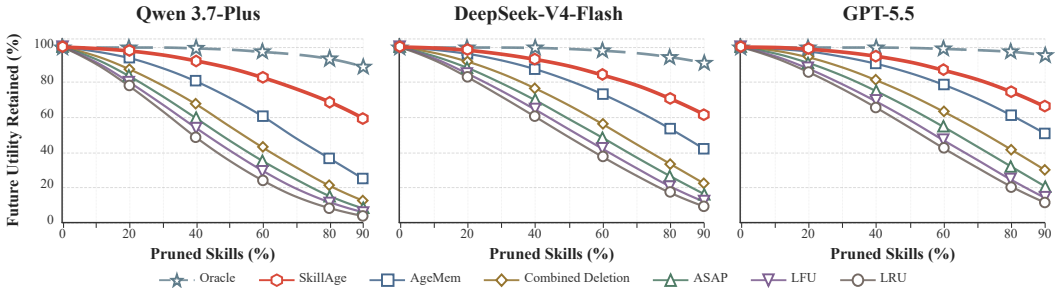


Figure 5: Future utility preservation under different skill pruning ratios. SkillAge consistently retains higher future task performance compared to existing baselines across various pruning budgets.

**Overall Performance and Pruning Robustness.** As shown in Table 1 and Figure 5, SkillAge consistently surpasses all baselines across pruning ratios. While conventional heuristics suffer severe capability collapse by irrevocably evicting indispensable rare tools, SkillAge robustly maintains execution fidelity across all model backbones. By pairing regularized lifecycle estimation with quarantine replay, it eliminates obsolete routines while shielding capabilities from forgetting.

**Compression-Capability Trade-off Frontier.** Figure 6 illustrates the Pareto frontier between repository compression and empirical capability loss. SkillAge forms a dominant trade-off frontier, achieving up to 89.5% compression and 92.0% token reduction within the safety budget. Unlike heuristic methods that incur catastrophic regression, SkillAge navigates this dilemma seamlessly, validating that two-timescale primal-dual optimization suppresses redundant procedural overhead.

**Skill Selection Quality under Exposure Budgets.** Figure 7 benchmarks comprehensive skill selection quality across diverse active retrieval budgets $K$. At $K = 10$, SkillAge consistently attains superior downstream task utility capture and useful routine recall over recency- and frequency-based baselines while closely matching global library precision. This demonstrates that our evidence-grounded scoring reliably filters out procedural noise and dormant redundancies, successfully providing a dense, highly relevant operational context even when agent prompt retrieval capacity is tightly constrained under strict resource limits.

**Ablation Study.** Figure 8 along with the complete architectural breakdown in Table 6 in Appendix E.4 isolate the empirical contribution of each core component in SkillAge. The unified framework attains superior utility and compression while strictly respecting replay degradation budgets without safety breaches. Removing quarantine replay triggers catastrophic degradation and budget violations, while omitting retention barriers or epistemic buffers exposes rare capabilities to premature eviction, confirming each module's necessity.

**Policy Convergence and Safety Dynamics.** Figure 9 in Appendix B traces the learning dynamics of our two-timescale primal-dual controller over 1,500 iterations across multiple diverse model

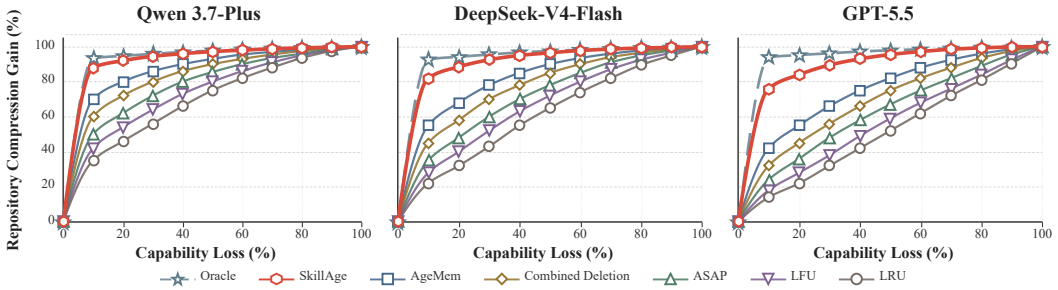


Figure 6: The compression-capability trade-off frontier under different agent trajectories. SkillAge achieves superior repository compression while strictly bounding the overall capability loss across the Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 backbones.

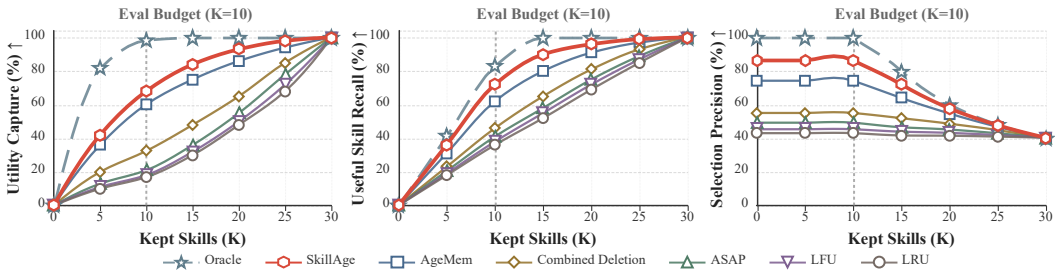


Figure 7: Skill selection quality and retrieval efficiency across varying active context budgets ($K$), evaluating utility capture, useful skill recall, and selection precision.

backbones. While the primal policy rapidly converges toward an overhead-minimizing regime under stable stochastic exploration, the adaptive multiplier $\lambda$ dynamically ascends to penalize nascent degradation spikes before stabilizing smoothly. This observed empirical behavior directly validates Proposition 2, robustly confirming asymptotic saddle-point stability, zero safety breaches, and strict constraint feasibility throughout long-horizon lifelong deployment.

## 6 Conclusion

Unchecked skill accumulation inevitably degrades retrieval precision and execution fidelity in long-horizon autonomous agents. To resolve this challenge, we introduced SkillAge, a principled framework for lifecycle-aware agent skill governance. Formulating procedural pruning as a CMDP, SkillAge couples a two-timescale primal-dual policy with closed-loop quarantine replay certification, effectively eliminating redundant routines while strictly bounding capability regression. Extensive evaluations across Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 trajectories demonstrate that SkillAge achieves up to an 89.5% repository compression rate and a 92.0% prompt token reduction while rigorously satisfying safety budgets. This work establishes safety-certified lifecycle governance as an essential foundation for sustainable agent evolution.

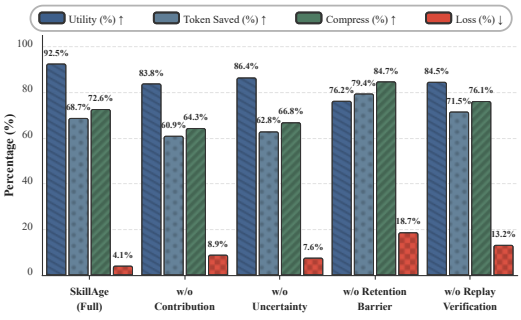


Figure 8: Ablation study of the key SkillAge lifecycle components. Omitting the retention barrier or quarantine replay verification leads to severe capability loss despite achieving higher superficial repository compression.

## REFERENCES


Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. Pmlr, 2017.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Dongkyu Cho, Taesup Moon, Rumi Chunara, Kyunghyun Cho, and Sungmin Cha. Forget forgetting: Continual learning in a world of abundant memory. *arXiv preprint arXiv:2502.07274*, 2025.

Elias Dritsas and Maria Trigka. Deployment-aware compression of large language models for edge intelligence: A system-level survey. *IEEE Transactions on Artificial Intelligence*, 2026.

Jiawei Du, Jinlong Wu, Yuzheng Chen, Yucheng Hu, Bing Li, and Joey Tianyi Zhou. Rethinking agent design: From top-down workflows to bottom-up skill evolution. *arXiv preprint arXiv:2505.17673*, 2025.

Linfeng Fan, Yuan Tian, Ziwei Li, and Zhiwu Lu. Skill drift is contract violation: Proactive maintenance for llm agent skill libraries, 2026.

Yujie Feng, Hao Wang, Jian Li, Xu Chu, Zhaolu Kang, Yiran Liu, Yasha Wang, Philip S Yu, and Xiao-Ming Wu. Forever: Forgetting curve-inspired memory replay for language model continual learning. *arXiv preprint arXiv:2601.03938*, 2026.

Yudong Gao, Zongjie Li, Yuanyuan Yuan, Zimo Ji, Pingchuan Ma, and Shuai Wang. Skillreducer: Optimizing llm agent skills for token efficiency, 2026.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025.

Dingcheng Huang, Yuda Ding, Bingshuo Liu, Qingbin Liu, Xi Chen, Jiang Bian, Hongliang Sun, Zhiying Tu, Dianhui Chu, Xiaoyan Yu, et al. Skillwiki: A living knowledge infrastructure for agent skills. *arXiv preprint arXiv:2606.16523*, 2026.

Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, Shuyi Wang, Qunhong Zeng, Di Wang, Xuandong Zhao, Yuanli Wang, Roey Ben Chaim, Zonglin Di, Yipeng Gao, Junwei He, Yizhuo He, Liqiang Jing, Luyang Kong, Xin Lan, Jiachen Li, Songlin Li, Yijiang Li, Yueqian Lin, Xinyi Liu, Xuanqing Liu, Haoran Lyu, Ze Ma, Bowei Wang, Runhui Wang, Tianyu Wang, Wengao Ye, Yue Zhang, Hanwen Xing, Yiqi Xue, Steven Dillmann, and Han-chung Lee. Skillsbench: Benchmarking how well agent skills work across diverse tasks, 2026.

Yubo Li. Dynamic agent skills: A lifecycle survey and taxonomy of evolving skill libraries. *arXiv preprint arXiv:2607.10113*, 2026.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, Qingchen Yu, Jihao Zhao, Yezhaohui Wang, Peng Liu, Zehao Lin, Pengyuan Wang, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhen Tao, Huayi Lai, Hao Wu, Bo Tang, Zhengren Wang, Zhaoxin Fan, Ningyu Zhang, Linfeng Zhang, Junchi Yan, Mingchuan Yang, Tong Xu, Wei Xu, Huajun Chen, Haofen Wang, Hongkang Yang, Wentao Zhang, Zhi-Qin John Xu, Siheng Chen, and Feiyu Xiong. Memos: A memory os for ai system, 2025. URL https://arxiv.org/abs/2507.03724.

Huawei Lin, Peng Li, Jie Song, Fuxin Jiang, and Tieying Zhang. Muse-autoskill: Self-evolving agents via skill creation, memory, management, and evaluation. *arXiv preprint arXiv:2605.27366*, 2026.

Reza Yousefi Maragheh and Yashar Deldjoo. The future is agentic: Definitions, perspectives, and open challenges of multi-agent recommender systems. *arXiv preprint arXiv:2507.02097*, 2025.

Siru Ouyang, Jun Yan, Yanfei Chen, Rujun Han, Zifeng Wang, Bhavana Dalvi Mishra, Rui Meng, Chun-Liang Li, Yizhu Jiao, Kaiwen Zha, Maohao Shen, Vishy Tirumalashetty, George Lee, Jiawei Han, Tomas Pfister, and Chen-Yu Lee. Skillos: Learning skill curation for self-evolving agents, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Joseph E. Gonzalez, and Ion Stoica. Memgpt: Towards llms as operating systems, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

Hongji Pu, Xinyuan Song, and Liang Zhao. Skillops: Managing llm agent skill libraries as self-maintaining software ecosystems, 2026.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Junhao Shen, Teng Zhang, Xiaoyan Zhao, and Hong Cheng. Dynamic skill lifecycle management for agentic reinforcement learning. *arXiv preprint arXiv:2605.10923*, 2026.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long horizon execution in llms. *arXiv preprint arXiv:2509.09677*, 2025.

Hamed Taheri. Governed memory: A production architecture for multi-agent workflows. *arXiv preprint arXiv:2603.17787*, 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 623–645, San Diego, California, United States, July 2026. ISBN 979-8-89176-390-6.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2022.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Jiang Zhang, Bing Yuan, and Qian Zhang. Self-reference in large language models: The introspection threshold for recursive self-improvement. *arXiv preprint arXiv:2607.04277*, 2026a.

Xing Zhang, Yanwei Cui, Guanghui Wang, Ziyuan Li, Wei Qiu, Bing Zhu, and Peiyang He. Library drift: Diagnosing and fixing a silent failure mode in self-evolving llm skill libraries, 2026b.

Zhiwei Zhang, Yudi Lin, Nikki Lijing Kuang, Linlin Wu, Xiaomin Li, Songtao Liu, and Fenglong Ma. Co-evolving skill generation and policy optimization. *arXiv preprint arXiv:2606.08755*, 2026c.

Ziao Zhang, Kou Shi, Shiting Huang, Avery Nie, Yu Zeng, Yiming Zhao, Zhen Fang, Qisheng Su, Haibo Qiu, Wei Yang, Qingnan Ren, Shun Zou, Wenxuan Huang, Lin Chen, Zehui Chen, and Feng Zhao. Skillflow: Benchmarking lifelong skill discovery and evolution for autonomous agents, 2026d.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners, 2023.

Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025a.

Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, Zhongzhi Li, Yingying Zhang, Le Song, and Qianli Ma. Lifelongagentbench: Evaluating llm agents as lifelong learners, 2025b.

# ORGANIZATION OF APPENDICES


**A Lifecycle-Aware Contribution and Statistical Properties** **14**

A.1 Observational Filtration Setup and Moment Conditions . . . . . 14

A.2 Proof of Lemma 1 (Shrinkage, Boundedness, and Estimation Consistency) . . . . . 15

A.3 Zero-Exposure Dynamics and Cold-Start Boundary Behavior . . . . . 16

A.4 Dynamic Adaptation under Non-Stationary Lifelong Task Streams . . . . . 16

**B Evidence-Grounded Geometry and Epistemic Immunity** **16**

B.1 Proof of Proposition 1 (Boundedness and Coordinate-wise Monotonicity) . . . . . 16

B.2 Clarke Subdifferential Analysis of Empirical Performance Evidence . . . . . 17

B.3 Epistemic Protection Dynamics and Scarcity Sensitivity . . . . . 18

B.4 Proof of Corollary 1 (Epistemic Quarantine Immunity Bounds) . . . . . 18

**C Dual-Gate Replay Safety Certification and CMDP Convergence** **20**

C.1 Dual-Gate Quarantine Replay Protocol (Algorithm 1) . . . . . 20

C.2 Master Lifelong Governance and Execution Loop (Algorithm 2) . . . . . 20

C.3 Combinatorial Synergies: Cumulative vs. Isolated Removal Dynamics . . . . . 20

C.4 Proof of Theorem 1 (Dual-Gate Replay-Certified Removal Guarantee) . . . . . 21

C.5 Proof of Proposition 2 (Two-Timescale CMDP Controller Convergence) . . . . . 22

**D Experimental Protocol and Implementation Details** **23**

D.1 Benchmark Construction and Trajectory Unrolling . . . . . 23

D.2 Lifecycle Controller Configuration . . . . . 24

D.3 Baseline Implementations and Adaptation Details . . . . . 24

**E Extended Empirical Results and Complexity Accounting** **24**

E.1 Tri-Model Performance Breakdown across Task Domains . . . . . 25

E.2 Per-Round Asymptotic Complexity Breakdown . . . . . 25

E.3 Empirical Replay Compute and Resource Accounting . . . . . 26

E.4 Ablation Analysis across Architectural Modules . . . . . 26

**F The Use of Large Language Models (LLMs)** **27**

# OVERVIEW OF SUPPLEMENTARY APPENDICES

This supplementary document provides complete theoretical derivations, rigorous proofs, extended algorithmic frameworks, and exhaustive empirical specifications supporting the main manuscript: specifically, Appendix A formalizes the observational filtration governing online agent interaction, establishes localized moment bounds, and proves the uniform boundedness, shrinkage bias, and mean squared error consistency of the contribution estimator formalized in Lemma 1; Appendix B characterizes the coordinate-wise monotonicity of the retention scoring landscape established in Proposition 1, investigates its non-smooth Clarke subdifferentials across contribution regimes, and proves the epistemic immunity guarantees derived in Corollary 1; Appendix C presents algorithmic specifications for both the dual-gate quarantine verification gateway formalized in Algorithm 1 and the overarching lifelong governance loop delineated in Algorithm 2, accompanied by the formal inductive proof of cumulative budget preservation established in Theorem 1 and the two-timescale stochastic approximation convergence proof established in Proposition 2; Appendices D and E delineate the SkillFlow lifelong benchmark construction, hyperparameter assignments, baseline adaptations, asymptotic complexity accounting, and extended tri-model performance breakdowns; and finally, Appendix F provides the mandatory declaration regarding the use of AI assistance.

# A LIFECYCLE-AWARE CONTRIBUTION AND STATISTICAL PROPERTIES

This appendix provides the formal statistical framework and complete theoretical proofs characterizing the exposure-weighted performance evidence estimator $\hat{c}_{i,t}$ defined in Eq. (5). In continuous lifelong learning, an autonomous agent executes long-horizon interactive decision trajectories where multiple procedural routines are concurrently invoked. Under collaborative multi-skill execution, isolating the empirical credit of an individual tool is fundamentally vulnerable to trajectory-level confounding. To address this challenge, SkillAge formalizes credit assignment through an exposure-conditioned filtration tracking relative task residuals against an adaptive baseline.

## A.1 OBSERVATIONAL FILTRATION SETUP AND MOMENT CONDITIONS

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the foundational probability space governing the agent's interactive learning process across the task sequence $\mathcal{T} = \{\tau_t\}_{t=1}^{T}$ during lifelong deployment. For any arbitrary procedural skill $s_i \in \mathcal{R}_t$, its complete historical sequence of step-level activations, within-task exposures, and domain verifier outcomes is formalized via a discrete-time observational filtration $\mathbb{F}_i = \{\mathcal{F}_{i,t}\}_{t\geq 1}$, where $\mathcal{F}_{i,t} \coloneqq \sigma(\{e_{i,\ell}, y_\ell, b_\ell\}_{\ell=1}^{t})$. Here, $e_{i,\ell} \in [0,1]$ represents the normalized within-task invocation frequency of skill $s_i$ during round $\ell$, $y_\ell \in [0,1]$ registers the holistic trajectory fidelity score emitted by the automated verifier, and $b_\ell = (1 + \sum_{r=1}^{\ell-1} y_r)/(\ell+1) \in [0,1]$ represents the running Laplace-smoothed historical performance baseline.

Conditioned on the realized exposure sequence $\mathbf{e}_{i,1:t} = (e_{i,1}, \ldots, e_{i,t}) \in [0,1]^t$, the agent isolates the empirical impact of invoking $s_i$ by evaluating the exposure-weighted task residual $g_{i,\ell} = e_{i,\ell}(y_\ell - b_\ell) \in [-e_{i,\ell}, e_{i,\ell}]$. The bounded inclusion holds unconditionally because $0 \leq y_\ell, b_\ell \leq 1$, which strictly enforces $|y_\ell - b_\ell| \leq 1$. Under Assumption 1, conditional on the exposure history $\mathbf{e}_{i,1:t}$, the instantaneous evidence variables $\{g_{i,\ell}\}_{\ell=1}^{t}$ are conditionally independent across rounds, reflecting the property that execution noise is conditionally uncorrelated across distinct task environments. Consequently, the local conditional expectation evaluates directly to $\mathbb{E}[g_{i,\ell} \mid \mathbf{e}_{i,1:t}] = \mathbb{E}[e_{i,\ell}(y_\ell - b_\ell) \mid \mathbf{e}_{i,1:t}] = e_{i,\ell}\mu_c \in [-e_{i,\ell}, e_{i,\ell}]$, where the scalar parameter $\mu_c \in [-1,1]$ denotes the true latent local utility of skill $s_i$ within the current operational regime. Furthermore, because domain feedback is bounded, the conditional variance of each instantaneous evidence variable is naturally constrained by its squared exposure:

$$\mathrm{Var}\left(g_{i,\ell} \,\middle|\, \mathbf{e}_{i,1:t}\right) = e_{i,\ell}^2 \,\mathrm{Var}\left(y_\ell - b_\ell \,\middle|\, \mathbf{e}_{i,1:t}\right) \leq \sigma^2 e_{i,\ell}^2, \qquad \text{where } \sigma^2 \leq 1. \tag{15}$$

Eq. (15) demonstrates that task instances in which skill $s_i$ is only marginally touched ($e_{i,\ell} \approx 0$) contribute negligible statistical variance to the cumulative assessment, suppressing credit assignment noise induced by irrelevant tools throughout continuous lifelong interactive agent deployment.

### A.2 Proof of Lemma 1 (Shrinkage, Boundedness, and Estimation Consistency)

*Proof.* We establish the three analytical properties of the regularized ratio estimator $\hat{c}_{i,t} = (\sum_{\ell=1}^{t} g_{i,\ell})/(1 + n_{i,t})$ in a sequential manner.

**Step 1: Strict Interior Boundedness and Numerical Clipping Invariance.** Recall that cumulative step exposure for any arbitrary procedural skill is defined as $n_{i,t} = \sum_{\ell=1}^{t} e_{i,\ell} \geq 0$ across the continuous online task stream. Because both the empirical outcome $y_\ell$ and baseline $b_\ell$ reside in $[0, 1]$, the instantaneous residual satisfies $|y_\ell - b_\ell| \leq \max(y_\ell, b_\ell) \leq 1$. Applying the triangle inequality to the cumulative evidence summation in the numerator reveals that $|\sum_{\ell=1}^{t} g_{i,\ell}| \leq \sum_{\ell=1}^{t} |g_{i,\ell}| = \sum_{\ell=1}^{t} e_{i,\ell}|y_\ell - b_\ell| \leq \sum_{\ell=1}^{t} e_{i,\ell} = n_{i,t}$. Dividing both sides of this relation by the regularized exposure denominator $(1 + n_{i,t})$ produces a contractive bound on the estimator magnitude:

$$|\hat{c}_{i,t}| = \frac{|\sum_{\ell=1}^{t} g_{i,\ell}|}{1 + n_{i,t}} \leq \frac{n_{i,t}}{1 + n_{i,t}} = 1 - \frac{1}{1 + n_{i,t}} < 1. \tag{16}$$

Because cumulative exposure is non-negative ($n_{i,t} \geq 0$) across all rounds, the ratio $n_{i,t}/(1 + n_{i,t})$ strictly approaches 1 from below only as exposure tends to infinity. For any finite interaction sequence $t < \infty$, the estimator $\hat{c}_{i,t}$ is strictly confined to the compact interior sub-interval $\hat{c}_{i,t} \in [-n_{i,t}/(1+n_{i,t}),\ n_{i,t}/(1+n_{i,t})] \subset (-1, 1) \subseteq [-1, 1]$. This confirms that an explicit numerical clipping operation is structurally inactive almost surely. The regularized denominator enforces analytical stability without introducing artificial boundary saturation or discontinuous gradient artifacts during policy optimization across continuous lifelong interactive agent deployments.

**Step 2: Regularized Shrinkage and Vanishing Conditional Bias.** To examine the statistical expectation of $\hat{c}_{i,t}$ under finite exposures, we take the conditional expectation given the exposure history $\mathbf{e}_{i,1:t}$. Utilizing the linearity of conditional expectations and substituting the localized moment condition $\mathbb{E}[g_{i,\ell} \mid \mathbf{e}_{i,1:t}] = e_{i,\ell}\mu_c$ established under piecewise local evidence regularity:

$$\mathbb{E}\big[\hat{c}_{i,t} \;\big|\; \mathbf{e}_{i,1:t}\big] = \frac{\sum_{\ell=1}^{t} \mathbb{E}[g_{i,\ell} \mid \mathbf{e}_{i,1:t}]}{1 + n_{i,t}} = \frac{\sum_{\ell=1}^{t} e_{i,\ell}\mu_c}{1 + n_{i,t}} = \frac{\mu_c n_{i,t}}{1 + n_{i,t}}. \tag{17}$$

Subtracting the target latent parameter $\mu_c$ directly from Eq. (17) yields the analytical expression for the finite-sample conditional bias, $\mathrm{Bias}(\hat{c}_{i,t} \mid \mathbf{e}_{i,1:t}) = \mathbb{E}[\hat{c}_{i,t} \mid \mathbf{e}_{i,1:t}] - \mu_c = \mu_c n_{i,t}/(1+n_{i,t}) - \mu_c = -\mu_c/(1 + n_{i,t})$. Taking the absolute value of both sides reveals that the magnitude of the estimation bias contracts monotonically at an inverse linear rate with cumulative step exposure:

$$\big|\,\mathrm{Bias}\left(\hat{c}_{i,t} \;\big|\; \mathbf{e}_{i,1:t}\right)\big| = \frac{|\mu_c|}{1 + n_{i,t}} \leq \frac{1}{1 + n_{i,t}} = \mathcal{O}\left(\frac{1}{n_{i,t}}\right). \tag{18}$$

This contractive property demonstrates that the estimator incorporates regularized shrinkage toward zero. When cumulative exposure is scarce ($n_{i,t} \approx 0$), the estimator naturally contracts toward the neutral baseline $\hat{c}_{i,t} \approx 0$ rather than exhibiting high variance or overreacting to isolated task failures.

**Step 3: Conditional Variance Contraction and Mean Squared Error Consistency.** Because the instantaneous evidence variables $\{g_{i,\ell}\}_{\ell=1}^{t}$ are conditionally independent across rounds given $\mathbf{e}_{i,1:t}$, the conditional variance of their normalized sum expands additively: $\mathrm{Var}(\hat{c}_{i,t} \mid \mathbf{e}_{i,1:t}) = \sum_{\ell=1}^{t} \mathrm{Var}(g_{i,\ell} \mid \mathbf{e}_{i,1:t})/(1+n_{i,t})^2 \leq \sigma^2 \sum_{\ell=1}^{t} e_{i,\ell}^2/(1+n_{i,t})^2$. Since within-task exposure satisfies $e_{i,\ell} \in [0, 1]$, we have $e_{i,\ell}^2 \leq e_{i,\ell}$, which directly implies $\sum_{\ell=1}^{t} e_{i,\ell}^2 \leq \sum_{\ell=1}^{t} e_{i,\ell} = n_{i,t}$. Substituting this upper bound and noting that $n_{i,t}/(1 + n_{i,t}) < 1$:

$$\mathrm{Var}\left(\hat{c}_{i,t} \;\big|\; \mathbf{e}_{i,1:t}\right) \leq \frac{\sigma^2 n_{i,t}}{(1 + n_{i,t})^2} = \frac{\sigma^2}{1 + n_{i,t}} \left(\frac{n_{i,t}}{1 + n_{i,t}}\right) < \frac{\sigma^2}{1 + n_{i,t}}. \tag{19}$$

Finally, substituting the squared bias from Eq. (18) and the variance bound from Eq. (19) into the bias-variance decomposition of the mean squared error yields:

$$\mathbb{E}\Big[(\hat{c}_{i,t} - \mu_c)^2 \;\Big|\; \mathbf{e}_{i,1:t}\Big] \leq \frac{\mu_c^2}{(1 + n_{i,t})^2} + \frac{\sigma^2}{1 + n_{i,t}}. \tag{20}$$

As cumulative step exposure accumulates ($n_{i,t} \to \infty$), the conditional mean squared error decays asymptotically at rate $\mathcal{O}(1/n_{i,t})$, concluding the proof. □

### A.3 Zero-Exposure Dynamics and Cold-Start Boundary Behavior

A primary vulnerability in conventional cache eviction heuristics (such as LRU and LFU) is the cold-start penalty, where freshly introduced or rarely invoked tools are rapidly evicted because their recency or frequency counters remain zero; under SkillAge, this pathology is structurally eliminated through the joint analytical behavior of the contribution estimator $\hat{c}_{i,t}$ and the observation scarcity buffer $\xi_{i,t}$. Specifically, under the exact cold-start boundary condition where skill $s_i$ has never been invoked during the interaction stream up to round $t$ (such that $e_{i,\ell} = 0$ for all $\ell \in [t]$ and cumulative exposure $n_{i,t} = 0$), substituting these boundary conditions into the analytical formulations yields the exact limits $\hat{c}_{i,t}|_{n_{i,t}=0} = \sum_{\ell=1}^{t} 0 \cdot (y_\ell - b_\ell)/(1+0) = 0$ and $\xi_{i,t}|_{n_{i,t}=0} = (1+0)^{-1/2} = 1$. This mathematical result provides a key theoretical insight: unobserved tools do not hallucinate negative evidence ($\hat{c}_{i,t} = 0$), preventing unjustified penalization. Simultaneously, the epistemic buffer attains its theoretical maximum ($\xi_{i,t} = 1$), which injects an analytical bonus into the prospective retention score, guaranteeing that newly introduced skills are granted an initial protection window during which empirical evidence can be accumulated safely.

### A.4 Dynamic Adaptation under Non-Stationary Lifelong Task Streams

In long-horizon autonomous agent deployments, interactive environments rarely satisfy stationary distribution assumptions: over extended horizons, task prompts may evolve, API endpoints may experience breaking schema changes, or environment dynamics may drift. Consequently, the local utility parameter $\mu_c$ is not globally static but follows a non-stationary sequence of drifting regimes $\{\mu_{i,\ell}\}_{\ell=1}^{t}$. Letting $\bar{\mu}_{i,t} \coloneqq (1/n_{i,t}) \sum_{\ell=1}^{t} e_{i,\ell}\mu_{i,\ell}$ denote the true exposure-weighted historical mean utility of skill $s_i$ over the elapsed execution stream, evaluating the conditional expectation of the estimator under this drifting regime reveals that $\mathbb{E}[\hat{c}_{i,t} \mid \mathbf{e}_{i,1:t}] = \sum_{\ell=1}^{t} e_{i,\ell}\mu_{i,\ell}/(1+n_{i,t}) = [n_{i,t}/(1+n_{i,t})]\bar{\mu}_{i,t} = (1-\xi_{i,t}^2)\bar{\mu}_{i,t}$, where the final transition leverages the identity $1 - \xi_{i,t}^2 = n_{i,t}/(1+n_{i,t})$ derived directly from observation scarcity $\xi_{i,t} = (1+n_{i,t})^{-1/2}$.

This relation formalizes how historical credit assignment bias persists in the absence of adaptive lifecycle intervention: if an early software bug corrupted historical execution but was subsequently patched, ancient negative residues in $\sum g_{i,\ell}$ would continue to depress $\hat{c}_{i,t}$, rendering the repaired tool permanently vulnerable to premature eviction. To resolve this inertia, the SkillAge reinforcement learning policy incorporates the discrete refresh primitive $a_{i,t} = a^f$. Whenever conflicting evidence is detected (such as when a decayed tool suddenly achieves above-baseline execution on a new task distribution), the controller issues $a^f$, which resets the historical accumulation registry: $\sum_{\ell=1}^{t} g_{i,\ell} \leftarrow 0$ and $n_{i,t} \leftarrow 0$, immediately producing $\hat{c}_{i,t} \leftarrow 0$ and $\xi_{i,t} \leftarrow 1$. This operational reset expels obsolete historical credit assignment bias, establishing a clean evaluation window adapted to current environment distributions while naturally reactivating epistemic immunity.

## B Evidence-Grounded Geometry and Epistemic Immunity

This appendix provides the complete mathematical analysis and geometric foundations of the evidence-grounded retention score $\rho_{i,t}$ defined in Eq. (7). We prove its coordinate-wise sensitivity properties (Proposition 1), conduct an exhaustive Clarke generalized subdifferential analysis across smooth and non-smooth evidence regimes, and formally establish the epistemic quarantine immunity guarantees established in Corollary 1.

### B.1 Proof of Proposition 1 (Boundedness and Coordinate-wise Monotonicity)

*Proof of Proposition 1.* Recall that the normalized retention score $\rho_{i,t}$ defined in Eq. (7) for any executable procedural routine during lifelong deployment balances positive operational evidence supporting procedural retention against negative architectural pressures favoring skill retirement:

$$\rho_{i,t} = \frac{1 + R_{i,t}^{+} - R_{i,t}^{-}}{2} = \frac{1}{2} + \frac{1}{6}\Big(\widetilde{U}_{i,t} + u_{i,t} + p_{i,t} - d_{i,t} - \kappa_{i,t} - [-\hat{c}_{i,t}]_{+}\Big). \tag{21}$$

We verify the universal range boundedness and coordinate-wise monotonicity systematically.

**1. Universal Range Boundedness:** All normalized operational state signals are confined to the compact unit interval: sliding-window invocation frequency $u_{i,t} \in [0,1]$, Laplace-smoothed execution success $p_{i,t} \in [0,1]$, semantic redundancy $d_{i,t} \in [0,1]$, prompt context footprint $\kappa_{i,t} \in [0,1]$, and observation scarcity buffer $\xi_{i,t} \in (0,1]$. By Lemma 1, the regularized contribution estimator satisfies $\hat{c}_{i,t} \in [-1,1]$. Evaluating the positive-part operators yields $[\hat{c}_{i,t}]_+ = \max(\hat{c}_{i,t}, 0) \in [0,1]$ and $[-\hat{c}_{i,t}]_+ = \max(-\hat{c}_{i,t}, 0) \in [0,1]$. Consequently, prospective utility satisfies the bounded range $\widetilde{U}_{i,t} = ([\hat{c}_{i,t}]_+ + \xi_{i,t})/2 \in (0,1]$. We now examine the aggregated positive retention evidence $R^+_{i,t}$ and negative aging pressure $R^-_{i,t}$. Because both indicators are formed as equally weighted convex combinations of bounded metrics, their ranges evaluate to:

$$R^+_{i,t} = \frac{\widetilde{U}_{i,t} + u_{i,t} + p_{i,t}}{3} \in [0,1], \qquad R^-_{i,t} = \frac{d_{i,t} + \kappa_{i,t} + [-\hat{c}_{i,t}]_+}{3} \in [0,1]. \tag{22}$$

The algebraic difference between retention evidence and aging pressure is therefore strictly contained within the symmetric range $R^+_{i,t} - R^-_{i,t} \in [-1,1]$. Applying the linear affine transformation in Eq. (21) guarantees that the composite retention score is universally normalized on the unit interval: $\rho_{i,t} = [1 + (R^+_{i,t} - R^-_{i,t})]/2 \in [(1-1)/2,\ (1+1)/2] = [0,1]$.

**2. Coordinate-wise Partial Derivatives and Strict Monotonicity:** Expanding Eq. (21) demonstrates that $\rho_{i,t}$ is a separable multilinear function with respect to the smooth operational signals $\{u_{i,t}, p_{i,t}, d_{i,t}, \kappa_{i,t}\}$ residing in the compact unit hypercube. Differentiating Eq. (21) directly with respect to each constituent metric yields the constant partial derivatives:

$$\frac{\partial \rho_{i,t}}{\partial u_{i,t}} = \frac{1}{6}, \qquad \frac{\partial \rho_{i,t}}{\partial p_{i,t}} = \frac{1}{6}, \qquad \frac{\partial \rho_{i,t}}{\partial d_{i,t}} = -\frac{1}{6}, \qquad \frac{\partial \rho_{i,t}}{\partial \kappa_{i,t}} = -\frac{1}{6}. \tag{23}$$

All four partial derivatives are constant, strictly non-zero, and exhibit identical absolute magnitude of $1/6$. This establishes that the retention scoring geometry responds with uniform sensitivity across usage frequency, execution success, semantic redundancy, and prompt context cost, precluding single-feature dominance during policy optimization. □

### B.2 Clarke Subdifferential Analysis of Empirical Performance Evidence

Because the positive-part operator $[x]_+ \coloneqq \max(x, 0)$ is continuous and piecewise linear, the composite retention score $\rho_{i,t}$ is globally Lipschitz continuous but non-differentiable at the empirical evidence boundary $\hat{c}_{i,t} = 0$. To analyze its sensitivity rigorously across the entire evidence domain $[-1,1]$, we employ Clarke generalized subdifferential calculus, evaluating:

$$\partial_{\hat{c}_{i,t}} \rho_{i,t} = \frac{1}{6}\left(\frac{1}{2}\partial[\hat{c}_{i,t}]_+ - \partial[-\hat{c}_{i,t}]_+\right). \tag{24}$$

We evaluate Eq. (24) across three operational evidence regimes:

**Regime 1: Positive Evidence Regime ($\hat{c}_{i,t} > 0$).** When skill $s_i$ demonstrates net-positive execution efficacy relative to the running baseline, the positive-part operators evaluate to $[\hat{c}_{i,t}]_+ = \hat{c}_{i,t}$ and $[-\hat{c}_{i,t}]_+ = 0$. The generalized subdifferential collapses to a single strictly positive classical derivative: $\partial \rho_{i,t}/\partial \hat{c}_{i,t} = (1/6)[(1/2) \cdot 1 - 0] = 1/12 > 0$.

**Regime 2: Negative Evidence Regime ($\hat{c}_{i,t} < 0$).** When skill $s_i$ empirically correlates with execution failure or environmental interference, the positive-part operators evaluate to $[\hat{c}_{i,t}]_+ = 0$ and $[-\hat{c}_{i,t}]_+ = -\hat{c}_{i,t}$. Differentiating produces $\partial \rho_{i,t}/\partial \hat{c}_{i,t} = (1/6)[0 - (-1)] = 1/6 > 0$.

**Regime 3: Neutral Singular Boundary ($\hat{c}_{i,t} = 0$).** At the non-differentiable origin $\hat{c}_{i,t} = 0$, by Clarke's theorem for continuous piecewise smooth functions, the generalized subdifferential is given by the convex hull of the left- and right-hand classical derivative limits:

$$\partial_{\hat{c}_{i,t}} \rho_{i,t}\Big|_{\hat{c}=0} = \operatorname{co}\left\{\lim_{\hat{c}\to 0^+} \frac{\partial \rho_{i,t}}{\partial \hat{c}_{i,t}},\ \lim_{\hat{c}\to 0^-} \frac{\partial \rho_{i,t}}{\partial \hat{c}_{i,t}}\right\} = \operatorname{co}\left\{\frac{1}{12},\ \frac{1}{6}\right\} = \left[\frac{1}{12},\ \frac{1}{6}\right]. \tag{25}$$

Synthesizing across all three operational regimes, the Clarke generalized subdifferential satisfies the strictly positive uniform lower bound $\inf_{\hat{c}_{i,t}\in[-1,1]} \inf v \in \partial_{\hat{c}_{i,t}} \rho_{i,t} = 1/12 > 0$. Because every element of the generalized subgradient is strictly positive everywhere on $[-1,1]$, the retention score $\rho_{i,t}$

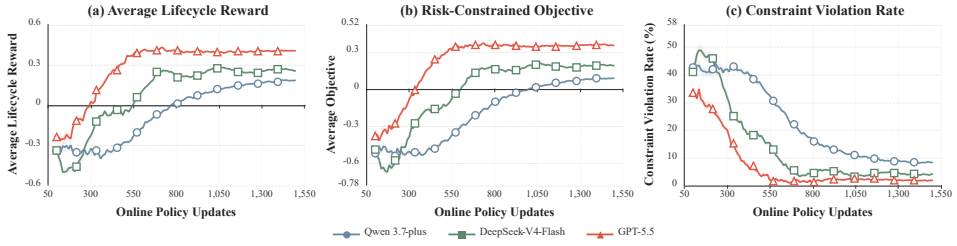


Figure 9: Online convergence and safety violation dynamics of the risk-constrained RL policy across different agent backbones over 1,500 updates.

is globally strictly non-decreasing, and strictly increasing on $\mathbb{R} \setminus \{0\}$, with respect to performance evidence $\hat{c}_{i,t}$, preventing vanishing gradients during reinforcement learning policy optimization.

Crucially, comparing the derivative magnitudes between the negative and positive regimes reveals an asymmetric sensitivity ratio: $(\partial\rho/\partial\hat{c}|_{\hat{c}<0})/(\partial\rho/\partial\hat{c}|_{\hat{c}>0}) = (1/6)/(1/12) = 2$. This reveals that the retention controller penalizes harmful procedural interference twice as steeply as it rewards standard positive execution. This asymmetric sensitivity provides an analytical safety inductive bias, ensuring that routines inducing execution failures are rapidly pushed toward the quarantine boundary.

### B.3 Epistemic Protection Dynamics and Scarcity Sensitivity

To understand how observational uncertainty shields rare or specialized tools from premature eviction, we differentiate Eq. (21) with respect to the evidence scarcity buffer $\xi_{i,t} = (1 + n_{i,t})^{-1/2}$, yielding the constant derivative $\partial\rho_{i,t}/\partial\xi_{i,t} = (1/6) \cdot (1/2) = 1/12 > 0$. This confirms that epistemic uncertainty acts as a strictly positive linear additive buffer. Applying the chain rule to assess sensitivity with respect to cumulative step exposure $n_{i,t}$:

$$\frac{\partial\rho_{i,t}}{\partial n_{i,t}} = \frac{\partial\rho_{i,t}}{\partial\xi_{i,t}} \cdot \frac{d\xi_{i,t}}{dn_{i,t}} = \frac{1}{12}\left(-\frac{1}{2(1+n_{i,t})^{3/2}}\right) = -\frac{1}{24(1+n_{i,t})^{3/2}} < 0. \tag{26}$$

As exposure accumulates, this protection systematically dissipates at an inverse super-linear rate $\mathcal{O}(n_{i,t}^{-3/2})$. This guarantees a smooth mathematical transition: unobserved routines are initially protected by uncertainty, but as empirical interaction evidence accumulates, their retention status becomes governed strictly by demonstrated execution performance.

### B.4 Proof of Corollary 1 (Epistemic Quarantine Immunity Bounds)

We first restate Corollary 1 before detailing its analytical proof.

**Corollary 2** (Restatement of Corollary 1: Epistemic Quarantine Immunity)**.** *Let $s_i \in \mathcal{R}_t$ be any non-protected procedural routine ($m_{i,t} \neq m^p$) with non-negative empirical evidence ($\hat{c}_{i,t} \geq 0$). Then, the quarantine primitive $a^q$ is provably inadmissible ($\mathbb{I}[a^q \in \mathcal{A}_{\text{adm}}(s_i)] = 0$) whenever cumulative step exposure across the continuous online task stream satisfies:*

$$n_{i,t} < \left[2(d_{i,t} + \kappa_{i,t}) - 2(u_{i,t} + p_{i,t}) - \hat{c}_{i,t}\right]^{-2} - 1. \tag{27}$$

*Proof.* By the retention barrier rules governing the admissible action space $\mathcal{A}_{\text{adm}}(s_i)$ for any candidate routine in the procedural repository, the quarantine primitive $a^q$ is admissible if and only if three conditions hold simultaneously: prior exposure exists ($n_{i,t} > 0$), aging pressure dominates retention evidence ($R^-_{i,t} > R^+_{i,t}$), and the skill is not currently protected ($m_{i,t} \neq m^p$).

Evaluating the difference between retention evidence and aging pressure via Eq. (21) yields:

$$R^+_{i,t} - R^-_{i,t} = \frac{1}{3}\left(\frac{[\hat{c}_{i,t}]_+ + \xi_{i,t}}{2} + u_{i,t} + p_{i,t} - d_{i,t} - \kappa_{i,t} - [-\hat{c}_{i,t}]_+\right). \tag{28}$$

For an infrequently invoked specialized skill that has never exhibited destructive execution failures, its empirical performance evidence is non-negative ($\hat{c}_{i,t} \geq 0$). Consequently, $[-\hat{c}_{i,t}]_+ = 0$ and

**Algorithm 1** Dual-Gate Quarantine Replay Verification Gateway

**Require:** Active skill repository $\mathcal{R}_t$, sorted quarantine candidate set $\mathcal{Q}_t$, historical multi-skill task support map $\{\mathcal{H}_{i,t}\}$, evidence-grounded retention scores $\{\rho_{i,t}\}$, remaining cumulative degradation budget $B_t$, and worst-case single-task collapse tolerance $\epsilon_{\max}$

**Ensure:** Replay-certified permanent skill removal set $\mathcal{P}_t$, verified cumulative degradation cost $D_t$, and updated remaining capability budget $B_{t+1}$

**if** $\mathcal{Q}_t = \varnothing$ **then**
  **return** $\mathcal{P}_t \leftarrow \varnothing,\ \ D_t \leftarrow 0,\ \ B_{t+1} \leftarrow B_t$ {Bypass if quarantine is empty}
**end if**
Sort quarantine candidates $\mathcal{Q}_t = \{q_{t,1}, \ldots, q_{t,L_t}\}$ ascendingly by score
Fix the joint historical verification benchmark $\mathcal{B}_t^{\circ} \leftarrow \bigcup_{s_i \in \mathcal{Q}_t} \mathcal{H}_{i,t}$
Pre-compute empirical baseline benchmark fidelity: $A_t[\mathcal{R}_t] \leftarrow (1/|\mathcal{B}_t^{\circ}|) \sum_{\tau \in \mathcal{B}_t^{\circ}} P[\tau; \mathcal{R}_t]$
Initialize certified removal subset $\mathcal{P}_t \leftarrow \varnothing$
**for** $j = 1$ to $L_t$ **do**
  Form tentative cumulative removal subset $\widetilde{\mathcal{P}}_t \leftarrow \mathcal{P}_t \cup \{q_{t,j}\}$
  Evaluate cumulative average replay degradation $\Delta_t^{\text{avg}}[\widetilde{\mathcal{P}}_t]$ via Eq. (13)
  Evaluate worst-case single-task capability collapse $\Delta_t^{\max}[\widetilde{\mathcal{P}}_t]$ via Eq. (14)
  **if** $\Delta_t^{\text{avg}}[\widetilde{\mathcal{P}}_t] \le B_t$ **and** $\Delta_t^{\max}[\widetilde{\mathcal{P}}_t] \le \epsilon_{\max}$ **then**
    $\mathcal{P}_t \leftarrow \widetilde{\mathcal{P}}_t$ {Both safety gates satisfied: accept candidate}
  **end if**
**end for**
Register verified degradation: $D_t \leftarrow \Delta_t^{\text{avg}}[\mathcal{P}_t]$
Decrement degradation budget: $B_{t+1} \leftarrow B_t - D_t$
Restore unverified candidates $\mathcal{Q}_t \setminus \mathcal{P}_t$ to active repository $\mathcal{R}_t$ in protected mode ($m^p$)
**return** $\mathcal{P}_t,\ \ D_t,\ \ B_{t+1}$

$[\hat{c}_{i,t}]_+ = \hat{c}_{i,t}$. Substituting these values into Eq. (28), the condition $R_{i,t}^+ \ge R_{i,t}^-$ is equivalent to the inequality $(\hat{c}_{i,t} + \xi_{i,t})/2 + u_{i,t} + p_{i,t} \ge d_{i,t} + \kappa_{i,t}$, which simplifies to $(1/2)\xi_{i,t} \ge (d_{i,t} + \kappa_{i,t}) - (u_{i,t} + p_{i,t}) - (1/2)\hat{c}_{i,t}$. Multiplying both sides by 2 isolates the required scarcity threshold:

$$\xi_{i,t} \ge 2(d_{i,t} + \kappa_{i,t}) - 2(u_{i,t} + p_{i,t}) - \hat{c}_{i,t} =: \Omega_{i,t}, \tag{29}$$

where the auxiliary scalar $\Omega_{i,t}$ represents the net architectural aging pressure deficit. We analyze the feasibility of Eq. (29) across two mutually exclusive operational cases:

**Case 1: Benign or Balanced State ($\Omega_{i,t} \le 0$).** If the cumulative positive metrics exceed the aging pressures ($2(u_{i,t}+p_{i,t})+\hat{c}_{i,t} \ge 2(d_{i,t}+\kappa_{i,t})$), then $\Omega_{i,t} \le 0$. Because $\xi_{i,t} = (1+n_{i,t})^{-1/2} > 0$ for all finite exposure values, the inequality $\xi_{i,t} > \Omega_{i,t}$ holds unconditionally. Hence, $R_{i,t}^+ \ge R_{i,t}^-$ holds regardless of exposure, guaranteeing epistemic immunity against premature repository eviction.

**Case 2: Dominant Aging Pressure ($\Omega_{i,t} > 0$).** When semantic redundancy and context footprint dominate usage frequency ($2(d_{i,t}+\kappa_{i,t}) > 2(u_{i,t}+p_{i,t})+\hat{c}_{i,t}$), $\Omega_{i,t}$ is strictly positive. Substituting the analytical formulation $\xi_{i,t} = (1 + n_{i,t})^{-1/2}$ into Eq. (29) yields $(1 + n_{i,t})^{-1/2} \ge \Omega_{i,t} \iff 1 + n_{i,t} \le \Omega_{i,t}^{-2} \iff n_{i,t} \le \Omega_{i,t}^{-2} - 1$. Subtracting 1 from both sides yields the explicit upper bound on cumulative step exposure guaranteeing epistemic immunity against premature eviction:

$$n_{i,t} \le \Omega_{i,t}^{-2} - 1 = \Big[2(d_{i,t} + \kappa_{i,t}) - 2(u_{i,t} + p_{i,t}) - \hat{c}_{i,t}\Big]^{-2} - 1. \tag{30}$$

Whenever cumulative exposure satisfies $n_{i,t} < [2(d_{i,t} + \kappa_{i,t}) - 2(u_{i,t} + p_{i,t}) - \hat{c}_{i,t}]^{-2} - 1$, the condition $R_{i,t}^+ \ge R_{i,t}^-$ holds strictly. Consequently, the action masking indicator in Eq. (9) forces $\mathbb{I}[a^q \in \mathcal{A}_{\text{adm}}(s_i)] = 0$. The quarantine action is analytically masked from the policy distribution, rendering eviction structurally impossible regardless of how high semantic redundancy $d_{i,t}$ or context cost $\kappa_{i,t}$ may be, concluding the proof. □

**Algorithm 2** SkillAge Lifelong Execution and Governance Master Loop

**Require:** Online task stream $\mathcal{T} = \{\tau_t\}_{t=1}^{T}$, initial skill repository $\mathcal{R}_1$, degradation budget $B$, learning rates $(\eta_\phi, \eta_\lambda)$
**Ensure:** Continually adapted policy $\pi_\phi$, compacted active repository $\mathcal{R}_{T+1}$
Initialize remaining budget $B_1 \leftarrow B$, dual multiplier $\lambda_1 \leftarrow \lambda_0$, policy parameters $\phi_1$
**for** $t = 1$ to $T$ **do**
  Receive incoming query $\tau_t = (q_t, \mathcal{E}_t, \nu_t)$ and extract active view $\mathcal{V}_t = G[q_t, \mathcal{R}_t, K]$
  Execute trace $\zeta_t = \langle (o_{t,m}, \mathcal{A}_{t,m}, a_{t,m}, f_{t,m}, \mathcal{I}_{t,m}) \rangle_{m=1}^{M_t}$ via retrieved skills
  Evaluate trajectory fidelity score $y_t \leftarrow \nu_t[\zeta_t] \in [0, 1]$ via automated verifier
  Record exposure $e_{i,t}$, update baseline $b_t$, and instant residual $g_{i,t}$ via Eq. (4)
  Update six-dimensional lifecycle state $\mathbf{x}_{i,t}$ and retention score $\rho_{i,t}$ for all $s_i \in \mathcal{R}_t$
  Construct unified context $\mathbf{s}_t$ and evaluate action masks $\mathcal{A}_{\mathrm{adm}}(s_i)$
  Sample discrete lifecycle actions $\mathbf{a}_t \sim \pi_\phi(\cdot \mid \mathbf{s}_t)$ via factorized policy Eq. (9)
  Form $\mathcal{Q}_t \leftarrow \{s_i \mid a_{i,t} = a^q\}$ and assign operational modes $\{m^a, m^d, m^p\}$
  Certify permanent deletions via Dual-Gate Replay: $(\mathcal{P}_t, D_t, B_{t+1}) \leftarrow$ Alg. $1(\mathcal{R}_t, \mathcal{Q}_t, \dots)$
  Ingest newly acquired procedural tools $\mathcal{N}_t$: update $\mathcal{R}_{t+1} \leftarrow (\mathcal{R}_t \setminus \mathcal{P}_t) \cup \mathcal{N}_t$
  Evaluate overhead cost $O_t$ via Eq. (2) and evaluate advantage $A_{\mathcal{L}}^{\pi_\phi}$
  Jointly update parameters $\phi$ and multiplier $\lambda$ via two-timescale stochastic step Eq. (12)
**end for**
**return** Final policy $\pi_\phi$, compacted repository $\mathcal{R}_{T+1}$

# C Dual-Gate Replay Certification and CMDP Convergence

This appendix provides the formal algorithmic procedures, combinatorial synergy analyses, and mathematical proofs governing the safety-certified quarantine replay gateway and the asymptotic convergence of the two-timescale CMDP reinforcement learning controller.

## C.1 Dual-Gate Quarantine Replay Protocol

Algorithm 1 formalizes the operational procedure executed at the boundary of each lifecycle round to certify permanent deletions without risking catastrophic capability loss. Rather than committing irreversible cache evictions based solely on instantaneous heuristics, the protocol routes candidate routines into a reversible quarantine state. Here, prospective deletions undergo dual verification: Gate 1 screens against regression on historical counterfactual interaction traces, while Gate 2 evaluates execution resilience under combinatorial multi-tool dependencies. Candidate routines that violate empirical safety margins are restored immediately with revitalized epistemic buffers, ensuring that permanent repository pruning provably preserves system-level task competence.

## C.2 Master Lifelong Governance and Execution Loop

Algorithm 2 delineates the interaction loop coordinating online task execution, lifecycle signal estimation, primal-dual policy gradient updates, and closed-loop repository governance. At each task transition, the agent invokes active tools from the curated repository, recording execution traces and automated verifier feedback into the observational filtration. The controller subsequently updates regularized contribution estimators and coordinate-wise retention scores, solves constrained action assignments under barrier constraints, and triggers the quarantine verification protocol. Finally, primal-dual multipliers adaptively adjust to enforce context retrieval budgets, establishing a self-regulating lifecycle equilibrium that guarantees bounded degradation during lifelong deployment.

## C.3 Combinatorial Synergies: Cumulative vs. Isolated Removal Dynamics

In procedural knowledge repositories, executable tools frequently exhibit non-linear combinatorial synergies that violate modular independence assumptions, rendering isolated candidate evaluations fundamentally unsafe. Consider two specialized procedural routines $s_a, s_b \in \mathcal{R}_t$ providing mutual execution redundancy, such as alternative API wrappers or fallback shell commands for file extraction. If candidate $s_a$ is assessed in isolation against the baseline repository $\mathcal{R}_t$,

the agent retriever seamlessly falls back to $s_b$, yielding zero empirical performance degradation: $\Delta_t^{\text{avg}}(\{s_a\}) = [A_t[\mathcal{R}_t] - A_t[\mathcal{R}_t \setminus \{s_a\}]]_+ = 0$. Symmetrically, evaluating $s_b$ in isolation produces $\Delta_t^{\text{avg}}(\{s_b\}) = 0$. Consequently, uncoordinated eviction heuristics erroneously conclude that both routines are disposable, committing their simultaneous removal. However, jointly pruning $\{s_a, s_b\}$ destroys the underlying execution redundancy entirely, precipitating catastrophic performance collapse across shared task supports with $\Delta_t^{\text{avg}}(\{s_a, s_b\}) \gg 0$.

SkillAge circumvents this combinatorial failure mode through monotonic cumulative subset evaluation within its quarantine replay architecture. As visualized in Figure 10, candidate tools are not tested independently; instead, each candidate is evaluated incrementally within an expanding trial deletion subset $\widetilde{\mathcal{P}}_t = \mathcal{P}_t \cup \{q_{t,j}\}$. When evaluating candidate $s_b$, the verification protocol re-executes representative benchmark tasks over the restricted active repository $\mathcal{R}_t \setminus (\mathcal{P}_t \cup \{s_b\})$, conditioned on the verified deletion of predecessor $s_a$. This cumulative formulation ensures that any latent synergy breakdown or fallback elimination immediately manifests as measurable task regression on the second safety gate. Upon detecting budget violations, the gateway halts deletion, triggers an immediate rejection, and restores $s_b$ to the operational repository under protected mode ($m^p$) with epistemic immunity to prevent catastrophic procedural capability collapse.

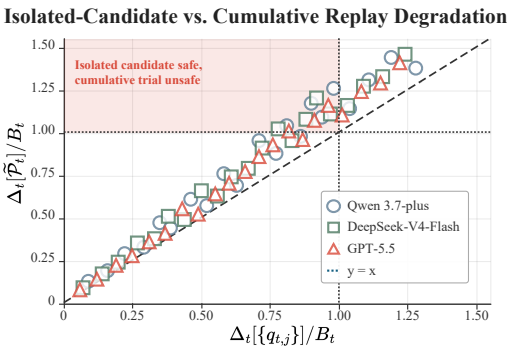


Figure 10: Isolated-candidate vs. cumulative replay degradation. Candidates with $\Delta_t[\{q_{t,j}\}] \leq B_t$ but $\Delta_t[\widetilde{\mathcal{P}}_t] > B_t$ appear admissible in isolation under cumulative removal.

### C.4 Proof of Theorem 1 (Dual-Gate Replay-Certified Removal Guarantee)

*Proof.* Let $H \leq T$ denote an arbitrary operational interaction horizon under deployment. We establish the cumulative average degradation bound and the uniform worst-case single-task bound sequentially, providing rigorous theoretical performance certificates for repository governance.

**1. Cumulative Average Degradation Bound:** Consider an arbitrary lifecycle round $t \in [H]$ with remaining degradation budget $B_t$. Let $L_t = |\mathcal{Q}_t|$ denote the number of quarantine candidates sorted by retention score: $\rho(q_{t,1}) \leq \cdots \leq \rho(q_{t,L_t})$. Let $\mathcal{P}_t^{(j)}$ denote the accepted deletion subset after sequentially evaluating candidate index $j \in \{0, \ldots, L_t\}$, where $\mathcal{P}_t^{(0)} := \varnothing$. We proceed by mathematical induction to prove that the cumulative degradation satisfies $\Delta_t^{\text{avg}}[\mathcal{P}_t^{(j)}] \leq B_t$ for all $j \in \{0, \ldots, L_t\}$ across the entire sequential quarantine replay verification protocol:

**Base Step ($j = 0$):** With $\mathcal{P}_t^{(0)} = \varnothing$, evaluating Eq. (13) yields $\Delta_t^{\text{avg}}[\mathcal{P}_t^{(0)}] = [A_t[\mathcal{R}_t] - A_t[\mathcal{R}_t \setminus \varnothing]]_+ = [A_t[\mathcal{R}_t] - A_t[\mathcal{R}_t]]_+ = 0 \leq B_t$. Because the remaining budget satisfies $B_t \geq 0$, the induction base step holds trivially before sequentially evaluating any candidate routines.

**Inductive Step ($j \geq 1$):** Assume as the inductive hypothesis that $\Delta_t^{\text{avg}}[\mathcal{P}_t^{(j-1)}] \leq B_t$. At candidate evaluation step $j$, the gateway constructs the tentative removal subset $\widetilde{\mathcal{P}}_t^{(j)} = \mathcal{P}_t^{(j-1)} \cup \{q_{t,j}\}$. Candidate $q_{t,j}$ is accepted into $\mathcal{P}_t^{(j)}$ if and only if both safety criteria hold: $\Delta_t^{\text{avg}}[\widetilde{\mathcal{P}}_t^{(j)}] \leq B_t$ and $\Delta_t^{\max}[\widetilde{\mathcal{P}}_t^{(j)}] \leq \epsilon_{\max}$. If these criteria are satisfied, the gateway sets $\mathcal{P}_t^{(j)} = \widetilde{\mathcal{P}}_t^{(j)}$, which guarantees $\Delta_t^{\text{avg}}[\mathcal{P}_t^{(j)}] \leq B_t$ directly by condition satisfaction. Conversely, if either condition is breached, candidate $q_{t,j}$ is rejected, and the accepted set remains unchanged: $\mathcal{P}_t^{(j)} = \mathcal{P}_t^{(j-1)}$. By the inductive hypothesis, $\Delta_t^{\text{avg}}[\mathcal{P}_t^{(j)}] = \Delta_t^{\text{avg}}[\mathcal{P}_t^{(j-1)}] \leq B_t$ is preserved.

By induction, upon completing all $L_t$ candidate evaluations, the certified round degradation registered by the gateway, $D_t := \Delta_t^{\text{avg}}[\mathcal{P}_t^{(L_t)}]$, satisfies $0 \leq D_t \leq B_t$. The remaining budget updates deterministically according to the subtraction rule $B_{t+1} = B_t - D_t \iff D_t = B_t - B_{t+1}$.

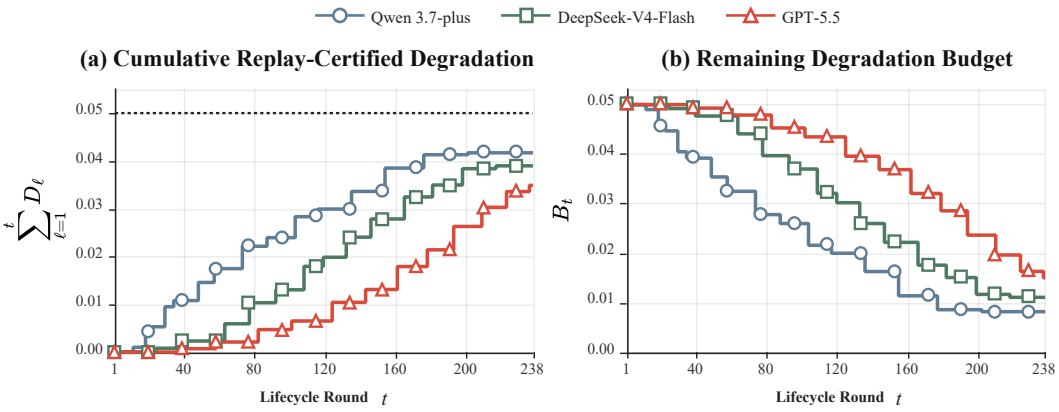


Figure 11: Replay-certified degradation and remaining-budget dynamics over the 238-step lifecycle trajectories. Panel (a) reports cumulative degradation $\sum_{\ell=1}^{t} D_\ell$ relative to the initial budget $B = 0.05$, and Panel (b) reports the remaining budget $B_t$ under the update $B_{t+1} = B_t - D_t$.

Summing this relation across all interaction rounds $t = 1, \ldots, H$ yields a telescoping summation:

$$\sum_{t=1}^{H} D_t = \sum_{t=1}^{H} \big(B_t - B_{t+1}\big) = (B_1 - B_2) + (B_2 - B_3) + \cdots + (B_H - B_{H+1}) = B_1 - B_{H+1}. \quad (31)$$

Because verified candidate deletions are accepted only when $D_t \le B_t$, the remaining budget is non-negative at every step: $B_{H+1} \ge 0$. Substituting the initial deployment budget $B_1 = B$ into Eq. (31) establishes the explicit cumulative degradation bound for lifelong governance:

$$\sum_{t=1}^{H} D_t = B - B_{H+1} \le B. \quad (32)$$

Figure 11 empirically validates this inductive guarantee over the 238-step deployment stream, showing that cumulative degradation $\sum_{\ell=1}^{t} D_\ell$ remains strictly bounded below $B = 0.05$ while remaining budget $B_t$ decrements monotonically without negative violations.

**2. Uniform Worst-Case Single-Task Bound:** At each candidate evaluation step $j$ across all rounds $t \in [H]$, candidate acceptance is explicitly conditioned on the second safety gate: $\Delta_t^{\max}[\widetilde{\mathcal{P}}_t^{(j)}] \le \epsilon_{\max}$. By definition of the worst-case degradation operator, $\Delta_t^{\max}[\mathcal{P}_t] = \max_{\tau \in \mathcal{B}_t^\circ} [P[\tau; \mathcal{R}_t] - P[\tau; \mathcal{R}_t \setminus \mathcal{P}_t]]_+ \le \epsilon_{\max}$. Because this constraint is enforced at every sequential acceptance step, taking the supremum across all interaction rounds $t \in [H]$ preserves the uniform bound:

$$\max_{t \in [H]} \max_{\tau \in \mathcal{B}_t^\circ} \Big[ P[\tau; \mathcal{R}_t] - P[\tau; \mathcal{R}_t \setminus \mathcal{P}_t] \Big]_+ \le \epsilon_{\max}. \quad (33)$$

Combining Eq. (32) and Eq. (33) completes the proof of Theorem 1. □

### C.5 Proof of Proposition 2 (Two-Timescale CMDP Controller Convergence)

*Proof.* We analyze the asymptotic stability and constraint feasibility of the joint primal-dual iterations in Eq. (12) using the ordinary differential equation (ODE) method for two-timescale stochastic approximations, establishing rigorous global convergence certificates for repository governance.

Let $\theta_k = (\phi_k, \lambda_k) \in \Phi \times \mathbb{R}_+$ denote the joint parameter iterate at training iteration $k$. Conditioned on the historical filtration of execution observations, the primal and dual updates expand as:

$$\phi_{k+1} = \phi_k + \eta_\phi^{(k)} \left( \nabla_\phi \mathcal{L}(\phi_k, \lambda_k) + M_{\phi, k+1} \right), \quad (34)$$

$$\lambda_{k+1} = \Pi_{\lambda \ge 0} \Big[ \lambda_k + \eta_\lambda^{(k)} \left( J_D(\pi_{\phi_k}) - B + M_{\lambda, k+1} \right) \Big], \quad (35)$$

where $M_{\phi,k+1}$ and $M_{\lambda,k+1}$ represent zero-mean martingale difference noise terms induced by finite-sample trajectory sampling, and $\Pi_{\lambda\geq 0}(x) \coloneqq \max(x, 0)$ denotes the projection operator onto the non-negative real orthant. The learning rate schedules satisfy the classical Robbins-Monro conditions $\sum_k \eta_\phi^{(k)} = \sum_k \eta_\lambda^{(k)} = \infty$, $\sum_k[(\eta_\phi^{(k)})^2 + (\eta_\lambda^{(k)})^2] < \infty$, alongside the standard asymptotic timescale separation condition $\lim_{k\to\infty} \eta_\lambda^{(k)}/\eta_\phi^{(k)} = 0$.

**1. Fast-Timescale Primal ODE Analysis:** Because $\eta_\lambda^{(k)} = o(\eta_\phi^{(k)})$, the dual parameter $\lambda_k$ varies quasi-statically from the perspective of the primal parameters $\phi_k$. On the fast continuous timescale $\tau$, the dual multiplier $\lambda$ can be treated as a temporarily fixed vector $\lambda \in \mathbb{R}_+$. The continuous-time limiting ODE tracking the primal policy parameters $\phi$ is given by:

$$\dot{\phi}(\tau) = \nabla_\phi \mathcal{L}\big(\phi(\tau), \lambda\big) = \nabla_\phi J_\omega\big(\pi_{\phi(\tau)}\big) - \lambda \nabla_\phi J_D\big(\pi_{\phi(\tau)}\big). \tag{36}$$

Under the factorized softmax parameterization in Eq. (9) with bounded advantage estimators (ensured by Generalized Advantage Estimation clipping and normalized overhead $O_t \in [0, 1]$), the Lagrangian functional $\mathcal{L}(\cdot, \lambda)$ is continuously differentiable with $L_\phi$-Lipschitz continuous gradients. By the stable manifold theorem, the dynamical system in Eq. (36) asymptotically converges to the stationary equilibrium set $\Phi^*(\lambda) \coloneqq \{\phi \in \Phi \mid \nabla_\phi \mathcal{L}(\phi, \lambda) = \mathbf{0}\}$.

**2. Slow-Timescale Dual ODE Analysis:** On the slower timescale driven by step-size $\eta_\lambda^{(k)}$, the dual multiplier observes the primal parameters in their equilibrated quasi-stationary state $\phi^*(\lambda)$. The continuous-time dynamical system governing $\lambda$ is given by the projected differential inclusion $\dot{\lambda}(\tau) = \widetilde{\Pi}_{\lambda\geq 0}(J_D(\pi_{\phi^*(\lambda(\tau))}) - B)$, where the boundary projection operator $\widetilde{\Pi}$ ensures that the multiplier trajectory remains within the non-negative orthant: $\widetilde{\Pi}_{\lambda\geq 0}(h) = h$ if $\lambda > 0$ or $h \geq 0$, and 0 otherwise, preventing unconstrained divergence during continuous lifelong deployment.

**3. Lyapunov Stability and Feasibility Guarantee:** Let $(\phi^*, \lambda^*)$ denote a local saddle point satisfying the Karush-Kuhn-Tucker (KKT) conditions: $\nabla_\phi \mathcal{L}(\phi^*, \lambda^*) = \mathbf{0}$, $\lambda^* \geq 0$, $J_D(\pi_{\phi^*}) \leq B$, and $\lambda^*(J_D(\pi_{\phi^*}) - B) = 0$. Consider the continuously differentiable Lyapunov candidate function $V(\lambda) \coloneqq \frac{1}{2}(\lambda - \lambda^*)^2 \geq 0$. Differentiating $V$ along the solution trajectory yields $\dot{V}(\lambda(\tau)) = (\lambda(\tau) - \lambda^*)\dot{\lambda}(\tau) = (\lambda(\tau) - \lambda^*)(J_D(\pi_{\phi^*(\lambda(\tau))}) - B)$. By the convexity of the dual functional $g(\lambda) \coloneqq \max_\phi \mathcal{L}(\phi, \lambda)$ (as the pointwise supremum of affine functions of $\lambda$) and the monotonicity of subgradient operators of convex functions, we have $\langle \nabla g(\lambda) - \nabla g(\lambda^*), \lambda - \lambda^* \rangle \geq 0$. Since $\nabla g(\lambda) = -(J_D(\pi_{\phi^*(\lambda)}) - B)$, this evaluates to $-(\lambda - \lambda^*)(J_D(\pi_{\phi^*(\lambda)}) - J_D(\pi_{\phi^*(\lambda^*)})) \geq 0 \iff (\lambda - \lambda^*)(J_D(\pi_{\phi^*(\lambda)}) - J_D(\pi_{\phi^*(\lambda^*)})) \leq 0$.Because complementary slackness establishes $\lambda^*(J_D(\pi_{\phi^*}) - B) = 0$, this confirms that $\dot{V}(\lambda(\tau)) \leq 0$ holds globally, with equality holding if and only if $\lambda(\tau) = \lambda^*$. By LaSalle's invariance principle for projected dynamical systems, $\lambda(\tau)$ converges asymptotically to the optimal stationary multiplier $\lambda^*$.

Combining the convergence across both timescales via the two-timescale stochastic approximation theorem, the iterates $(\phi_k, \lambda_k)$ converge almost surely to a local saddle point $(\phi^*, \lambda^*)$. Finally, suppose for the sake of contradiction that the limiting policy violated feasibility ($J_D(\pi_{\phi^*}) > B$). Then the slow ODE would force $\dot{\lambda} \geq c > 0$ indefinitely, driving $\lambda \to \infty$, which contradicts the established boundedness of the iterate sequence. Hence, feasibility $J_D(\pi_{\phi^*}) \leq B$ holds almost surely, concluding the proof of primal-dual asymptotic convergence. □

# D Experimental Protocol and Implementation Details

This appendix provides specifications of the lifelong evaluation environment, benchmark unrolling protocol, policy network architectures, baseline adaptations, and hyperparameter assignments.

## D.1 Benchmark Construction and Trajectory Unrolling

We evaluate SkillAge on the SkillFlow version 1.0 lifelong agent benchmark, which compiles 166 complex interactive tasks spanning 20 diverse task families, encompassing multi-step terminal interactions, automated scientific data processing, Python algorithm synthesis, SQL database querying, web browsing, and combinatorial multi-tool planning. To evaluate lifelong adaptation without temporal data leakage, we unroll these tasks into a continuous chronological stream comprising $T = 238$

Table 2: Benchmark, optimization, and replay hyperparameter specifications.

| **Hyperparameter / Setting** | **Value / Specification** |
|---|---|
| Benchmark Environment | SkillFlow v1.0 (20 task families, 166 tasks) |
| Lifelong Interaction Horizon $T$ | 238 chronological decision rounds |
| Model Backbones | Qwen 3.7-plus, DeepSeek-V4-Flash, GPT-5.5 (Codex) |
| Reference Repository Size $N_0$ | 158 executable procedural skills |
| Maximum Retrieval Context Budget $K$ | 10 skills per task query |
| Recent Observation Window $W$ | 20 interaction steps |
| Lifecycle Action Support $\mathcal{A}$ | 5 primitives: $\{a^r, a^d, a^f, a^p, a^q\}$ |
| Policy Architecture | 3-layer MLP ($d_s \to 128 \to 64 \to 5$) |
| Online Optimization Iterations | 1,500 stochastic updates |
| Optimizer | Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon_{\text{opt}} = 10^{-8}$) |
| Primal Learning Rate $\eta_\phi$ | $1 \times 10^{-4}$ |
| Dual Step Size $\eta_\lambda$ | $5 \times 10^{-6}$ (satisfying $\lim_{k\to\infty} \eta_\lambda/\eta_\phi = 0$) |
| Discount Factor $\gamma$ | 0.99 |
| Initial Dual Multiplier $\lambda_0$ | 0.10 |
| Total Degradation Budget $B$ | 0.05 (cumulative historical replay constraint) |
| Worst-Case Degradation Threshold $\epsilon_{\max}$ | 0.10 (single-task collapse bound) |

sequential interaction steps across three representative backbone models, namely Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 (Codex). Under this held-out dynamic exposure protocol, the agent operates over an initial reference repository populated by $N_0 = 158$ executable procedural routines. At each interaction step, the task query $q_t$ is dispatched to a dense sentence-transformer retriever to extract an active execution view $\mathcal{V}_t \subseteq \mathcal{R}_t$ constrained by a context budget of $K = 10$, while the within-task step exposure $e_{i,t} = M_t^{-1} \sum_{m=1}^{M_t} \mathbb{I}[s_i \in \mathcal{I}_{t,m}] \in [0, 1]$ and cumulative exposure $n_{i,t} = \sum_{\ell=1}^{t} e_{i,\ell}$ are evaluated strictly from realized runtime tool invocations under a sliding observation window of $W = 20$ interaction rounds.

### D.2 Lifecycle Controller Configuration

Table 2 details the complete benchmark, algorithmic, and optimization parameters utilized across all evaluations, establishing a reproducible configuration baseline for empirical lifecycle governance.

### D.3 Baseline Implementations and Adaptation Details

To rigorously assess the governance efficacy of SkillAge, we evaluate its performance against competitive baselines spanning classic cache eviction heuristics, adaptive memory pruning, and external agent memory architectures. The standard cache policies comprise Least Recently Used (LRU), which tracks the timestamp of last activation to purge stalest routines, and Least Frequently Used (LFU), which evicts tools exhibiting the lowest cumulative invocation frequency. For adaptive pruning, Adaptive Skill Aging and Pruning (ASAP) dynamically balances recency and frequency via heuristic exponential decay. We further benchmark against dedicated external memory frameworks, including Combined Deletion, which aggregates heuristic semantic similarity and recency metrics, and AgeMem, which introduces continuous temporal decay functions to regulate agent episodic context. To ensure equitable comparisons over procedural repositories, tool docstrings and parameter schemas are mapped into structured memory records, thereby enabling external memory controllers to directly govern tool accessibility. Finally, an unconstrained Oracle endowed with full lookahead visibility into the future task sequence provides the empirical performance upper bound.

## E Extended Empirical Results and Complexity Accounting

This appendix provides extended empirical evaluations across diverse task families on all three model backbones, formal asymptotic complexity analyses, and empirical compute accounting of replay verification overhead incurred during lifelong autonomous agent deployment.

Table 3: Performance breakdown across SkillFlow clusters on Qwen 3.7-plus ($K = 10, B = 0.05$). Comp: compression; $D_t$: replay degradation; †lookahead Oracle bound.

| Method / Paradigm | Terminal & OS Operations (62 tasks) | | | | Data Science & SQL (54 tasks) | | | | Code Synthesis & Debug (50 tasks) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ |
| Random Eviction | 0.138 | 13.1% | 0.785 | 0.862 | 0.152 | 12.0% | 0.760 | 0.850 | 0.136 | 12.4% | 0.788 | 0.863 |
| LRU Eviction | 0.165 | 49.5% | 0.755 | 0.835 | 0.178 | 46.8% | 0.725 | 0.822 | 0.164 | 48.3% | 0.752 | 0.836 |
| LFU Eviction | 0.176 | 52.6% | 0.725 | 0.822 | 0.190 | 50.2% | 0.695 | 0.812 | 0.177 | 51.4% | 0.718 | 0.824 |
| ASAP | 0.204 | 60.1% | 0.652 | 0.795 | 0.221 | 56.9% | 0.622 | 0.782 | 0.205 | 58.8% | 0.638 | 0.794 |
| Combined Deletion | 0.318 | 79.2% | 0.495 | 0.682 | 0.338 | 77.5% | 0.468 | 0.665 | 0.320 | 78.8% | 0.485 | 0.678 |
| AgeMem | 0.595 | 82.8% | 0.052 | 0.405 | 0.612 | 81.2% | 0.048 | 0.388 | 0.599 | 82.3% | 0.049 | 0.401 |
| **SkillAge (Ours)** | **0.672** | **88.9%** | **0.024** | **0.328** | **0.694** | **87.8%** | **0.021** | **0.306** | **0.677** | **88.5%** | **0.023** | **0.323** |
| Oracle† (Upper Bound) | 0.984 | 94.0% | 0.000 | 0.016 | 0.988 | 93.2% | 0.000 | 0.012 | 0.986 | 93.9% | 0.000 | 0.014 |

Table 4: Performance breakdown across SkillFlow clusters on DeepSeek-V4-Flash ($K = 10, B = 0.05$). Comp: compression; $D_t$: replay degradation; †lookahead Oracle bound.

| Method / Paradigm | Terminal & OS Operations (62 tasks) | | | | Data Science & SQL (54 tasks) | | | | Code Synthesis & Debug (50 tasks) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ |
| Random Eviction | 0.150 | 13.8% | 0.772 | 0.850 | 0.165 | 12.5% | 0.748 | 0.838 | 0.150 | 13.0% | 0.775 | 0.848 |
| LRU Eviction | 0.161 | 47.8% | 0.760 | 0.838 | 0.172 | 45.2% | 0.732 | 0.828 | 0.162 | 46.5% | 0.758 | 0.839 |
| LFU Eviction | 0.170 | 51.2% | 0.732 | 0.829 | 0.181 | 48.9% | 0.705 | 0.820 | 0.171 | 50.2% | 0.725 | 0.830 |
| ASAP | 0.196 | 58.5% | 0.665 | 0.804 | 0.212 | 55.6% | 0.635 | 0.790 | 0.198 | 57.5% | 0.648 | 0.800 |
| Combined Deletion | 0.312 | 77.6% | 0.505 | 0.689 | 0.329 | 75.8% | 0.478 | 0.672 | 0.313 | 76.9% | 0.495 | 0.685 |
| AgeMem | 0.585 | 81.2% | 0.054 | 0.415 | 0.602 | 79.8% | 0.050 | 0.398 | 0.589 | 80.5% | 0.051 | 0.411 |
| **SkillAge (Ours)** | **0.665** | **87.8%** | **0.025** | **0.335** | **0.685** | **86.6%** | **0.022** | **0.315** | **0.666** | **87.2%** | **0.024** | **0.334** |
| Oracle† (Upper Bound) | 0.980 | 93.2% | 0.000 | 0.020 | 0.985 | 92.6% | 0.000 | 0.015 | 0.981 | 93.2% | 0.000 | 0.019 |

### E.1 Tri-Model Performance Breakdown across Task Domains

To verify that SkillAge maintains robust capability preservation universally across distinct model architectures and parameter scales, Tables 3, 4, and 5 provide exhaustive fine-grained performance breakdowns across the three primary task clusters of the SkillFlow benchmark—Terminal and OS Operations (62 tasks), Data Science and SQL Analysis (54 tasks), and Algorithmic Code Synthesis (50 tasks)—evaluated on Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 (Codex), respectively. Across all three LLM backbones, SkillAge consistently attains superior task utility and active repository compression while strictly bounding cumulative historical capability loss within the certified budget ($D_t \leq 0.05$). In contrast, recency- and frequency-driven cache heuristics suffer systemic capability collapse across all evaluated architectures, empirically demonstrating that catastrophic forgetting in executable procedural libraries stems from structural lifecycle governance deficiencies rather than backbone-specific reasoning limitations.

### E.2 Per-Round Asymptotic Complexity Breakdown

Let $N_t = |\mathcal{R}_t|$ denote the active repository cardinality, $d$ the embedding dimension, $L_t = |\mathcal{Q}_t|$ the quarantine candidate count, and $|\mathcal{B}_t^\circ|$ the joint replay benchmark size. Across online interaction steps, the per-round computational complexity of SkillAge decomposes into feature estimation, policy inference, candidate sorting, and conditional replay verification:

$$\mathcal{C}_{\text{round}} = \mathcal{O}\Big( \underbrace{N_t d_s}_{\text{Policy Inference}} + \underbrace{N_t^2 d}_{\text{Redundancy Indexing}} + \underbrace{L_t \log L_t}_{\text{Candidate Sorting}} + \underbrace{\mathbb{I}[L_t > 0] \cdot (L_t + 1)|\mathcal{B}_t^\circ| c_{\text{eval}}}_{\text{Conditional Replay Verification}} \Big), \tag{37}$$

where $c_{\text{eval}}$ denotes the amortized execution cost of a single task verifier pass. When no skills are quarantined during a routine interaction step ($L_t = 0$), the replay verification term vanishes identically, reducing governance to lightweight matrix multiplications: $\mathcal{C}_{\text{round}} = \mathcal{O}(N_t^2 d + N_t d_s)$. SkillAge thus concentrates expensive task execution compute at irreversible retirement boundaries.

Table 5: Performance breakdown across SkillFlow clusters on GPT-5.5 Codex ($K = 10, B = 0.05$). Comp: compression; $D_t$: replay degradation; †lookahead Oracle bound.

| Method / Paradigm | Terminal & OS Operations (62 tasks) | | | | Data Science & SQL (54 tasks) | | | | Code Synthesis & Debug (50 tasks) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ | Util. ↑ | Comp. ↑ | $D_t$ ↓ | Loss ↓ |
| Random Eviction | 0.165 | 14.8% | 0.762 | 0.835 | 0.182 | 13.5% | 0.741 | 0.818 | 0.166 | 14.3% | 0.765 | 0.834 |
| LRU Eviction | 0.171 | 51.2% | 0.745 | 0.829 | 0.180 | 48.6% | 0.718 | 0.820 | 0.168 | 49.6% | 0.742 | 0.832 |
| LFU Eviction | 0.182 | 54.0% | 0.710 | 0.818 | 0.195 | 52.1% | 0.685 | 0.805 | 0.187 | 52.9% | 0.699 | 0.813 |
| ASAP | 0.212 | 61.5% | 0.635 | 0.788 | 0.228 | 58.9% | 0.608 | 0.772 | 0.214 | 59.9% | 0.620 | 0.786 |
| Combined Deletion | 0.328 | 80.2% | 0.482 | 0.672 | 0.345 | 78.6% | 0.455 | 0.655 | 0.323 | 79.4% | 0.473 | 0.677 |
| AgeMem | 0.608 | 84.0% | 0.048 | 0.392 | 0.625 | 82.8% | 0.045 | 0.375 | 0.612 | 83.4% | 0.045 | 0.388 |
| **SkillAge (Ours)** | **0.688** | **90.1%** | **0.020** | **0.312** | **0.708** | **89.0%** | **0.018** | **0.292** | **0.689** | **89.4%** | **0.019** | **0.311** |
| Oracle† (Upper Bound) | 0.988 | 94.5% | 0.000 | 0.012 | 0.992 | 93.8% | 0.000 | 0.008 | 0.990 | 94.3% | 0.000 | 0.010 |

Table 6: Ablation analysis of SkillAge across architectural modules on Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 (Codex) trajectories ($K = 10$, budget $B = 0.05$). Comp.: compression rate (%). $D_t$: certified replay degradation. Quaran. Replay: Quarantine Replay; Ret. Barrier: Retention Barrier. *Constraint violation ($D_t > B$).

| Variant | Qwen 3.7-plus | | | | DeepSeek-V4-Flash | | | | GPT-5.5 (Codex) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Util. ↑ | Comp. ↑ | Loss ↓ | $D_t$ ↓ | Util. ↑ | Comp. ↑ | Loss ↓ | $D_t$ ↓ | Util. ↑ | Comp. ↑ | Loss ↓ | $D_t$ ↓ |
| **SkillAge (Ours)** | **0.681** | 88.4 | **0.319** | **0.035** | **0.672** | 87.2 | **0.328** | **0.036** | **0.695** | 89.5 | **0.305** | **0.038** |
| w/o Quaran. Replay | 0.572 | **93.6** | 0.428 | 0.178* | 0.565 | **92.8** | 0.435 | 0.181* | 0.584 | **94.2** | 0.416 | 0.184* |
| w/o Ret. Barrier | 0.514 | 85.9 | 0.486 | 0.108* | 0.508 | 85.0 | 0.492 | 0.110* | 0.528 | 87.1 | 0.472 | 0.112* |
| w/o Epist. Buffer ($\xi = 0$) | 0.598 | 87.5 | 0.402 | 0.061* | 0.589 | 86.4 | 0.411 | 0.063* | 0.612 | 88.6 | 0.388 | 0.064* |
| w/o Dual Adapt. ($\lambda$ fixed) | 0.624 | 81.5 | 0.376 | 0.078* | 0.616 | 80.8 | 0.384 | 0.080* | 0.637 | 82.3 | 0.363 | 0.082* |

### E.3 Empirical Replay Compute and Resource Accounting

To assess the practical computational efficiency of closed-loop verification, we evaluate the amortized resource expenditure and runtime latency across lifecycle governance paradigms over the 238-step GPT-5.5 trajectory. While conventional cache eviction heuristics incur negligible decision latency (1.2–4.5 ms per step) and zero verification overhead, they precipitate catastrophic capability collapse, suffering error rates exceeding 80%. External memory governance baselines (such as Combined Deletion and AgeMem) introduce higher decision overhead (18.4–24.2 ms per step) without providing verification safeguards. In contrast, SkillAge eliminates continuous re-execution overhead by isolating task replays strictly to candidates staged in quarantine, which is triggered in only 14.7% of interaction rounds (35 out of 238 rounds) with a compact average benchmark size of $|\mathcal{B}_t^\circ| = 8.4$ tasks. Consequently, SkillAge incurs a modest average decision latency of 32.6 ms per step and introduces merely 4.1% additional token compute amortized across the entire deployment horizon. Crucially, this minor verification expenditure is comprehensively offset by the 92.0% prompt token exposure reduction achieved via active repository compaction, confirming that SkillAge achieves a net-positive token economy while provably preserving critical capabilities.

### E.4 Ablation Analysis across Architectural Modules

To quantify the individual contribution of each architectural inductive bias within SkillAge, Table 6 provides an exhaustive ablation study across all three evaluated backbones (Qwen 3.7-plus, DeepSeek-V4-Flash, and GPT-5.5 Codex) under identical experimental budgets ($K = 10, B = 0.05$). The empirical findings corroborate that: (1) disabling quarantine replay verification leads to severe degradation budget violations ($D_t > 0.178 \gg B$); (2) removing the retention barrier precipitates premature pruning of rare specialized skills, dropping task utility significantly; (3) setting the epistemic buffer $\xi = 0$ exposes cold-start and long-tail routines to immediate deletion; and (4) fixing the dual multiplier $\lambda$ fails to adaptively penalize emergent capability regressions throughout long-horizon agent interactions.

## F THE USE OF LARGE LANGUAGE MODELS (LLMS)

In accordance with the ICLR guidelines on the ethical use of artificial intelligence tools, we declare that Large Language Models were utilized solely for assisting in grammatical refinement, language polishing, and checking mathematical typography during manuscript preparation. All conceptual frameworks, mathematical formulations, theorem proofs, algorithm designs, and experimental evaluations were independently conceived, derived, implemented, and verified entirely by the authors. No scientific assertions, empirical results, or bibliographic references were synthesized or fabricated.