When to Look Again: Learning from Refresh Feedback For LLM Agents

Anonymous authors
Paper under double-blind review

Abstract

Large language model (LLM) agents reuse information from tools, retrieval, and memory over many steps while the underlying sources keep changing, so an agent can act on an outdated copy of information it once retrieved correctly. With a small refresh budget per step, the agent must decide which copies can wait. We model independent source changes with weighted staleness costs and a shared refresh budget. The model is a restless bandit whose closed-form Whittle index sets the refresh priority and a Lagrangian lower bound evaluates the scheduler. The index reflects both the risk of staleness and how long a refresh keeps the copy fresh, i.e., faster change raises priority at short ages but lowers the priority ceiling. When rates are learned from refreshes, an upper confidence bound (UCB) on a sources rate therefore lowers its priority and can delay the observations that would correct it. We instead maximize the index over plausible rates and rebuild the plausible set at every decision. For a two-source class, we prove that freezing the set can yield linear expected excess cost while rebuilding it keeps the excess bounded, and we give checkable conditions for bounded cost with multiple unknown sources. In numerical experiments, maximizing the index over plausible rates approaches the cost of a scheduler that knows the rates, while a UCB on the rate incurs persistent excess cost. With known rates, the index scheduler stays close to the lower bound across budgets and system sizes.

1 Introduction

An agent helping a customer may reuse an inventory count, a delivery estimate, and a policy document over several steps, each of which can become outdated while it works. Rechecking them consumes calls and time; leaving them unchecked risks decisions based on obsolete information, as time-sensitive question answering and dynamic agent benchmarks show (Vu et al., 2024; Froger et al., 2026). With a limited refresh budget, the agent must decide which sources can wait.

A natural rule is to revisit the copy most likely to be stale. However, a source that changes often also becomes stale again soon after a refresh, while a slower source stays fresh longer after one visit. So even at equal relevance, the value of a refresh depends on both current staleness and how long the new copy stays fresh.

We study refresh allocation over an existing working set whose changes remain hidden between visits. Adaptive retrieval can decide when more information is needed using confidence or learned critics (Asai et al., 2024; Jiang et al., 2023; Jeong et al., 2024; Su et al., 2024). Memory systems manage stored information (Packer et al., 2023; Chhikara et al., 2025; Yu et al., 2026), and stale-memory detection finds entries contradicted by incoming evidence (Chao et al., 2026). We ask how sources should share a limited number of visits over time, including when no new evidence signals a change.

In our model, sources change independently at constant rates, refreshes are exact, and stale copies incur fixed relevance-weighted costs. When rates are known, the time since the last refresh determines each copy's staleness probability. The problem becomes a restless multi-armed bandit (Whittle, 1988), since unvisited sources continue to change. A classical web-crawling index gives the refresh priority (Avrachenkov & Borkar, 2018). This index depends on the age of each copy. At equal relevance, a faster-changing source ranks higher one slot after a visit but has a lower priority

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.

ceiling, g_i/p_i, where g_i is its weight and p_i its change probability. Thus, after a long enough wait, the ranking reverses.

This reversal also matters when the agent must estimate the rates from its own visits. An upper estimate of a source's change rate lowers its priority ceiling and can postpone the next observation, so the estimate stays too high because the policy does not visit the source. To avoid this problem, we use index optimism, which takes the largest index supported by plausible rates, and compare it with rate optimism, which takes the largest plausible rate. Index optimism also requires revisable candidate sets. If an early observation excludes every rate that could win another visit and the set is frozen until that visit, the source is never visited again. Rebuilding the set at each decision brings an excluded rate back after any finite history, and we analyze when the resulting visits identify the rate.

We make four contributions.

- We prove threshold optimality and indexability, recover the closed-form web-crawling index, and obtain a Lagrangian lower bound (Sections 2–3). Whittle-Index Refresh Scheduling (WIRS) ranks sources by this index, which reflects both current staleness and how long a refresh keeps the copy fresh.
- Our learning scheduler UCB-WIRS applies index optimism. It maximizes the index over plausible rates and rebuilds the set of plausible rates at every decision. In experiments, its cost stays close to that of known-rate WIRS, while rate optimism incurs persistent excess cost. An LLM agent that answers from the maintained copies is also correct more often under UCB-WIRS than under rate optimism (Sections 4–5).
- For a two-source class with an exactly solvable hard-budget optimum, frozen candidate sets give linear expected excess while reopening gives a uniform-in-time bound (Section 4.1). Continuous rates admit a grid-spacing error bound, and a capacity margin and a common cycle mean give bounded excess with multiple unknown sources (Appendix B).
- Known-rate WIRS stays close to the lower bound across budgets and population sizes. It keeps the lowest cost under the tested bursty changes, but a rate-free age rule does better when the supplied rates have large errors (Section 5).

2 A MODEL OF BUDGETED REFRESH

The agent keeps copies of its sources while the sources change. It spends a shared refresh budget to reduce the cost of keeping stale copies.

Sources and copies. There are N sources, indexed by $i \in \mathcal{N} \triangleq \{1, \dots, N\}$, and time is slotted, $t \in \{0, 1, 2, \dots\}$. Let $Z_i(t) \in \{0, 1\}$ indicate whether source i has changed since its last refresh. We call the retained copy *stale* when $Z_i(t) = 1$. Staleness evolves as follows.

Assumption 1 (Markov change process)**.** *After the refresh decision and cost in slot t, a fresh copy becomes stale with probability $p_i \in (0, 1)$ before slot $t+1$. A stale copy remains stale until refreshed. Change events are independent across sources and slots.*

Refresh action and observations. In each slot, $a_i(t) = 1$ means that the agent refreshes source i, and $a_i(t) = 0$ means that it waits. An integer budget $1 \le B < N$ limits the number of visits,

$$\sum_{i=1}^{N} a_i(t) \le B. \tag{1}$$

The visit returns the source's content and reveals whether the previous copy was stale. This binary observation is the feedback used to learn change rates in Section 4. Between visits, the source is unobserved.

Assumption 2 (Noiseless refresh)**.** *A refresh returns the source's current content exactly.*

Thus a visit resets Z_i to zero before cost is incurred. This known reset, together with independent changes, lets the scheduler track staleness from elapsed time alone when the rates are known. Appendix F considers the belief state needed if the returned content can itself be stale.

Cost. After the slot's refreshes, the agent incurs

$$c\big(Z(t), a(t)\big) = \sum_{i=1}^{N} g_i \, Z_i(t^+), \tag{2}$$

where $Z_i(t^+)$ denotes post-refresh staleness and $g_i > 0$ is the expected per-slot loss from a stale copy of source i. The weight can combine use frequency with the consequence of using an outdated value. A refreshed source contributes zero in that slot. Calls consume the budget in (1); Remark 2 adds optional per-call penalties.

Assumption 3. *The weights $g_1, \ldots, g_N$ are known and constant over the horizon under consideration.*

Tracking staleness by age. Let $h_i(t) \in \mathbb{N}$ be the *age* of copy i, measured in slots since its last refresh. Immediately after a visit, $h_i = 0$; changes occur before the next decision, when the age is one. For known rates, this counter contains all the information needed to predict staleness.

Proposition 1. *With known change rates, the posterior staleness probability of source i given the agent's observation history is*

$$\omega_i(h) \triangleq \mathbb{P}\big(Z_i(t) = 1 \mid h_i(t) = h\big) = 1 - q_i^{\,h}, \tag{3}$$

$$q_i \triangleq 1 - p_i, \tag{4}$$

and the process $\big(h_1(t), \ldots, h_N(t)\big)$ is a controlled Markov chain whose expected per-slot cost equals $\sum_i g_i \, \omega_i(h_i(t^+))$.

Proposition 1 replaces hidden content histories with an observable vector of ages. The unknown-rate case additionally requires the visit observations used to estimate p_i.

Objective. A policy π maps observation histories to feasible refresh sets. Its objective is the long-run average staleness cost,

$$J(\pi) \triangleq \limsup_{T\to\infty} \frac{1}{T} \, \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \sum_{i=1}^{N} g_i \, \omega_i\big(h_i(t^+)\big) \right], \tag{5}$$

$$\text{s.t.} \quad \sum_i a_i(t) \le B, \forall t. \tag{6}$$

By Proposition 1 and the law of total expectation, the same objective is obtained by averaging the expected realized cost (2). We record realized costs in the experiments. Given the actions, ages evolve deterministically; the difficulty is allocating the shared budget, since a visit to one source can force another to wait. Whittle's relaxation assigns a price to each visit and yields a separate priority for each source (Whittle, 1988).

3 The value of a refresh

A visit removes the current expected loss $g_i\omega_i(h_i)$ and changes the copy's freshness in later slots. To value both effects, we compare complete refresh cycles. Relaxing (1) to a time-average constraint assigns each visit a price $\nu \ge 0$ and separates the sources. Under Assumptions 1–3, with known rates, a source with parameters (g, p) then chooses its refresh times to minimize

$$J_\nu(\sigma) \triangleq \limsup_{T\to\infty} \frac{1}{T} \, \mathbb{E}_\sigma \left[\sum_{t=0}^{T-1} \Big(g \, \omega\big(h(t^+)\big) + \nu \, a(t) \Big) \right]. \tag{7}$$

Refreshing every τ slots incurs expected staleness $S(\tau) \triangleq \sum_{h=0}^{\tau-1} \omega(h) = \tau - \frac{1-q^\tau}{p}$ and one visit price per cycle. Theorem 1 shows that optimizing over these cycles, together with the option never to refresh, solves the single-source problem (7).

Theorem 1 (Threshold optimality). *For every $\nu \geq 0$, the single-source problem* (7) *admits an optimal stationary policy of threshold type, which refreshes if and only if $h \geq \tau^\star(\nu)$, where*

$$A_\nu(\tau) \triangleq \frac{g\,S(\tau) + \nu}{\tau} \tag{8}$$

$$\tau^\star(\nu) \in \arg\min_{\tau \in \mathbb{N}_{\geq 1} \cup \{\infty\}} A_\nu(\tau), \tag{9}$$

where $A_\nu(\infty) \triangleq \lim_{\tau\to\infty} A_\nu(\tau) = g$ and $\tau^\star(\nu) = \infty$ means "never refresh". Let $m(\tau) \triangleq g\big[\tau\,\omega(\tau) - S(\tau)\big]$, strictly increasing with $m(0) = 0$ and $\sup_\tau m(\tau) = g/p$. For $\nu < g/p$ the minimizers in (8) *form the interval $\{\tau \in \mathbb{N}_{\geq 1} : m(\tau{-}1) \leq \nu \leq m(\tau)\}$; for $\nu \geq g/p$ the unique minimizer is $\tau = \infty$; in either case the minimizer set moves monotonically in ν.*

Proof sketch. The cost of waiting increases with age, whereas a refresh always leads to the same next state. A vanishing-discount argument formalizes this comparison and gives an optimal threshold policy. For a finite threshold, the cycle average is $A_\nu(\tau)$. The identity $A_\nu(\tau{+}1) - A_\nu(\tau) = \frac{m(\tau)-\nu}{\tau(\tau+1)}$ then characterizes its minimizers because m is strictly increasing. The full argument, including the never-refresh case, is in Appendix A. □

As visits become more expensive, the optimal policy waits longer. The set of ages at which waiting is optimal therefore expands with the price. We use Whittle indexability to describe it. At each age, the index is the price at which refreshing now and waiting one more slot have equal value.

Theorem 2 (Indexability and the refresh index). *Problem* (5) *is indexable, and the Whittle index of source i at age h is*

$$W_i(h) = g_i \sum_{k=0}^{h-1} \left(q_i^{\,k} - q_i^{\,h}\right) = g_i \left[\frac{1 - q_i^{\,h}}{p_i} - h\,q_i^{\,h}\right], \qquad W_i(0) = 0. \tag{10}$$

W_i is strictly increasing in h, with $W_i(h{+}1) - W_i(h) = g_i\,p_i\,(h{+}1)\,q_i^{\,h} > 0$, and $W_i(h) \uparrow g_i/p_i$ as $h \to \infty$.

Proof sketch. For $h \geq 1$, the index is the refresh price at which thresholds h and $h{+}1$ tie. Solving $A_\nu(h) = A_\nu(h{+}1)$ gives $\nu = m(h) = g\,[h\,\omega(h) - S(h)]$, which yields (10). The increment follows from the closed form, and the limit from $q^h \to 0$ and $h\,q^h \to 0$. Appendix A gives the full proof. □

WIRS evaluates (10) for every source and refreshes up to B sources with the largest positive indices. Each evaluation takes constant time, followed by a top-B selection. With known rates, scheduling requires no policy training or value iteration. Appendix A.5 relates (10) to earlier crawling and age-of-information indices and adds optional per-call penalties.

Remark 1 (Non-monotone dependence on the change rate). At equal weight, a faster-changing source has higher priority with a small age but a lower ceiling. Its copy is likely to become stale sooner, and a refresh keeps it fresh for a shorter time. A slower source eventually takes priority if both wait long enough, giving the crossing curves in Figure 1.

Age alone cannot express this tradeoff. For $p_i h \ll 1$, $W_i(h) \approx g_i p_i\,h(h{+}1)/2$, so an old copy of a rarely changing source can have low priority. At large ages the index saturates, but weighted age $g_i h$ keeps growing. A higher rate estimate therefore lowers the highest priority the source can reach while it waits for its next visit (Section 4).

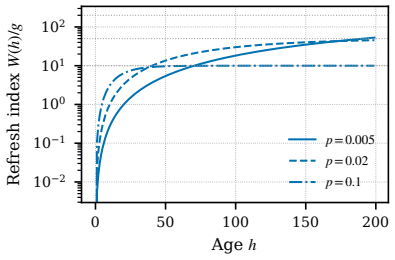


Figure 1: Normalized refresh index vs. age for three change rates, with dotted lines at the ceilings $1/p$.

A lower bound for evaluating schedules. The relaxation that yields the priorities also bounds the cost of every feasible policy from below, so the top-B schedule can be evaluated without solving the coupled problem.

Theorem 3 (Lagrangian lower bound). *For any $\nu \geq 0$, let $f_i^\star(\nu) \triangleq \inf_{\tau \in \mathbb{N}_{\geq 1}} \frac{g_i S_i(\tau)+\nu}{\tau}$, which is attained for $\nu < g_i/p_i$ and equals g_i otherwise. Then every policy π feasible for* (5) *satisfies*

$$J(\pi) \geq \max_{\nu \geq 0} \sum_{i=1}^{N} f_i^\star(\nu) - \nu B. \tag{11}$$

Proof sketch. For a feasible policy, subtracting the nonnegative budget slack $\nu\left(B - \sum_i a_i(t)\right) \geq 0$ from its cost gives a lower bound. The resulting cost, with price ν per visit, separates across sources, and each source's part is bounded below in the long run by $f_i^\star(\nu)$. Maximizing over ν gives (11). Appendix A handles the time-average limit using bounded relative values. □

The gap to the bound combines policy suboptimality with relaxation slack, so a small gap certifies near-optimality. Under additional conditions, Whittle policies and related alternatives are asymptotically optimal as the number of sources grows (Weber & Weiss, 1990; Verloop, 2016; Gast et al., 2023; Hong et al., 2023; Yan et al., 2025).

4 Learning unknown change rates

When rates are unknown, a visit both refreshes a copy and gives information about its source. At age h, it returns a Bernoulli$(\omega_i(h))$ observation y that indicates whether at least one change occurred, without its count or timing. The observation distribution depends on how long the policy waited. Rate estimation and scheduling are therefore coupled through the chosen visit times.

A higher rate estimate lowers the priority ceiling. A high rate estimate $\tilde{p}_i$ bounds a source's index by $g_i/\tilde{p}_i$, however long it waits (Remark 1), so other sources can keep winning the visits and leave no evidence to correct the estimate (Proposition 3). We therefore apply optimism directly to the index.

Optimism over plausible rates. UCB-WIRS retains the log-likelihood $\ell_i(p)$ of every point in a finite grid $\mathcal{G} \subset (0,1)$; only visits add likelihood terms. Before every decision it rebuilds $\mathcal{C}_i(t) = \{p \in \mathcal{G} : \ell_i(p) \geq \max_{p'} \ell_i(p') - \beta_t\}$ with $\beta_t = \alpha \log(\max\{t, 2\})$ from observations strictly before t, so excluded points can return. Here $\alpha > 0$ is fixed, with default $\alpha = 2$; the learning guarantees below assume $\alpha > 1$. At decision time t, the policy uses

$$\widetilde{W}_i(h) \;=\; \max_{p \in \mathcal{C}_i(t)} W_i(h; p), \tag{12}$$

Maximizing $W_i(h; p)$ can select an interior rate, whereas rate optimism always uses the largest one, $\tilde{p}_i = \max \mathcal{C}_i(t)$. A grid scan costs $O(N|\mathcal{G}|)$ index evaluations per slot with $O(N|\mathcal{G}|)$ stored likelihoods. Likelihoods are evaluated in closed form as $h \log(1-p)$ for no change and $\log[1-(1-p)^h]$ for a change.

Frozen and reopened sets. After finitely many changed observations, a *frozen* rule that rebuilds its set only after each visit can exclude every candidate able to win another visit, and then never visit the source again. Because likelihood differences stay fixed between visits while β_t grows, rebuilding at every decision restores the excluded candidate after any finite history and forces another visit (Proposition 4, Appendix A.8).

4.1 From recovery to learning guarantees

We now give conditions under which reopening yields visits at bounded ages that identify the rate and keep the expected excess bounded. In this subsection there are two sources and $B = 1$. Source 1 has known parameters, source 2 has unknown true rate p_*, and $c \triangleq g_1 p_1 \geq g_2$. We use Assumptions 1–3, exact likelihoods, fresh initialization, ties in favor of source 1, and Algorithm 1 with decision-time rebuilding and a fixed cutoff coefficient $\alpha > 1$. Write $S_p(L) = \sum_{h=0}^{L-1}[1-(1-p)^h]$.

Proposition 2 (Exact two-source optimum). *Under $c \geq g_2$, the optimal average cost over all feasible, possibly randomized and history-dependent policies is*

$$\rho_*(p) = \min_{L \in \mathbb{N}_{\geq 2} \cup \{\infty\}} J_L(p), \qquad J_L(p) = \frac{c + g_2 S_p(L)}{L}, \quad J_\infty(p) = g_2. \tag{13}$$

Algorithm 1 UCB-WIRS for refresh scheduling with unknown rates

Input. Weights g_i, budget B, rate grid $\mathcal{G}$, cutoff coefficient $\alpha > 0$ with default 2. Initialize $\ell_i(p) \leftarrow 0$ and $h_i \leftarrow 0$ for all sources.
for $t = 0, 1, 2, \ldots$ **do**
  Rebuild each $\mathcal{C}_i(t)$ from all stored likelihoods with $\beta_t = \alpha \log(\max\{t, 2\})$; compute $\widetilde{W}_i(h_i)$ by (12) and refresh the top-B with positive index.
  for each refreshed source i **do**
    Observe $y \sim \text{Bernoulli}(\omega_i(h_i))$; update $\ell_i(p)$ += $y \log \omega(h_i; p) + (1 - y) \log(1 - \omega(h_i; p))$; set $h_i \leftarrow 0$.
  end for
  $h_i \leftarrow h_i + 1$ for all i.
end for

A policy visiting source 2 *every minimizing finite* L *slots and source* 1 *otherwise attains this value. Known-rate WIRS implements such a policy, or never visits source* 2 *when* $c \geq g_2/p$.

At price $\nu = c$ the periodic policy attains the relaxation bound of Theorem 3, so in this class the hard-budget optimum is known exactly. For a horizon T, let $R_T \triangleq \mathbb{E}[\sum_{t<T} \sum_i g_i Z_i(t^+)] - T\rho_*(p_*)$ denote the expected excess over it; R_T can be negative for small T because both copies start fresh.

Theorem 4 (Learning and cost on a finite grid)**.** *Let* $p_* \in \mathcal{G}$*, where* $\mathcal{G} \subset (0, 1)$ *is finite. If* $c < g_2/p_*$*, then almost surely the true rate is eventually always retained, source* 2*'s subsequent inter-visit intervals are bounded by*

$$H_* \triangleq \min\{h \geq 1 : W_2(h; p_*) > c\}, \tag{14}$$

and $\mathcal{C}_2(t) = \{p_*\}$ *permanently after a time with finite expectation. Moreover,* $R_T = O(1)$ *in this case. If* $c \geq g_2/p_*$*, the visit frequency instead converges to zero almost surely and* $R_T/T \to 0$.

Likelihood ratios bound the expected number of slots at which the true rate is excluded; a counting argument charges every long wait to these slots, and bounded-age observations then separate the remaining candidates. The explicit bound in Appendix C separates only candidates that would cause a strictly premature refresh and ignores rates that give the same decisions.

Corollary 1 (Linear versus bounded expected excess)**.** *If* $p_* \in \mathcal{G}$ *and* $p_* < g_2/c \leq \max \mathcal{G}$*, the frozen-set index policy has* $R_T = \Omega(T)$*, whereas reopening has* $R_T = O(1)$*, for every finite grid satisfying these inequalities.*

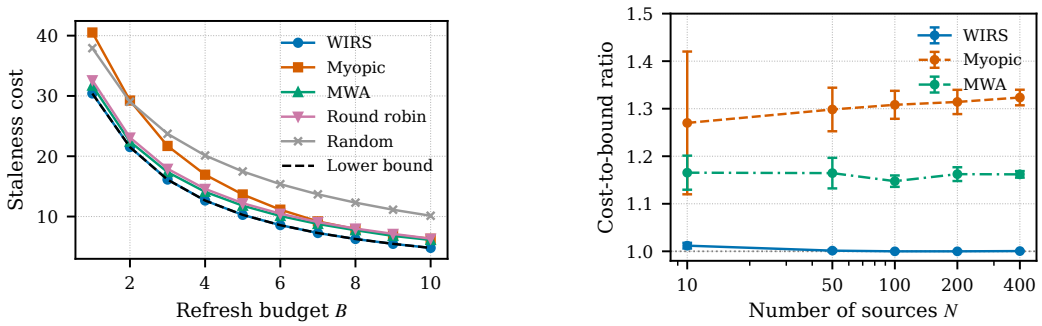
A finite run of changed observations can remove every useful candidate; freezing then gives limiting cost $g_2 > \rho_*(p_*)$, and a bounded relative value prevents other paths from cancelling this loss. The bounds hold for every cutoff $\beta_t = \alpha \log(\max\{t, 2\})$ with $\alpha > 1$. Appendix B extends the analysis to continuous rates, with a grid-spacing error bound, and to multiple unknown sources under a checkable capacity margin (Theorem 6).

5 Experiments

We first supply the true rates to test the index, and then hide them to compare index optimism with rate optimism. Synthetic sources let us measure staleness directly and compare costs with a lower bound. Perturbed dynamics and rate estimates test robustness to model error.

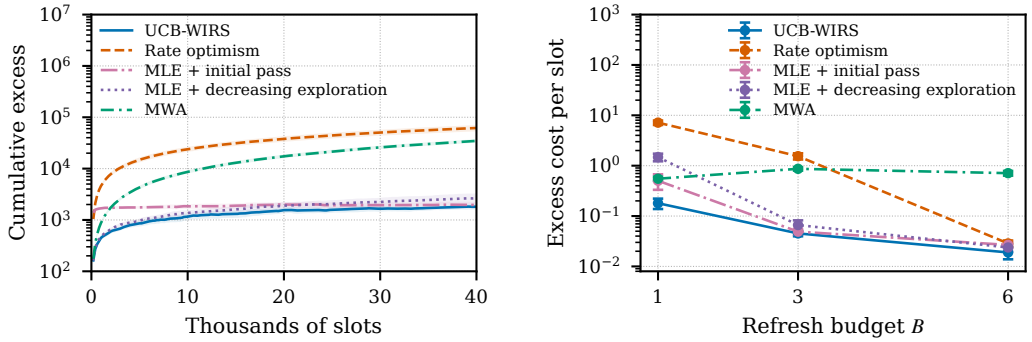
Sources and measurement. Unless stated otherwise, rates p_i are drawn log-uniformly on $[0.005, 0.2]$ and weights g_i uniformly on $[0.5, 2]$. Known-rate experiments run $T = 4 \times 10^4$ slots and exclude a 10% burn-in from steady-state averages. Learning experiments use $N = 30$ sources on independent instances with four paired change trajectories each, and accumulate cost from the first slot. Appendix H gives the configurations and numerical details.

Comparisons. *Myopic* ranks the immediate expected saving $g_i \omega_i(h_i)$, using the rates but omitting future benefit. *Maximum Weighted Age (MWA)* ranks $g_i h_i$ without rate estimates. *Round robin* and *Random* use no priorities. With hidden rates, *rate optimism* uses the largest plausible rate under



(a) (b)

Figure 2: Comparison of WIRS with myopic, MWA, and other baselines. (a) Staleness cost vs. refresh budget. (b) Cost-to-bound ratio vs. number of sources.



(a) (b)

Figure 3: Comparison of UCB-WIRS with rate optimism and other learners. (a) Cumulative excess over time at $B = 3$. (b) Excess cost per slot vs. refresh budget.

the same likelihood-ratio sets as UCB-WIRS, and *MLE-WIRS* plugs the grid maximum-likelihood estimate (MLE) into (10), optionally after an initial round-robin pass or with forced exploration (Appendix H.1).

WIRS vs. myopic and age-based rules. WIRS stays essentially at the lower bound across budgets (Figure 2a). Myopic scheduling uses the same rates and weights yet costs clearly more. It ignores how long each refresh keeps a copy fresh, so removing the most staleness now leaves more stale copies later. MWA is competitive when the budget is small but falls behind as the budget grows, because its age score keeps growing even when a copy is likely still fresh (Remark 1). Round robin and random also have larger gaps (Appendix H).

Scaling with the number of sources. We next hold the budget fraction fixed at $B = N/10$ and draw independent populations of increasing size (Figure 2b). WIRS stays close to the bound at every size, and its gap shrinks as N grows. Myopic and MWA keep a nearly constant gap above the bound.

Index optimism reduces the cost of learning. With rates hidden, UCB-WIRS approaches the cost of known-rate WIRS, and its cumulative excess grows slowly (Figure 3a). Rate optimism keeps paying extra cost in every period, so its excess grows steadily and becomes much larger. The gap persists because, as explained in Section 4, a high rate estimate suppresses the visits that would correct it. The advantage of index optimism is largest when the budget is tight and narrows as the budget grows (Figure 3b). MLE-WIRS with an initial pass is the strongest non-optimistic alternative. It matches index optimism at the middle budget and costs more at the smallest and largest budgets (Table 1).

Robustness to model error. With changes following Assumption 1, we perturb the supplied rates by log-normal noise of scale σ (Figure 4). The cost of WIRS rises slowly with moderate errors. MWA uses no rate estimates, so it is unaffected and overtakes WIRS once the errors become large. Myopic is much more sensitive to the same errors. Under hidden two-regime bursts with the same mean rate, WIRS keeps the lowest cost, and the advantage over MWA narrows as the bursts become stronger (Figure 5, Appendix H).

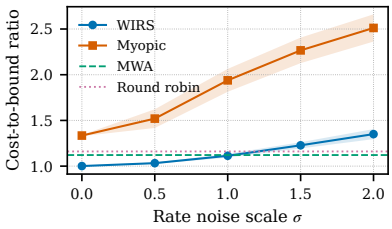


Figure 4: Cost-to-bound ratio vs. the noise scale of the supplied rates.

LLM agent task. An LLM agent, Qwen2.5-7B-Instruct with a lookup tool, answers questions from copies of 30 supplier records whose price and stock change at the rates of the main learning instances (Appendix I). Starting without observations, the scheduler refreshes at most B copies per slot. At regular intervals a customer orders from one supplier, chosen in proportion to its weight, and the agent returns the order total or `OUT OF STOCK`. The accuracy of the agent follows the staleness of the copies it reads. Index optimism answers about as many questions correctly as known-rate WIRS and clearly more than rate optimism, in every instance and at both tested budgets. The other schedulers rank in the same order as their staleness cost. With every copy fresh, most errors disappear, so stale copies cause most of the mistakes of the agent.

6 Related work

Freshness metrics and refresh scheduling. Mathematically, our model is closest to web-crawl scheduling. Refresh-allocation work studies how frequent changes shorten the benefit of a visit (Cho & Garcia-Molina, 2003; Azar et al., 2018) and derives revisit indices (Avrachenkov & Borkar, 2018); Appendix A.5 maps (10) to the known crawling index. Kolobov et al. (2019) learn update schedules from binary crawl observations and guard against data starvation through a positive allocation floor under an average-rate budget, and Kolobov et al. (2020) add non-resetting probes. In contrast, we analyze a hard slot budget and show when useful visits alone give enough information (Appendix G). Related observation-scheduling models measure age of information, age of incorrect information, uncertainty, estimation error, or cache freshness (Hsu, 2018; Tripathi & Modiano, 2024; Maatouk et al., 2020; Chen et al., 2022; Wang et al., 2020; Ornee & Sun, 2026; Koley & Singh, 2026); charging for a superseded copy gives the saturating index studied here.

Restless bandits and learning. Whittle's relaxation (Whittle, 1988) yields tractable source priorities for a computationally hard scheduling problem (Papadimitriou & Tsitsiklis, 1999), and our threshold results follow the restart-arm structure of Akbarzadeh & Mahajan (2019; 2022). Under unknown dynamics, indices can be learned through Q-learning or neural approximations (Avrachenkov & Borkar, 2022; Nakhleh et al., 2021; Nakhleh & Hou, 2022; Xiong & Li, 2023; Chen et al., 2026), and Wang et al. (2023) study optimistic models with fully observed states, including direct index maximization within a confidence region. Here the index has a closed form and evidence arrives only through refreshes, so learning reduces to estimating source rates. We study where to apply optimism when the index is non-monotone in the rate, and compare frozen and reopened candidate sets when the policy chooses its own observations.

Retrieval, memory, and tools in LLM agents. Adaptive retrieval uses confidence, learned critics, or bandits to decide when and how to retrieve (Asai et al., 2024; Jiang et al., 2023; Jeong et al., 2024; Su et al., 2024; Tang et al., 2025), and bandit methods also tune other LLM components such as speculative decoding (Hou et al., 2025). Memory systems manage storage, revision, and forgetting (Packer et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Xu et al., 2025; Yu et al., 2026), stale-memory detection identifies contradicted entries (Chao et al., 2026), and tool-use studies examine unnecessary calls (Ross et al., 2025). Our scheduler decides which retained copies to revisit under a limited visit budget.

7 SCOPE AND LIMITATIONS

Source dynamics, retrieval, and costs. Tracking staleness by age alone requires independent, memoryless changes and exact refreshes. Correlated sources, reverting values, and event-driven changes may require additional state, and noisy or delayed retrieval requires a belief state (Appendix F). The weights are fixed and known, whereas an agent's priorities can change during a task and must be calibrated. The budget treats calls as equal-sized visits; fixed call penalties fit Remark 2, but heterogeneous token costs and overlapping latencies require an allocation rule that reflects those resources.

Scope of the learning guarantees. The cost separation is proved for a two-source class with a known competitor. With multiple unknown sources, bounded cost holds under a capacity margin and a common cycle mean, and global optimality holds when the relaxation bound is tight.

From staleness to task performance. Our agent task uses one 7B model, synthetic record changes, and a scheduler outside the model. Letting the agent choose its own visits, and evaluating in dynamic environments such as Gaia2 and τ-bench and memory tasks such as LoCoMo (Froger et al., 2026; Yao et al., 2025; Maharana et al., 2024), requires defining observable sources and corrective visits and calibrating relevance weights to task loss.

8 CONCLUSION

A classical freshness index ranks refreshes by both current staleness and how long a refresh keeps the copy fresh, and it explains why priority reverses with the change rate as copies age. When rates are unknown, an upper rate estimate can suppress the visits needed to correct it. Maximizing the index over plausible rates avoids this problem. Rebuilding the candidate sets at every decision reduces the linear expected excess of frozen sets to bounded excess. This holds in a two-source class and, under capacity and cycle conditions, with multiple unknown sources. In an LLM-agent task, these gains carry over to answer accuracy.

AI USE STATEMENT

[TODO: Required] In this work, we used generative AI tools for searching for materials and polishing the paper writing. Additionally, we used generative AI tools for helping with writing the code for the experiment and simulation sections.

REFERENCES

Nima Akbarzadeh and Aditya Mahajan. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *IEEE Conference on Decision and Control (CDC)*, pp. 7294–7300, 2019.

Nima Akbarzadeh and Aditya Mahajan. Conditions for indexability of restless bandits and an algorithm to compute Whittle index. *Advances in Applied Probability*, 54(4):1164–1192, 2022.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11511.

Konstantin Avrachenkov and Vivek S. Borkar. Whittle index policy for crawling ephemeral content. *IEEE Transactions on Control of Network Systems*, 5(1):446–455, 2018. arXiv:1503.08558.

Konstantin E. Avrachenkov and Vivek S. Borkar. Whittle index based Q-learning for restless bandits with average reward. *Automatica*, 139:110186, 2022. arXiv:2004.14427.

Yossi Azar, Eric Horvitz, Eyal Lubetzky, Yuval Peres, and Dafna Shahaf. Tractable near-optimal policies for crawling. *Proceedings of the National Academy of Sciences*, 115(32):8099–8103, 2018.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Gongpu Chen, Soung Chang Liew, and Yulin Shao. Uncertainty-of-information scheduling: A restless multiarmed bandit framework. *IEEE Transactions on Information Theory*, 68(9):6151–6173, 2022. arXiv:2102.06384.

Gongpu Chen, Soung Chang Liew, and Deniz Gündüz. GINO-Q: Learning an asymptotically optimal index policy for restless multi-armed bandits. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 20032–20040, 2026. arXiv:2408.09882.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *European Conference on Artificial Intelligence (ECAI)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000, 2025. arXiv:2504.19413.

Junghoo Cho and Hector Garcia-Molina. Effective page refresh policies for Web crawlers. *ACM Transactions on Database Systems*, 28(4):390–426, 2003.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Ménard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.17158.

Nicolas Gast, Bruno Gaujal, and Chen Yan. Exponential asymptotic optimality of Whittle index policy. *Queueing Systems*, 104(1–2):107–150, 2023. arXiv:2012.09064.

Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Restless bandits with average reward: Breaking the uniform global attractor assumption. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12810–12844, 2023. arXiv:2306.00196.

Yunlong Hou, Fengzhuo Zhang, Cunxiao Du, Xuan Zhang, Jiachun Pan, Tianyu Pang, Chao Du, Vincent Y. F. Tan, and Zhuoran Yang. BanditSpec: Adaptive speculative decoding via bandit algorithms. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24045–24079, 2025. arXiv:2505.15141.

Yu-Pin Hsu. Age of information: Whittle index for scheduling stochastic arrivals. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2634–2638, 2018. arXiv:1801.03422.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 7036–7050, 2024. arXiv:2403.14403.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7969–7992, 2023. arXiv:2305.06983.

Ankita Koley and Chandramani Singh. Fresh caching of dynamic contents using restless multi-armed bandits over wireless access. *IEEE Transactions on Networking*, 34:1083–1095, 2026. arXiv:2404.12468.

Andrey Kolobov, Yuval Peres, Cheng Lu, and Eric Horvitz. Staying up to date with online content changes using reinforcement learning for scheduling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 581–591, 2019.

Andrey Kolobov, Sébastien Bubeck, and Julian Zimmert. Online learning for active cache synchronization. In *International Conference on Machine Learning (ICML)*, pp. 5371–5380, 2020. arXiv:2002.12014.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.

Ali Maatouk, Saad Kriouile, Mohamad Assaad, and Anthony Ephremides. The age of incorrect information: A new performance metric for status updates. *IEEE/ACM Transactions on Networking*, 28(5):2215–2228, 2020.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13851–13870, 2024. arXiv:2402.17753.

Rahul Meshram, D. Manjunath, and Aditya Gopalan. On the Whittle index for restless multiarmed hidden Markov bandits. *IEEE Transactions on Automatic Control*, 63(9):3046–3053, 2018.

Khaled Nakhleh and I-Hong Hou. DeepTOP: Deep threshold-optimal policy for MDPs and RMABs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 28734–28746, 2022. arXiv:2209.08646.

Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I-Hong Hou, and Srinivas Shakkottai. NeurWIN: Neural Whittle index network for restless bandits via deep RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 828–839, 2021. arXiv:2110.02128.

Tasmeen Zaman Ornee and Yin Sun. Remote estimation of Gauss-Markov processes over multiple channels: A Whittle index policy. *IEEE Transactions on Networking*, 34:5818–5830, 2026. arXiv:2305.04809.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. When2Call: When (not) to call tools. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 3391–3409, 2025. arXiv:2504.18851.

Linn I. Sennott. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Operations Research*, 37(4):626–633, 1989.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12991–13013, 2024. arXiv:2403.10081.

Xiaqiang Tang, Jian Li, Nan Du, and Sihong Xie. Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs. In *AAAI Conference on Artificial Intelligence*, pp. 12658–12666, 2025. arXiv:2412.07618.

Vishrant Tripathi and Eytan Modiano. A Whittle index approach to minimizing functions of age of information. *IEEE/ACM Transactions on Networking*, 32(6):5144–5158, 2024.

Ina Maria Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability*, 26(4):1947–1995, 2016.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024. arXiv:2310.03214.

Jiazheng Wang, Xiaoqiang Ren, Yilin Mo, and Ling Shi. Whittle index policy for dynamic multichannel allocation in remote state estimation. *IEEE Transactions on Automatic Control*, 65(2): 591–603, 2020.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic Whittle index policy: Online learning for restless bandits. In *AAAI Conference on Artificial Intelligence*, pp. 10131–10139, 2023. arXiv:2205.15372.

Richard R. Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

Peter Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.

Guojun Xiong and Jian Li. Finite-time analysis of Whittle index based Q-learning for restless multi-armed bandits with neural network function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 29048–29073, 2023. arXiv:2310.02147.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 17577–17604, 2025. arXiv:2502.12110.

Chen Yan, Weina Wang, and Lei Ying. Achieving $\tilde{\mathcal{O}}(1/N)$ optimality gap in restless bandits through Gaussian approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 51190–51212, 2025. arXiv:2410.15003.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.12045.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 21457–21483, 2026. arXiv:2601.01885.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. arXiv:2305.10250.

A Proofs

A.1 Proof of Proposition 1

Fix a source and drop the subscript i. Immediately after a refresh, $Z = 0$ with probability one by Assumption 2. If the source is not refreshed for h consecutive slots, then by Assumption 1 the event $\{Z = 0\}$ requires the change event to fail in each of the h slots. This event has probability q^h, independent of everything else in the system. Hence $\mathbb{P}(Z(t) = 1 \mid h(t) = h) = 1 - q^h$, which is (3). Refreshes reset h to 0 and, by independence in Assumption 1, reveal no information about other sources. The age vector is therefore a controlled Markov chain. By the tower property, the expected value of the realized cost (2) conditional on the ages equals $\sum_i g_i \omega_i(h_i(t^+))$. ■

A.2 Proof of Theorem 1

Fix $\nu \geq 0$ and a single source with parameters (g, p). Write $c(h) \triangleq g\,\omega(h) = g(1 - q^h)$, which is nonnegative, bounded by g, and strictly increasing in h.

Step 1. An optimal stationary policy exists and satisfies the average-cost optimality equation (ACOE). Consider first the γ-discounted problem, $\gamma \in (0, 1)$, with post-decision cost as in (7). The state space is $\mathbb{N}$, the action space is $\{0, 1\}$, costs are bounded, and the active action resets the state to 0. Hence the discounted value function V_γ exists and satisfies

$$V_\gamma(h) \;=\; \min\Big\{ \underbrace{\nu + \gamma V_\gamma(1)}_{\text{refresh}},\; \underbrace{c(h) + \gamma V_\gamma(h{+}1)}_{\text{wait}} \Big\}, \tag{15}$$

where a refreshed source has post-decision cost $c(0) = 0$ and age 1 at the next decision. Value iteration from $V \equiv 0$ preserves monotonicity in h, because each iterate is the minimum of a constant refresh value and a nondecreasing wait value. Hence V_γ is nondecreasing. Bounded one-step costs also give $V_\gamma(h) \le (\nu + g)/(1-\gamma)$.

The relative values satisfy a stronger bound. At $h = 0$, both actions lead to age 1, but refreshing pays ν extra, so $V_\gamma(0) = \gamma V_\gamma(1)$. Refreshing immediately from any age gives $V_\gamma(h) \le \nu + \gamma V_\gamma(1) = \nu + V_\gamma(0)$. Combining this inequality with monotonicity yields

$$0 \;\le\; V_\gamma(h) - V_\gamma(0) \;\le\; \nu \qquad \text{for all } h \text{ and all } \gamma \in (0,1). \tag{16}$$

Write $w_\gamma(h) \triangleq V_\gamma(h) - V_\gamma(0)$ and rearrange (15) as $(1-\gamma)V_\gamma(0) + w_\gamma(h) = \min\{\nu + \gamma w_\gamma(1),\ c(h) + \gamma w_\gamma(h{+}1)\}$. By (16) and $0 \le (1-\gamma)V_\gamma(0) \le \nu + g$, a diagonal argument over the countable state space extracts a sequence $\gamma_k \uparrow 1$ along which $(1-\gamma_k)V_{\gamma_k}(0) \to \rho$ and $w_{\gamma_k} \to v$ pointwise, with v nondecreasing and $0 \le v \le \nu$. Since the minimum is over two actions whose values converge pointwise, the limit satisfies the ACOE

$$\rho + v(h) \;=\; \min\big\{\nu + v(1),\ \ c(h) + v(h{+}1)\big\}, \tag{17}$$

whose minimizing actions define an optimal stationary policy, and ρ is the optimal average cost. This follows from a standard verification argument under bounded costs and the uniformly bounded relative values (16) (Sennott, 1989).

Step 2. The optimal stationary policy is a threshold. Define $\Delta(h) \triangleq \big(c(h) + v(h{+}1)\big) - \big(\nu + v(1)\big)$, so waiting minimizes the right-hand side of (17) when $\Delta(h) \le 0$. Since c is strictly increasing and v is nondecreasing, Δ is increasing in h. Choosing to wait at ties therefore gives a waiting region $\{0, 1, \dots, \tau - 1\}$ and an optimal policy that refreshes if and only if $h \ge \tau$. Here $\tau \in \mathbb{N}_{\ge 1} \cup \{\infty\}$, with $\tau = \infty$ denoting no refreshes. The region contains age zero because $\Delta(0) = v(1) - \nu - v(1) = -\nu \le 0$. Step 3 compares the average costs of all thresholds.

Step 3. The optimal threshold solves (8) **and is monotone in ν.** Under threshold $\tau < \infty$, the age process is a deterministic cycle $1, 2, \dots, \tau$ of length τ with post-decision ages $0, 1, \dots, \tau{-}1$. Each cycle incurs staleness cost $g\,S(\tau) = g\sum_{h=0}^{\tau-1}\omega(h)$ and one refresh charge ν. By the renewal-reward theorem, its average cost is $A_\nu(\tau) = \frac{gS(\tau)+\nu}{\tau}$. The never-refresh policy has average cost $g = \lim_{\tau\to\infty} A_\nu(\tau) = A_\nu(\infty)$, so it suffices to minimize A_ν over $\mathbb{N}_{\ge 1}\cup\{\infty\}$. A direct computation gives the exact difference identity

$$A_\nu(\tau{+}1) - A_\nu(\tau) \;=\; \frac{m(\tau) - \nu}{\tau(\tau+1)}, \qquad m(\tau) = g\big[\tau\,\omega(\tau) - S(\tau)\big] = g\sum_{k<\tau}\big[\omega(\tau) - \omega(k)\big],$$

and m is strictly increasing in τ with $m(0) = 0$, since $m(\tau) - m(\tau{-}1) = \tau\big[\omega(\tau) - \omega(\tau{-}1)\big]g = g\,p\,\tau\,q^{\tau-1} > 0$. Moreover, $m(\tau) = g\big[\tau(1-q^\tau) - S(\tau)\big] \uparrow g/p$. Hence A_ν is unimodal. It weakly decreases while $m(\tau) \le \nu$ and increases strictly once $m(\tau) > \nu$. If $\nu < g/p$, some τ with $m(\tau) \ge \nu$ exists and the set of minimizers is exactly the nonempty interval $\{\tau :\ m(\tau{-}1) \le \nu \le m(\tau)\}$, whose endpoints are nondecreasing in ν because m is nondecreasing. Every minimizer has average cost below $A_\nu(\infty) = g$, since A_ν eventually increases towards its limit g. If $\nu \ge g/p$, then $m(\tau) < \nu$ for all finite τ because $m(\tau) < g/p$ strictly, so A_ν is strictly decreasing on $\mathbb{N}_{\ge 1}$ with limit g. Consequently, no finite minimizer exists and $A_\nu(\tau) > g$ for all τ, which is equivalent to $\nu > g(1-q^\tau)/p$. The unique minimizer is therefore $\tau = \infty$. ■

A.3 Proof of Theorem 2

Indexability requires the passive set $\Pi(\nu) \triangleq \{h :$ waiting optimal at $h\}$ to grow monotonically to all of $\mathbb{N}$ as ν increases, starting here from $\Pi(0) = \{0\}$ because age 0 is always passive. We say waiting is optimal at h if some optimal stationary *threshold* policy waits at h. Equivalently, waiting is optimal at h if it attains the minimum in (17) at h, that is, if $h < \tau_+(\nu)$, where $\tau_+(\nu) \in \mathbb{N}_{\ge 1}\cup\{\infty\}$ is the largest minimizer in (8). By Theorem 1, $\tau_+(\nu) = \max\{\tau : m(\tau{-}1) \le \nu\}$ for $\nu < g/p$ and $\tau_+(\nu) = \infty$ otherwise. In both cases τ_+ is nondecreasing in ν, so $\Pi(\nu) = \{0, \dots, \tau_+(\nu) - 1\}$ grows monotonically. At $\nu = 0$, we have $m(0) = 0$ and $m(\tau) > 0$ for $\tau \ge 1$, so $\tau_+(0) = 1$ and

$\Pi(0) = \{0\}$. Since $\sup_\tau m(\tau) = \lim_{\tau\to\infty} m(\tau) = g/p$, every age is passive once $\nu \geq g/p$, so $\Pi(\nu) \uparrow \mathbb{N}$. This is Whittle's indexability condition.

The Whittle index of age h is $W(h) = \inf\{\nu : \ h \in \Pi(\nu)\} = \inf\{\nu : \ \tau_+(\nu) > h\}$. By the interval characterization, $\tau_+(\nu) > h$ if and only if $m(h) \leq \nu$. Indeed, if $m(h) \leq \nu$ then $\tau = h{+}1$ satisfies $m(\tau{-}1) \leq \nu$, so $\tau_+ \geq h{+}1$. Conversely, $\tau_+ \geq h{+}1$ forces $m(h) \leq \nu$. Therefore $W(h) = m(h) = g\big[h\,\omega(h) - S(h)\big]$. Substituting $\omega(h) = 1 - q^h$ and $S(h) = h - \frac{1-q^h}{p}$ gives

$$W(h) = g\Big[h(1-q^h) - h + \tfrac{1-q^h}{p}\Big] = g\Big[\tfrac{1-q^h}{p} - h\,q^h\Big] = g\sum_{k=0}^{h-1}\big(q^k - q^h\big),$$

where the last equality uses $\sum_{k<h} q^k = \frac{1-q^h}{p}$. Since $(q^h - q^{h+1})/p = q^h$, the closed form gives

$$W(h{+}1) - W(h) = g\big[q^h + h\,q^h - (h{+}1)\,q^{h+1}\big] = g\,p\,(h{+}1)\,q^h > 0,$$

which is also $m(h{+}1){-}m(h)$ from Step 3 of the proof of Theorem 1. Hence W is strictly increasing in h. For the limit, $(1{-}q^h)/p \to 1/p$ and $h\,q^h \to 0$ as $h \to \infty$, so $W(h) \to g/p$. Finally, $W(0) = 0$ from the empty sum. ■

A.4 Proof of Theorem 3

Let π be feasible for (5) and $\nu \geq 0$. Since $\sum_i a_i(t) \leq B$ almost surely, for every T,

$$\begin{aligned}\frac{1}{T}\mathbb{E}\sum_{t<T}\sum_i g_i\omega_i(h_i(t^+)) \ &\geq \ \frac{1}{T}\mathbb{E}\sum_{t<T}\sum_i \big(g_i\omega_i(h_i(t^+)) + \nu a_i(t)\big) - \nu B \\ &= \ \sum_i \frac{1}{T}\mathbb{E}\sum_{t<T}\big(g_i\omega_i(h_i(t^+)) + \nu a_i(t)\big) - \nu B.\end{aligned}$$

Fix a source i and let (ρ_i, v_i) solve its ACOE (17) with $0 \leq v_i \leq \nu$, as constructed in the proof of Theorem 1. By Theorem 1, $\rho_i = f_i^\star(\nu)$. The policy π induces for source i a refresh sequence $\{a_i(t)\}$ that may be randomized, history-dependent, and coupled to the other sources. However, the age h_i evolves *deterministically* given a_i, resetting to 0 on refresh and incrementing otherwise, and the coupling randomness is non-anticipating. Because (17) takes a minimum over both actions, along every trajectory and for every t,

$$g_i\omega_i\big(h_i(t^+)\big) + \nu a_i(t) + v_i\big(h_i(t{+}1)\big) \ \geq \ \rho_i + v_i\big(h_i(t)\big).$$

Summing over $t < T$, taking expectations, and telescoping,

$$\mathbb{E}\sum_{t<T}\big[g_i\omega_i(h_i(t^+)) + \nu a_i(t)\big] \ \geq \ T\rho_i + \mathbb{E}\,v_i(h_i(0)) - \mathbb{E}\,v_i(h_i(T)) \ \geq \ T\rho_i - \nu,$$

so the source's time average is at least $\rho_i - \nu/T = f_i^\star(\nu) - \nu/T$. Combining over sources, taking $\limsup_{T\to\infty}$, and then maximizing over $\nu \geq 0$ yields (11). The maximum is attained because the dual is concave with value 0 at $\nu = 0$ and tends to $-\infty$. ■

A.5 Existing refresh indices and per-refresh prices

On reachable crawl states, Theorem 1 of Avrachenkov & Borkar (2018) gives (10) under the change of variables $q \leftrightarrow \alpha$, $g \leftrightarrow u/C$. The age-arm index of Tripathi & Modiano (2024) gives the same formula after adjusting the timing convention. Restart-arm results provide related threshold and indexability properties (Akbarzadeh & Mahajan, 2019; 2022).

At equal weight the index is non-monotone in the rate for every fixed $h \geq 2$. Indeed, $W_i(1) = g_i p_i$ favors fast sources at age one, while the ceiling g_i/p_i favors slow ones. With unequal weights, the ceilings g_i/p_i determine the limiting comparison. These properties recover the allocation tradeoff studied in web crawling (Cho & Garcia-Molina, 2003; Azar et al., 2018).

Remark 2 (Per-refresh prices). If the objective includes a fixed penalty $\kappa_i \geq 0$ for refreshing source i, the single-source price becomes $\nu + \kappa_i$ and the net index is $W_i(h) - \kappa_i$. This prices individual calls while retaining the count budget (1).

A.6 Proof of Proposition 3

Proposition 3 (Rate optimism can suppress all future observations)**.** *Consider two sources with $B = 1$. Source 1 has known (g_1, p_1) and initial age at least one. Source 2 has a nonempty finite plausible-rate set $\mathcal{C}_2 \subset (0, 1)$, retained until its next visit. Write $u = \max \mathcal{C}_2$. If some $p_0 \in \mathcal{C}_2$ satisfies $g_2/u \leq g_1 p_1 < g_2/p_0$, scheduling by the largest plausible rate never visits source 2. Scheduling by the largest plausible index visits it in finite time.*

The rate-optimistic index stays below g_2/u, while repeatedly visiting source 1 leaves its index at $g_1 p_1$. Index optimism can instead use p_0, whose limiting priority is higher. The construction holds source 2's set fixed between visits, as the frozen rule does.

Corollary 2 (A positive cost of starvation)**.** *Under Proposition 3, suppose the true rate is $p_2 = p_0$. Rate optimism has average cost g_2. A feasible policy visiting source 2 every $L \geq 2$ slots and source 1 otherwise improves this cost by*

$$\Gamma_L = \frac{(g_2/p_0)\big[1 - (1 - p_0)^L\big] - g_1 p_1}{L} > 0 \tag{18}$$

whenever $(1 - p_0)^L < 1 - g_1 p_1 p_0/g_2$. Such a finite L exists.

Corollary 2 gives a case where missing observations strictly increase cost. It specifies the true rate because ignoring a sufficiently volatile source can be optimal.

Proof of Proposition 3. For every finite age h, $W_2(h; u) < g_2/u \leq g_1 p_1$. Source 1, whose initial age is at least one, has index at least $W_1(1; p_1) = g_1 p_1$. Rate optimism therefore visits source 1 first and returns it to age one at the next decision. Induction gives the same strict priority comparison forever. Source 2 receives no observations, so its set stays unchanged.

For index optimism, suppose instead that source 2 is never visited. Source 1 is then visited at every slot and has priority $g_1 p_1$ after the first decision. Meanwhile source 2's age tends to infinity, and its optimistic priority is at least $W_2(h; p_0) \to g_2/p_0 > g_1 p_1$. At a finite age it must exceed source 1's priority, a contradiction. This proves the first-visit claim. □

A.7 Proof of Corollary 2

Under permanent starvation, source 1 contributes zero cost and source 2's staleness probability tends to one, giving average cost g_2. In the periodic policy, source 1 is skipped once per L slots. For $L \geq 2$, it was refreshed in the preceding slot, so this omission costs $g_1 p_1$. Source 2 contributes $g_2 S_2(L)$ per cycle. Thus the periodic average is $[g_1 p_1 + g_2 S_2(L)]/L$, and subtraction from g_2 gives (18). The strict inequality is equivalent to $(1 - p_0)^L < 1 - g_1 p_1 p_0/g_2$. Its right side is positive by Proposition 3, and its left side tends to zero, so a finite qualifying $L \geq 2$ exists. ■

A.8 Proof of Proposition 4

Proposition 4 (Recovery after a finite history)**.** *Consider two sources with $B = 1$. Let source 1 have known (g_1, p_1) and let source 2 use Algorithm 1 with any fixed cutoff coefficient $\alpha > 0$. If some $p_0 \in \mathcal{G}$ satisfies $g_2/p_0 > g_1 p_1$, source 2 is visited infinitely often on every path with finite log-likelihoods after finite histories. Following a visit at s, define $\Delta_s(p_0) = \max_{p \in \mathcal{G}} \ell_{2,s^+}(p) - \ell_{2,s^+}(p_0)$, where s^+ denotes the likelihood immediately after that visit, and $H = \min\{h \geq 1 : W_2(h; p_0) > g_1 p_1\}$. Its next visit occurs no later than*

$$\max\left\{s + 2,\ \left\lceil \exp\big(\Delta_s(p_0)/\alpha\big) \right\rceil,\ s + H\right\}. \tag{19}$$

Proof. After a finite history at source 2's last visit s, every stored likelihood is finite. If no further visit occurs, these values remain fixed. Candidate p_0 belongs to the decision-time set once $\alpha \log(\max\{t, 2\}) \geq \Delta_s(p_0)$, which holds by $t = \max\{s + 1, \lceil \exp(\Delta_s(p_0)/\alpha) \rceil\}$. This step relies on rebuilding the set from the full grid.

Let T_s denote (19). Suppose source 2 is not visited before T_s. Source 1 is visited at every intervening slot, and $T_s \geq s + 2$ ensures that its decision-time age is one. Source 2 has age $T_s - s \geq H$, and p_0

is back in its plausible set. Hence its optimistic index is at least $W_2(H; p_0) > g_1 p_1$, forcing a visit by T_s. The same argument from the initial finite likelihood gives a first visit. Applying it after each visit proves infinitely many visits. The argument is pathwise. ■

The recovery bound grows exponentially in the likelihood gap. Without new observations, a candidate with gap Δ returns only once $\max\{t, 2\} \geq \exp(\Delta/\alpha)$. Long waits also make observations weakly informative. For one observation at age h, the Fisher information about p is $h^2(1-p)^{h-2}/[1-(1-p)^h]$, which tends to zero as h grows. Under the two-source assumptions of Section 4.1, including $\alpha > 1$, suppose that $p_* \in \mathcal{G}$ and $c < g_2/p_*$. Then Theorem 4 strengthens Proposition 4. It shows that visit intervals are almost surely eventually bounded and that the plausible set eventually equals $\{p_*\}$.

B Off-grid rates and multiple unknown sources

B.1 Continuous rates and decision accuracy

The two-source guarantees of Section 4.1 assume that the true rate is a grid point. We retain the other assumptions, including a fixed cutoff coefficient $\alpha > 1$, but now let p_* lie strictly between grid points and bound the cost of using a finite grid.

Let $p_L < p_* < p_U$ be adjacent grid points. The likelihood ratio of an exterior grid point to its nearest bracketing point is a supermartingale under p_*.

Theorem 5 (Off-grid decisions). *If $c < g_2/p_*$, the average cost satisfies*

$$0 \leq J(\text{UCB-WIRS}) - \rho_*(p_*) \leq g_2 \log(p_U/p_L). \tag{20}$$

If also $c < g_2/p_U$, visit intervals are eventually bounded and $\mathcal{C}_2(t) \subseteq \{p_L, p_U\}$ permanently after a time with finite expectation. If, for an integer $L \geq 2$,

$$W_2(L-1; p) < c < W_2(L; p) \quad \textit{for every } p \in [p_L, p_U], \tag{21}$$

then the algorithm eventually uses the true optimal threshold L, and $R_T = O(1)$.

Condition (21) concerns the entire interval because the index is not monotone in the rate. For a bracket that crosses g_2/c, the cost bound (20) still holds (Appendix C.5).

B.2 Information and cost with multiple unknown sources

We now allow arbitrary N and $1 \leq B < N$, with all source rates unknown. For each source i whose true rate is covered by the finite grid, let $\mathcal{B}_i = \{p_i\}$ when $p_i \in \mathcal{G}$. Otherwise, let $\mathcal{B}_i = \{p_{L,i}, p_{U,i}\}$ be its adjacent bracketing points. We retain the independent change model, exact likelihoods, and decision-time reopening with $\beta_t = \alpha \log(\max\{t, 2\})$ for fixed $\alpha > 1$, with default $\alpha = 2$. We drop the two-source weight condition $g_1 p_1 \geq g_2$.

For a priority level $c > 0$, define each competitor's full-grid envelope and required age by

$$U_j(h) = \max_{p \in \mathcal{G}} W_j(h; p), \qquad d_j(c) = \min\{h \geq 1 : U_j(h) \geq c\}, \tag{22}$$

with $d_j(c) = \infty$ if the set is empty and $1/\infty = 0$. For target i, choose

$$0 < c_i < \min_{p \in \mathcal{B}_i} \frac{g_i}{p}, \qquad \min_{p \in \mathcal{B}_i} W_i(H_i; p) > c_i, \qquad \Lambda_i(c_i) = \sum_{j \neq i} \frac{1}{d_j(c_i)}. \tag{23}$$

A finite integer H_i exists by the strict ceiling inequality. For an off-grid source, that inequality is $c_i < g_i/p_{U,i}$.

Proposition 5 (Sourcewise information under a capacity margin). *If $\Lambda_i(c_i) < B$, then almost surely all sufficiently late inter-visit intervals of source i are at most*

$$D_i = H_i + \left\lceil \frac{N-1}{B - \Lambda_i(c_i)} \right\rceil + 1, \tag{24}$$

and eventually $\varnothing \neq \mathcal{C}_i(t) \subseteq \mathcal{B}_i$. The other sources can remain unknown and need not satisfy their own capacity conditions.

A competitor needs at least $d_j(c_i)$ slots between visits scoring at least c_i. Intermediate refreshes only postpone its next such visit. Thus $\Lambda_i(c_i)$ bounds the average number of such competitor visits per slot. A positive budget margin prevents them from blocking an old target indefinitely. Taking envelopes over the full grid covers competitors whose plausible sets are temporarily overoptimistic. Appendix D accounts for arbitrary ages at the start of the counting window and gives a sharper integer bound.

We call source i certified if $\Lambda_i(c_i) < B$. For every cutoff coefficient $\alpha > 1$, each certified source also has a finite expected time to permanent localization. When all sources are on-grid and certified, the expected number of actions that differ from known-rate WIRS at the learner's own ages is finite. To bound the cost, we must also control the schedule reached after learning.

Theorem 6 (Multi-source bounded learning cost)**.** *Suppose every source is on-grid and satisfies the capacity condition, with exact likelihoods, fresh initialization, $\alpha > 1$, and a fixed source-ID tie rule. Known-rate WIRS has a finite absorbing age graph. If all recurrent cycles in this graph have the same average cost ρ_W, then*

$$\sup_{T \geq 1} \left| \mathbb{E} \sum_{t<T} \sum_{i} g_i Z_i(t^+) - T\rho_W \right| < \infty.$$

If ρ_W also equals the relaxation lower bound in Theorem 3, it is the globally optimal hard-budget average cost.

Appendix E constructs the finite graph and a bounded relative value from its common cycle mean. The construction allows multiple recurrent cycles and bounds the cost of every action that differs from known-rate WIRS.

C Proofs of the learning guarantees

C.1 Timing, filtration, and exact likelihoods

We consider the two-source model of Section 4.1. Source 1 has known parameters, source 2 has fixed true rate p_*, and $c = g_1 p_1 \geq g_2$. Both copies are fresh at $t = 0$. All likelihoods initially equal zero. A decision at time t uses only observations obtained strictly before t. The observation produced by that decision is incorporated before the decision at $t+1$. Indices are compared exactly, with ties assigned to source 1.

Let $\mathcal{F}_t$ be the decision-time observation history, including past actions, received observations, and any policy randomization. Changes and losses enter $\mathcal{F}_t$ only through observations. Write

$$\mu_p(h) = 1 - (1-p)^h, \qquad P_p(y \mid h) = \mu_p(h)^y [1 - \mu_p(h)]^{1-y}, \quad y \in \{0, 1\}. \tag{25}$$

If source 2 is visited at age h, its observation has conditional distribution $P_{p_*}(\cdot \mid h)$. The age is $\mathcal{F}_t$-measurable. Since changes after the previous reset are independent of $\mathcal{F}_t$, this conditional law holds for adaptively chosen visit times. Every visit occurs at a positive age.

Throughout these proofs, $\ell_t(p)$ denotes the sum of exact log probabilities from source 2 observations strictly before t. In particular, an unchanged observation at age h contributes $h \log(1-p)$. All probabilities in (25) are strictly between zero and one at finite positive ages, so every finite-history likelihood is finite. We abbreviate $\mathcal{C}_2(t)$ by $\mathcal{C}_t$.

C.2 Proof of Proposition 2

Set the Lagrangian visit price to $\nu = c$. Since $W_1(1; p_1) = g_1 p_1 = c$, the threshold characterization in Theorem 1 gives

$$f_1^*(c) = c, \qquad \text{with thresholds 1 and 2 both optimal.} \tag{26}$$

For source 2, $W_2(1; p) = g_2 p < g_2 \leq c$. Consequently, its optimal threshold at visit price c is either at least two or infinite. Thus

$$f_2^*(c) = \min_{L \in \mathbb{N}_{\geq 2} \cup \{\infty\}} \frac{c + g_2 S_p(L)}{L},$$

where the value at infinity is g_2. Theorem 3 applies to all feasible randomized and history-dependent policies and yields

$$J(\pi) \geq f_1^*(c) + f_2^*(c) - c = f_2^*(c).$$

For a finite minimizing $L \geq 2$, visiting source 2 every L slots and source 1 otherwise is feasible. Every source 2 visit is isolated, so source 1 was refreshed in the preceding slot and incurs expected cost exactly c in the skipped slot. Source 2 incurs $g_2 S_p(L)$ per cycle. This policy therefore attains $f_2^*(c)$. If the minimizing threshold is infinite, always visiting source 1 attains g_2. This proves (13) and global optimality.

Known-rate WIRS cannot visit source 2 in consecutive slots, because at age one its index is below c, while source 1's index is at least c whenever its decision age is positive. At all remaining decisions relevant to a source 2 visit, source 1 has age one and index c. If $c < g_2/p$, WIRS therefore uses

$$H(p) = \min\{h \geq 1 : W_2(h; p) > c\} \geq 2.$$

The threshold characterization shows that this is an optimal threshold. If $W_2(h; p) = c$, waiting selects the larger of the two adjacent minimizing thresholds. If $c \geq g_2/p$, every finite-age source 2 index is strictly below c, and WIRS never visits it.

The arguments below use $\beta_t = \alpha \log(\max\{t, 2\})$ with fixed $\alpha > 1$ and default $\alpha = 2$. Write $H(p) = \min\{h \geq 1 : W_2(h; p) > c\}$, with $H(p) = \infty$ when $c \geq g_2/p$.

C.3 Finite expected time to localization

Take the basic reference set $\mathcal{B}_0 = \{p_*\}$ if $p_* \in \mathcal{G}$. Otherwise let $p_L < p_* < p_U$ be adjacent grid points and take $\mathcal{B}_0 = \{p_L, p_U\}$. A reference set $\mathcal{R}$ may be any superset of $\mathcal{B}_0$ in $\mathcal{G}$ such that every $b \in \mathcal{R}$ is useful, that is, $c < g_2/b$. Set

$$\mathcal{Q} = \mathcal{G} \setminus \mathcal{R}, \quad M = |\mathcal{Q}|, \quad D = \max_{b \in \mathcal{R}} H(b) \geq 2, \quad A = 2(D+1), \quad z_\alpha = \sum_{t=2}^{\infty} t^{-\alpha}. \tag{27}$$

For each $p \in \mathcal{Q}$, its anchor is $b(p) = p_*$ on-grid, and is p_L for $p < p_L$ or p_U for $p > p_U$ off-grid. The adjacency assumption makes this assignment exhaustive. More generally, the same anchors may be retained after inserting grid points between them, provided every intervening grid point belongs to $\mathcal{R}$. Every exterior point then still lies outside the bracket. For $M > 0$ define

$$a_h(p) = \mathbb{E}_{p_*} \sqrt{P_p(Y \mid h)/P_{b(p)}(Y \mid h)}. \tag{28}$$

$$r_p = \max_{1 \leq h \leq D} a_h(p) \in (0, 1), \qquad a_p = r_p^{1/(2D)}. \tag{29}$$

$$\Phi_\alpha(a) = \sum_{t=2}^{\infty} t^{\alpha/2} a^t, \qquad \Psi_\alpha(a) = \sum_{t=2}^{\infty} t^{1+\alpha/2} a^t. \tag{30}$$

These are finite, instance-dependent constants. Since no target visit occurs before age $\min_{p \in \mathcal{G}} H(p)$, the maximum defining r_p may be restricted to these attainable ages.

Theorem 7 (Finite expected localization time)**.** *Under the preceding assumptions, let*

$$K_\infty = \sum_{t=0}^{\infty} \mathbf{1}\{\mathcal{C}_t \cap \mathcal{R} = \varnothing\}, \qquad \tau_{\mathcal{R}} = \inf\{n \geq 0 : \mathcal{C}_t \subseteq \mathcal{R} \text{ for every } t \geq n\}.$$

Then $\mathbb{E}K_\infty \leq M z_\alpha$, and

$$\mathbb{E} \sum_{t=0}^{\infty} \mathbf{1}\{\mathcal{C}_t \not\subseteq \mathcal{R}\} \leq 2 + AMz_\alpha + \sum_{p \in \mathcal{Q}} r_p^{-1} \Phi_\alpha(a_p), \tag{31}$$

$$\mathbb{E}\tau_{\mathcal{R}} \leq 2 + AMz_\alpha + \sum_{p \in \mathcal{Q}} r_p^{-1} \Psi_\alpha(a_p) < \infty. \tag{32}$$

If $M = 0$, both random quantities on the left are zero and no affinity constant is needed. In particular, a useful on-grid truth is identified permanently after a time with finite expectation. A useful off-grid bracketing pair admits the analogous localization guarantee.

Proof. Write

$$L_t(p) = \exp\{\ell_t(p) - \ell_t(b(p))\}. \tag{33}$$

On-grid this is a mean-one nonnegative martingale. Off-grid, putting $\theta = 1-(1-p_*)^h$ and defining θ_b, θ_p similarly, its conditional multiplicative mean is

$$1 + \frac{(\theta - \theta_b)(\theta_p - \theta_b)}{\theta_b(1-\theta_b)} < 1. \tag{34}$$

Thus it is a nonnegative supermartingale. If $\mathcal{C}_t$ excludes every point of $\mathcal{R}$, a likelihood maximizer p lies in $\mathcal{Q}$ and exceeds its excluded anchor by more than the cutoff. Hence

$$\mathbb{P}\{\mathcal{C}_t \cap \mathcal{R} = \varnothing\} \le Mt^{-\alpha}, \quad t \ge 2. \tag{35}$$

At times zero and one all likelihoods are equal. Summing over $t \ge 2$ proves the bound on $\mathbb{E}K_\infty$.

We next prove a deterministic counting bound. At decision t, let $h_1, \dots, h_n$ be the completed target inter-visit intervals, including the interval from the virtual reset at zero to its first actual visit. All these visits occurred strictly before t. The unfinished age is a_t, so $t = \sum_{k=1}^n h_k + a_t$ exactly. Let $N_D(t)$ count the completed intervals of length at most D, and let K_t count reference failures at decisions strictly before t.

In a completed interval of length $h > D$, the target was skipped at ages $D, \dots, h-1$. In each such slot, its competitor was refreshed in the preceding slot and has index c, because these ages are at least two and every positive-time decision refreshes one source. If a reference point were present, its index would exceed c, a contradiction. These $h - D$ slots are therefore reference failures, and $h \le (D+1)(h-D)$. The unfinished interval similarly has at least $(a_t - D)^+$ such failures before t, so its age is at most D plus its failure count. Disjoint intervals use disjoint slots. Consequently, on every trajectory,

$$t \le DN_D(t) + (D+1)K_t + D. \tag{36}$$

In particular,

$$K_\infty \le t/A \quad \Longrightarrow \quad N_D(t) \ge t/(2D) - 1. \tag{37}$$

All $a_h(p)$ lie strictly between zero and one. On-grid, $a_h(p)$ is the affinity of two distinct Bernoulli laws. Off-grid, the bound follows from Cauchy–Schwarz and the strict conditional likelihood-ratio mean above. At each visit divide the square-root likelihood increment by its predictable mean $a_h(p)$. Thus

$$Z_t(p) = \frac{\sqrt{L_t(p)}}{\prod_{k:\, s_k < t} a_{h_k}(p)}$$

is a mean-one nonnegative martingale, since each visit age h_k is $\mathcal{F}_{s_k}$-measurable. At a fixed $t \ge 2$, an exterior point in $\mathcal{C}_t$ has $L_t(p) \ge t^{-\alpha}$. If also $N_D(t) \ge m$, the denominator is at most r_p^m, because all other factors are at most one. Markov's inequality gives

$$\mathbb{P}\{L_t(p) \ge t^{-\alpha},\ N_D(t) \ge m\} \le t^{\alpha/2} r_p^m. \tag{38}$$

The same bound holds with $p \in \mathcal{C}_t$ in place of the likelihood-ratio event. Let

$$F_t = \{\mathcal{C}_t \not\subseteq \mathcal{R},\ N_D(t) \ge t/(2D) - 1\}.$$

Since the bound holds for every real m, taking $m = t/(2D) - 1$ yields

$$\mathbb{P}(F_t) \le \sum_{p \in \mathcal{Q}} r_p^{-1} t^{\alpha/2} a_p^t, \qquad \{\mathcal{C}_t \not\subseteq \mathcal{R}\} \subseteq \{K_\infty > t/A\} \cup F_t. \tag{39}$$

For integer $A = 2(D+1)$, the number of integers $t \ge 2$ with $t < Ak$ is at most Ak. Tonelli's theorem therefore gives

$$\sum_{t=2}^{\infty} \mathbb{P}\{K_\infty > t/A\} \le A\mathbb{E}K_\infty.$$

Together with (39) and the two initial slots, this proves (31).

Finally, we bound the expectation of the last exterior time. Set $J = \sup(\{t \geq 2 : F_t\} \cup \{0\})$. Pathwise, every exterior slot is either strictly before AK_∞ or at most J, apart from the two initial slots. Hence $\tau_{\mathcal{R}} \leq 2 + AK_\infty + J$. Since

$$\mathbb{E}J \leq \sum_{t=2}^{\infty} t\mathbb{P}(F_t) \leq \sum_{p\in\mathcal{Q}} r_p^{-1}\Psi_\alpha(a_p),$$

(32) follows. The last reference-failure time also has finite expectation, as it is no later than the last exterior time. □

For the default $\alpha = 2$, the constants have the elementary forms

$$z_2 = \frac{\pi^2}{6} - 1, \qquad \Phi_2(a) = \frac{a}{(1-a)^2} - a, \qquad \Psi_2(a) = \frac{a(1+a)}{(1-a)^3} - a. \tag{40}$$

After the last reference failure, every complete interval has length at most D, apart from the one containing that time, because a longer interval would contain a reference failure by (36). Positive Fisher information at these finitely many ages then accumulates at least linearly.

C.4 A DECISION-RELEVANT COST BOUND

Suppose the true rate is useful, set $H_* = H(p_*)$, and define

$$L_{\min} = \begin{cases} H_* - 1, & W_2(H_* - 1; p_*) = c, \\ H_*, & W_2(H_* - 1; p_*) < c. \end{cases}$$

The optimal thresholds for the true rate are exactly the integers in $[L_{\min}, H_*]$, which contains at most two consecutive values. In the on-grid case, we choose the safe reference set

$$\mathcal{R} = \{p \in \mathcal{G} : H(p) \in [L_{\min}, H_*]\}. \tag{41}$$

In the off-grid case, we assume condition (21). We then set $H_* = L_{\min} = L$ and $\mathcal{R} = \{p \in \mathcal{G} : H(p) = L\}$. This set contains both anchors. More generally, the proof below allows ties at the endpoints whenever the tie-broken thresholds of the endpoints and of every chosen reference lie in the optimal threshold set of the true rate.

With these choices, use the constants of Theorem 7 and define the strictly premature candidates

$$\mathcal{Q}_- = \{p \in \mathcal{G} : H(p) < L_{\min}\} \subseteq \mathcal{Q}.$$

Theorem 8 (Decision-relevant bounded cost)**.** *Under either safe-reference construction above, put $\rho = \rho_*(p_*)$. For every $T \geq 1$,*

$$-c \leq R_T \leq (g_2 - \rho)Mz_\alpha + c\left[AMz_\alpha + \sum_{p\in\mathcal{Q}_-} r_p^{-1}\Phi_\alpha(a_p)\right] \quad \text{if } \mathcal{Q}_- \neq \varnothing. \tag{42}$$

If $\mathcal{Q}_- = \varnothing$, the sharper bound is

$$-c \leq R_T \leq (g_2 - \rho)Mz_\alpha. \tag{43}$$

In particular, $\sup_T R_T < \infty$ for every $\alpha > 1$, including the default $\alpha = 2$. Adding arbitrarily close grid points with an already optimal threshold changes neither M nor the affinity terms of (42)*.*

Proof. At age one every plausible target index is $g_2 p < g_2 \leq c$, so there are no consecutive target visits. Consequently, the visit price c equals the competitor's conditional expected loss, and

$$\mathbb{E}\sum_{t<T}\sum_i g_i Z_i(t^+) = \mathbb{E}\sum_{t<T} d(h_t, b_t), \qquad d(h,b) = bc + (1-b)g_2[1-(1-p_*)^h]. \tag{44}$$

Here b_t indicates a target refresh, $h_{t+1} = 1$ after a refresh and $h_{t+1} = h_t + 1$ otherwise, and $h_0 = 0$.

A bounded relative value for the single-source problem with visit price c and true rate p_* is

$$v(h) = \begin{cases} h\rho - g_2 S_{p_*}(h), & 0 \leq h \leq H_*, \\ c, & h \geq H_*. \end{cases}$$

The threshold inequalities imply $g_2[1-(1-p_*)^{H_*-1}] \le \rho \le g_2[1-(1-p_*)^{H_*}]$. Thus $v(0)=0$, $v(1)=\rho$, $0 \le v \le c$, and

$$\rho + v(h) = \min\{c + v(1),\ g_2[1-(1-p_*)^h] + v(h+1)\}.$$

The residual $\delta_t = d(h_t, b_t) + v(h_{t+1}) - \rho - v(h_t)$ is nonnegative. A premature refresh has residual $c - v(h_t) \le c$, and has positive residual only for $h_t < L_{\min}$. A skipped refresh has positive residual only for $h_t \ge H_*$, in which case it equals $g_2[1-(1-p_*)^{h_t}] - \rho \le g_2 - \rho$. In particular $\delta_t \le c$.

If a reference point is retained, the learner cannot skip at any age $h_t \ge H_*$, because every reference point has threshold at most H_*, and the competitor then has age one. If a strictly premature refresh occurs, some retained candidate must have $H(p) \le h_t < L_{\min}$, so it belongs to $\mathcal{Q}_-$. Consequently

$$\delta_t \le (g_2 - \rho)\mathbf{1}\{\mathcal{C}_t \cap \mathcal{R} = \varnothing\} + c\,\mathbf{1}\{\mathcal{C}_t \cap \mathcal{Q}_- \ne \varnothing\}. \tag{45}$$

The residuals at times zero and one are zero regardless of the likelihood set. In the tie case $W_2(H_*-1; p_*) = c$, a refresh at age H_*-1 is optimal and has zero residual.

Telescoping gives

$$R_T = \mathbb{E}\sum_{t<T} \delta_t - \mathbb{E}v(h_T) \ge -c. \tag{46}$$

Together with (45), it gives

$$R_T \le (g_2 - \rho)\mathbb{E}K_\infty + c\sum_{t=2}^{T-1} \mathbb{P}\{\mathcal{C}_t \cap \mathcal{Q}_- \ne \varnothing\}. \tag{47}$$

For nonempty $\mathcal{Q}_-$, repeat (39) with the union restricted to $\mathcal{Q}_-$, using the same global count and the same K_∞. Summing only over $t \ge 2$ gives

$$\sum_{t=2}^{\infty} \mathbb{P}\{\mathcal{C}_t \cap \mathcal{Q}_- \ne \varnothing\} \le AMz_\alpha + \sum_{p \in \mathcal{Q}_-} r_p^{-1}\Phi_\alpha(a_p).$$

Combine this with (45) to prove both displayed bounds. When $M = 0$, all actions are optimal and $-c \le R_T \le 0$. □

Corollary 3 (Other logarithmic cutoffs)**.** *The conclusions of Theorem 4 for a useful true rate and of Theorem 5 under* (21) *hold with $R_T = O(1)$ for every fixed cutoff coefficient $\alpha > 1$.*

Proof. Apply Theorems 7 and 8. The constants z_α, $\Phi_\alpha(a)$ and $\Psi_\alpha(a)$ are finite. □

C.5 A BRACKETING INTERVAL THAT CROSSES THE NO-VISIT BOUNDARY

The assumption that every reference point is useful fails when $p_L < p_* < p^{\mathrm{crit}} = g_2/c \le p_U$. In this case we bound the average cost directly.

Theorem 9 (Cost approximation across the no-visit boundary)**.** *Suppose $p_L < p_* < p^{\mathrm{crit}} \le p_U$ are as above, with adjacent grid points p_L, p_U. Put $L_0 = H(p_L) < \infty$, $\rho = \rho_*(p_*)$, and*

$$\epsilon_{\mathrm{cross}} = \max\{J_{L_0}(p_*) - \rho,\ g_2 - \rho\}. \tag{48}$$

The reopening learner satisfies

$$0 \le \liminf_{T\to\infty} \frac{R_T}{T} \le \limsup_{T\to\infty} \frac{R_T}{T} \le \epsilon_{\mathrm{cross}} \le g_2 \max\{\log(p_*/p_L), \log(p^{\mathrm{crit}}/p_*)\} \le g_2 \log(p_U/p_L). \tag{49}$$

Proof. Call a completed target interval short if its length is less than L_0, and let $N_{<L_0}(t)$ count these intervals strictly before t. Neither bracketing point can trigger such a visit, because p_L waits until age L_0 and the index under p_U never exceeds c. A short visit therefore requires an exterior candidate in $\mathcal{C}_t$. As before, each exterior point is anchored at its nearest bracketing endpoint. The anchored likelihood ratios and the affinity factors remain valid without assuming useful endpoints. These factors are below one at every positive finite age. They have a common maximum $r < 1$ over $1 \le h < L_0$ and the finitely many exterior candidates.

For every fixed $\varepsilon > 0$, the argument for (38) gives

$$\mathbb{P}\{\text{a short visit at } t,\ N_{<L_0}(t) \geq \varepsilon t\} \leq (|\mathcal{G}| - 2)t^{\alpha/2} r^{\varepsilon t}.$$

Without exterior candidates there are no short visits. The bound is summable in t. By the Borel–Cantelli lemma, after a finite random time no short visit occurs while the number of earlier short intervals exceeds εt. At an increment, the ratio $N_{<L_0}(t)/t$ can exceed ε by at most $1/t$. Between increments, it decreases. Applying this to rational $\varepsilon \downarrow 0$ proves $N_{<L_0}(t)/t \to 0$ almost surely.

For the true rate,

$$J_{L+1}(p_*) - J_L(p_*) = \frac{W_2(L; p_*) - c}{L(L+1)}.$$

The index strictly increases in age (Theorem 2), so J_L first decreases and then increases towards g_2. Consequently $\sup_{L \geq L_0} J_L(p_*) = \max\{J_{L_0}(p_*), g_2\}$. Each completed long cycle has average age-based cost at most $\rho + \epsilon_{\text{cross}}$. The total duration of short cycles is at most $(L_0 - 1)N_{<L_0}(T) = o(T)$. A possibly long unfinished cycle has per-slot cost at most $g_2 \leq \rho + \epsilon_{\text{cross}}$, apart from a single refresh boundary of cost at most c. These statements also cover the case of finitely many visits. They prove the pathwise upper bound for the age-based average cost. These averages are bounded by c, so the reverse Fatou inequality and the cost identity (44) transfer the bound to R_T/T. The bounded relative value at the useful true rate gives $R_T \geq -c$.

Finally, every J_L and their infimum ρ_* are nondecreasing in p and Lipschitz in $\log p$ with constant g_2. Indeed, $0 \leq J_L'(p) = (g_2/L)\sum_{h=1}^{L-1} h(1-p)^{h-1} \leq g_2/p$. Since L_0 is optimal at p_L, $J_{L_0}(p_*) - \rho_*(p_*) \leq g_2 \log(p_*/p_L)$. Since $\rho_*(p^{\text{crit}}) = g_2$, $g_2 - \rho_*(p_*) \leq g_2 \log(p^{\text{crit}}/p_*)$. This proves the remaining inequalities. □

C.6 The on-grid no-visit and critical cases

It remains to prove Theorem 4 when $p_* \in \mathcal{G}$ and $c \geq g_2/p_*$. Proposition 2 then gives $\rho_*(p_*) = g_2$. The threshold H_* is infinite, including at equality, so we bound the visit frequency instead.

Let $N_2(t)$ count source 2 visits strictly before t, and let

$$\{p \in \mathcal{G} : g_2/p > c\}$$

be the useful grid points. A source 2 visit at a positive time requires at least one such point in $\mathcal{C}_t$, because every other point has finite-age index strictly below c, whereas source 1 has index at least c. Each useful point differs from p_*. If there are no useful grid points, source 2 is never visited and the claim is immediate.

Fix $\varepsilon > 0$ and set $D_\varepsilon = \lceil 2/\varepsilon \rceil$. The sum of all completed inter-visit intervals before t is at most t. If $N_2(t) \geq \varepsilon t$, at least $\varepsilon t/2 - 1$ of these observations have ages at most D_ε. The affinities of the useful candidates with p_* have a common maximum $r_\varepsilon < 1$ over this finite age set. If the algorithm visits source 2 at $t \geq 2$, a useful candidate lies in $\mathcal{C}_t$, and its likelihood ratio to p_* is at least $t^{-\alpha}$. Applying (38), now with D_ε, and taking a union bound gives

$$\mathbb{P}\{\text{source 2 visited at } t,\ N_2(t) \geq \varepsilon t\} \leq (|\mathcal{G}| - 1)t^{\alpha/2} r_\varepsilon^{\varepsilon t/2 - 1}.$$

This bound is summable in t. By the Borel–Cantelli lemma, almost surely only finitely many source 2 visits occur at times t with $N_2(t) \geq \varepsilon t$. The count $N_2(t)$ does not change between visits and increases by one at each visit. Taking a countable sequence $\varepsilon \downarrow 0$ therefore yields $N_2(t)/t \to 0$ almost surely.

Since $0 \leq N_2(T)/T \leq 1$, dominated convergence gives $\mathbb{E}N_2(T)/T \to 0$. For any action trajectory, the deterministic age-based freshness sum satisfies

$$\sum_{t<T} g_2(1-p_*)^{h_2(t^+)} \leq \frac{g_2}{p_*}[N_2(T) + 1],$$

because each segment beginning with a fresh copy contributes at most $\sum_{h \geq 0} g_2(1-p_*)^h = g_2/p_*$. We take expectations and apply the conditional staleness formula together with (44). This gives the following bounds on the expected realized cost

$$g_2 + \left(c - \frac{g_2}{p_*}\right)\frac{\mathbb{E}N_2(T)}{T} - \frac{g_2}{p_* T} \leq \frac{1}{T}\mathbb{E}\sum_{t<T}\sum_{i=1}^{2} g_i Z_i(t^+) \leq g_2 + c\frac{\mathbb{E}N_2(T)}{T}.$$

Both sides converge to g_2, proving $R_T/T \to 0$.

C.7 Proof of Corollary 1

The frozen rule updates its likelihood set immediately after each source 2 visit, with β_t evaluated at the time of that visit. It then holds the set fixed until the next visit. Write $p^\dagger = g_2/c$, $u = \max \mathcal{G}$, and $\mathcal{G}_- = \{p \in \mathcal{G} : p < p^\dagger\}$. By hypothesis, $\mathcal{G}_-$ contains p_* and $u \geq p^\dagger$. Let $v = \max \mathcal{G}_-$ and $K = \max_{p \in \mathcal{G}_-} H(p) < \infty$.

Consider the observation path on which every visit to source 2 reports a change. On this path u always maximizes the likelihood. Before $\mathcal{G}_-$ is entirely excluded, a retained useful candidate forces the next visit within K slots. Thus each visit up to and including the first complete exclusion, if it occurs, has time $s_n \leq nK$. Each observation increases the log-likelihood advantage of u over v by at least

$$\delta = \min_{1 \leq h \leq K} \log \frac{1-(1-u)^h}{1-(1-v)^h} > 0.$$

Since the likelihood of a changed observation increases with the rate, after n visits the advantage of u over every point of $\mathcal{G}_-$ is at least $n\delta$. Choose a deterministic finite n_0 with $n_0\delta > \alpha \log(\max\{n_0 K, 2\})$. Such an n_0 exists because $\delta > 0$ and α is fixed. If complete exclusion has not occurred before the n_0th visit, that visit exists, occurs by $n_0 K$, and excludes every useful point. Hence complete exclusion occurs at a finite time s on this path. Let E be the event that every visit up to time s reports a change. On this path, each visit age is determined by the preceding observations. Independence of changes after each reset therefore gives $\mathbb{P}(E) = \prod_k [1-(1-p_*)^{h_k}] > 0$.

On E, every retained rate is at least $p^\dagger$, so its finite-age index is strictly below $g_2/p^\dagger = c$. Source 1 has priority at least c after initialization, so source 2 is never visited again. Moreover, $c < g_2/p_*$, so Proposition 2 gives $\rho_*(p_*) < g_2$.

The frozen rule, like the reopening rule, never visits source 2 consecutively. It therefore satisfies the cost identity (44). For every trajectory of ages and actions and every prefix length n, the single-source average-cost inequality also gives

$$\sum_{t<n} d(h_2(t), b_t) - n\rho_*(p_*) \geq -c.$$

This bound concerns the deterministic age-based cost functional.

On E, after the visit at s, source 2 remains unvisited. For $T \geq s+1$, its subsequent age-based cost is $g_2 \sum_{k=1}^{T-s-1} [1-(1-p_*)^k]$. Applying the prefix bound at $s+1$ and bounding the remaining freshness sum by $1/p_*$ shows that the total age-based excess on E is at least

$$(g_2 - \rho_*(p_*))(T-s-1) - g_2/p_* - c.$$

On E^c it is at least $-c$. Splitting the expectation of this age-based functional over E and E^c and then applying (44) yields

$$R_T^{\text{frozen}} \geq \mathbb{P}(E)(g_2 - \rho_*(p_*))(T-s-1) - \mathbb{P}(E) g_2/p_* - c. \tag{50}$$

Its linear coefficient is strictly positive. On the same instance, Theorem 4 gives $R_T = O(1)$ for the reopening rule.

C.8 Off-grid approximation when thresholds differ

Suppose $p_L < p_* < p_U$ are adjacent grid points and $c < g_2/p_U$. Both reference rates are then useful. Theorem 7 establishes eventually bounded waits and $\mathcal{C}_t \subseteq \{p_L, p_U\}$ almost surely. This proves the localization part of Theorem 5.

After localization, every complete inter-visit cycle of source 2 has a length L in the range

$$\min\{H(p_L), H(p_U)\} \leq L \leq \max\{H(p_L), H(p_U)\}. \tag{51}$$

Before the earlier of these thresholds, neither candidate can outrank source 1. By the later threshold, either retained candidate does. Because the set can expand between visits as β_t increases, L can differ from both endpoint thresholds while remaining in the range (51).

Every integer L in this range is optimal for some $\bar{p} \in [p_L, p_U]$. To see this, choose endpoints p_a, p_b with $H(p_a) \le L \le H(p_b)$. Then $W_2(L; p_a) > c$ and $W_2(L-1; p_b) \le c$. If either endpoint already satisfies

$$W_2(L-1; p) \le c \le W_2(L; p),$$

it provides $\bar{p}$. Otherwise $W_2(L-1; p_a) > c$ and $W_2(L; p_b) < c$. Continuity along the interval between these endpoints gives a point with $W_2(L; \bar{p}) = c$, and strict increase of the index in age (Theorem 2) gives $W_2(L-1; \bar{p}) < c$. The threshold characterization then shows that L is optimal for $\bar{p}$. All these thresholds are finite since $c < g_2/p$ throughout the bracketing interval.

For a finite $L \ge 2$, differentiation of (13) gives

$$0 \le \frac{\partial J_L(p)}{\partial p} = \frac{g_2}{L} \sum_{h=1}^{L-1} h(1-p)^{h-1} \le \frac{g_2}{p}. \tag{52}$$

The last inequality uses $hp(1-p)^{h-1} \le 1$, the probability of exactly one success among h Bernoulli trials. The infinite-threshold cost is constant. Integrating (52) shows that every J_L, and hence their infimum ρ_*, satisfies

$$|J_L(p) - J_L(p')| \le g_2 |\log(p/p')|, \qquad |\rho_*(p) - \rho_*(p')| \le g_2 |\log(p/p')|.$$

The functions J_L and ρ_* are also nondecreasing in p. Let $\bar{p}$ be a rate for which a realized cycle length L is optimal. If $\bar{p} \le p_*$, monotonicity of ρ_* yields

$$J_L(p_*) - \rho_*(p_*) \le J_L(p_*) - J_L(\bar{p}) \le g_2 \log(p_*/\bar{p}).$$

If $\bar{p} \ge p_*$, monotonicity of J_L instead gives

$$J_L(p_*) - \rho_*(p_*) \le \rho_*(\bar{p}) - \rho_*(p_*) \le g_2 \log(\bar{p}/p_*).$$

Both bounds are at most $g_2 \log(p_U/p_L)$. After a finite random prefix, the age-based average cost is a duration-weighted average of these cycle costs plus an incomplete final cycle of uniformly bounded length. Its pathwise limit superior is therefore at most $\rho_*(p_*) + g_2 \log(p_U/p_L)$. The age-based averages are bounded, so the reverse Fatou inequality transfers this upper bound to the limit superior of their expectations. Equation (44) transfers it to expected realized costs. Proposition 2 supplies the lower bound, proving (20).

If (21) holds, all retained reference points and the true rate have the same optimal threshold. Theorem 8, with $D = L$, proves the bounded-excess assertion of Theorem 5. If instead the bracket crosses g_2/c, Theorem 9 proves the same average-cost bound. Together these arguments cover every useful true rate in Theorem 5.

D Proof of the capacity condition

We prove Proposition 5 for a fixed target i. The proof holds for any cutoff $\beta_t = \alpha \log(\max\{t, 2\})$ with fixed $\alpha > 1$. Competitors have unknown rates, may violate their own capacity conditions, and may use arbitrary evolving subsets of their full rate grid in their optimistic indices. All decisions follow the top-B positive-index rule.

A reference rate eventually remains available. We use the filtration of all received observations and decisions strictly before time t. Conditional on this history, a target visit of age h yields a Bernoulli observation with mean $\mu_{p_i}(h) = 1 - (1-p_i)^h$. This conditional law holds under any adaptive schedule because changes after the most recent reset are independent of the history. Write $P_p(\cdot \mid h)$ for the Bernoulli law with mean $\mu_p(h)$ and $\ell_{i,t}(p)$ for the exact target log-likelihood accumulated strictly before t.

For an on-grid target, compare each $p \notin \mathcal{B}_i$ with the anchor $b(p) = p_i$. For an off-grid target, use $b(p) = p_{L,i}$ if $p < p_{L,i}$ and $b(p) = p_{U,i}$ if $p > p_{U,i}$. In the on-grid case,

$$\exp\{\ell_{i,t}(p) - \ell_{i,t}(b(p))\}$$

is a nonnegative martingale of initial value one. In the off-grid case it is a nonnegative supermartingale. With $\theta = \mu_{p_i}(h)$, $\theta_b = \mu_{b(p)}(h)$, and $\theta_p = \mu_p(h)$, its one-observation increment has conditional expectation

$$1 + \frac{(\theta - \theta_b)(\theta_p - \theta_b)}{\theta_b(1-\theta_b)} < 1,$$

because the exterior point and the true rate lie on opposite sides of the anchor.

If $\mathcal{C}_i(t) \cap \mathcal{B}_i = \varnothing$, a maximum-likelihood point is exterior and exceeds its anchor in log-likelihood by more than β_t. Markov's inequality and a union bound therefore give, for $t \geq 2$,

$$\mathbb{P}\{\mathcal{C}_i(t) \cap \mathcal{B}_i = \varnothing\} \leq (|\mathcal{G}| - |\mathcal{B}_i|)t^{-\alpha}. \tag{53}$$

Since $\sum_{t=2}^{\infty} t^{-\alpha} < \infty$ for $\alpha > 1$, Borel–Cantelli implies that after a finite random time, at least one reference point is retained at every decision. This reference point may vary over time.

Counting high-priority competitor visits. Abbreviate $d_j = d_j(c_i)$, $\Lambda_i = \Lambda_i(c_i)$, and put

$$M_i = \#\{j \neq i : d_j < \infty\}.$$

Consider any L consecutive decision slots beginning at t_0. Suppose that the target's age is at least H_i at t_0, that a reference point is retained at every slot, and that the target is never visited in the window. The target's priority is then strictly greater than c_i throughout. Consequently, under any tie rule, every slot contains B competitor visits with priorities at least c_i.

The optimistic index of competitor j is bounded by U_j regardless of its current likelihood set. A visit scoring at least c_i therefore requires age at least d_j. Between two such visits its elapsed time is at least d_j, because after the first visit the age resets to zero, and any intervening lower-priority refresh only makes the age at the second visit smaller. For finite d_j, at most $1 + \lfloor (L-1)/d_j \rfloor$ high-priority visits fit into L decision slots. The leading one accounts for an arbitrary competitor age at the start of the window. A competitor with $d_j = \infty$ supplies no such visits. Hence

$$BL \leq \sum_{j \neq i:\, d_j < \infty} \left[1 + \left\lfloor \frac{L-1}{d_j} \right\rfloor \right] \leq M_i + (L-1)\Lambda_i. \tag{54}$$

Since $\Lambda_i < B$, the number of consecutive eligible slots in which the target can be omitted is at most

$$R_i = \left\lfloor \frac{M_i - \Lambda_i}{B - \Lambda_i} \right\rfloor. \tag{55}$$

Here $0 \leq \Lambda_i \leq M_i$, so R_i is nonnegative. In particular, the target must be visited by slot $t_0 + R_i$, because omitting it also in that slot would create $R_i + 1$ omitted slots and contradict (54).

Following any visit at s after reference retention has become permanent, the target is either visited before age H_i or reaches that age at $s + H_i$ and is visited by $s + H_i + R_i$. Thus a sharper eventual inter-visit bound is $H_i + R_i$. Since $M_i \leq N - 1$ and $M_i - \Lambda_i \leq N - 1$, the sharper bound implies (24). These eventual bounds apply to intervals that begin after the random last time at which no reference point is retained. A counting window started after that time also guarantees a first target visit.

Bounded intervals remove exterior grid points. For $p \notin \mathcal{B}_i$, set

$$a_h(p) = \mathbb{E}_{p_i} \sqrt{\frac{P_p(Y \mid h)}{P_{b(p)}(Y \mid h)}}.$$

For an on-grid target this is a strict Bernoulli affinity. In the off-grid case, the strict supermartingale increment bound and Cauchy–Schwarz give the same conclusion. Hence $0 < a_h(p) < 1$ for every finite positive h. With any deterministic eventual bound D_i from (24), the finite maximum

$$r_i = \max_{p \in \mathcal{G} \setminus \mathcal{B}_i} \max_{1 \leq h \leq D_i} a_h(p)$$

is strictly below one when exterior points exist. The square root of the anchored likelihood ratio, divided by the product of its factors $a_h(p)$ over target visits, is a mean-one nonnegative martingale. If at least m observations before t have ages at most D_i, Markov's inequality gives

$$\mathbb{P}\left\{ e^{\ell_{i,t}(p) - \ell_{i,t}(b(p))} \geq t^{-\alpha},\ \#\{\text{target observations before } t \text{ with age } \leq D_i\} \geq m \right\} \leq t^{\alpha/2} r_i^m.$$

For a fixed integer s_0, consider the event that a reference point is retained at every time after s_0. For sufficiently large $t \geq 2s_0$, the same deterministic capacity argument in the recent window $[\lceil t/2 \rceil, t)$ gives at least

$$\max\left\{0, \left\lfloor \frac{t - \lceil t/2 \rceil}{D_i} \right\rfloor - 2\right\}$$

bounded-age target observations. The first observation crossing the window boundary is discarded. An exterior point in $\mathcal{C}_i(t)$ has anchored likelihood ratio at least $t^{-\alpha}$, because its anchor's likelihood is at most the maximum. Thus the probability that some exterior point belongs to $\mathcal{C}_i(t)$ and the fixed-s_0 retention event holds is bounded by a constant times $t^{\alpha/2} r_i^{\lfloor t/(2D_i)\rfloor - 3}$, which is summable. By Borel–Cantelli, these joint events occur only finitely often. Taking the countable union over s_0 and using (53) proves that eventually $\mathcal{C}_i(t) \subseteq \mathcal{B}_i$. Likelihood sets are nonempty because they contain a maximizer. If there are no exterior points, this conclusion is immediate.

Because the argument uses only the target's reference retention and the competitors' full-grid score bounds, its conclusions hold almost surely for all certified sources simultaneously. If every source is certified and on-grid, the sets eventually equal the true singletons, so the learner selects the same sources as known-rate WIRS at the learner's current ages, with the same tie rule.

E Finite expected localization time and a multi-source cost certificate

Each certified source has a finite expected localization time. When every source is on-grid and certified, a verifiable cycle condition on known-rate WIRS turns this into a cost guarantee. Throughout this section, ties are broken by a fixed source-ID ordering. We retain independent changes, fresh initialization, exact likelihoods, and decision-time reopening with $\beta_t = \alpha \log(\max\{t, 2\})$, $\alpha > 1$.

E.1 Charging long intervals to missing references

For a target i, let $\mathcal{B}_i$ be its true grid singleton or its adjacent bracketing pair, as in Proposition 5. Choose c_i, H_i as in (23), and put

$$d_{ij} = \min\{h \geq 1 : U_j(h) \geq c_i\}, \qquad \Lambda_i = \sum_{j \neq i} d_{ij}^{-1} < B,$$
$$M_i = \#\{j \neq i : d_{ij} < \infty\}, \qquad \eta_i = B - \Lambda_i.$$

The convention $1/\infty = 0$ applies. Define

$$\kappa_i = B/\eta_i, \qquad A_i = H_i + M_i/\eta_i, \qquad \widehat{D}_i = \lceil 2A_i \rceil, \qquad m_i = |\mathcal{G} \setminus \mathcal{B}_i|, \qquad z_\alpha = \sum_{t=2}^{\infty} t^{-\alpha}.$$

Let $K_i(t)$ count decisions $s < t$ at which no reference point is retained. Let $N_i(t)$ count observations strictly before t whose visit age is at most $\widehat{D}_i$.

Lemma 1 (Global information count)**.** *On every history,*

$$t \leq \widehat{D}_i N_i(t) + 2\kappa_i K_i(t) + A_i. \tag{56}$$

Proof. In a complete target interval of length $h > H_i$, consider the $L = h - H_i$ skipped slots at ages $H_i, \ldots, h-1$. If no reference point is retained at k of them, the remaining $L - k$ slots require B competitor visits with score greater than c_i. Counting all high-score competitor visits in the entire window, including those in slots without a reference point, gives

$$B(L-k) \leq \sum_{j : d_{ij} < \infty} \left[1 + \left\lfloor \frac{L-1}{d_{ij}} \right\rfloor\right] \leq M_i + (L-1)\Lambda_i.$$

The leading one permits arbitrary initial competitor ages. Intermediate low-score visits cannot shorten the separation d_{ij} of high-score visits. Thus $h \leq A_i + \kappa_i k$. The same bound applies to the unfinished age. Intervals with $h \leq H_i$ require no window count.

We treat the fresh start at time zero as a reset. Completed visit ages plus the unfinished age sum to t exactly. A completed interval with $h > \widehat{D}_i \geq 2A_i$ satisfies $h < 2\kappa_i k$. All shorter intervals contribute at most $\widehat{D}_i N_i(t)$. The unfinished interval contributes at most $A_i + \kappa_i k$. The slots without a reference point counted for different intervals are disjoint, which proves the result. □

For $p \notin \mathcal{B}_i$, use the same anchor $b_i(p)$ as in the proof of the capacity condition. The anchor is the true rate for an on-grid target and the adjacent endpoint between p and the true rate for an off-grid target. Write

$$a_{i,h}(p) = \mathbb{E}_{p_i}\sqrt{P_p(Y \mid h)/P_{b_i(p)}(Y \mid h)}, \quad r_i = \max_{p \notin \mathcal{B}_i} \max_{1 \le h \le \widehat{D}_i} a_{i,h}(p) < 1, \quad a_i = r_i^{1/(2\widehat{D}_i)}.$$

When $m_i = 0$, the conclusions below are immediate. Terms multiplied by m_i are then interpreted as zero, and r_i and a_i need not be defined.

Theorem 10 (Finite expected localization under capacity)**.** *For a certified source i, define $\tau_i = \inf\{n : \mathcal{C}_i(t) \subseteq \mathcal{B}_i$ for all $t \ge n\}$. Then*

$$\mathbb{E}\sum_{t \ge 0} \mathbf{1}\{\mathcal{C}_i(t) \not\subseteq \mathcal{B}_i\} \le 2 + 4\kappa_i m_i z_\alpha + m_i r_i^{-A_i/\widehat{D}_i} \sum_{t=2}^{\infty} t^{\alpha/2} a_i^t < \infty, \tag{57}$$

$$\mathbb{E}\tau_i \le 2 + 4\kappa_i m_i z_\alpha + m_i r_i^{-A_i/\widehat{D}_i} \sum_{t=2}^{\infty} t^{1+\alpha/2} a_i^t < \infty. \tag{58}$$

If every source is on-grid and certified, the expected total number of actions different from known-rate WIRS at the learner's current state is finite.

Proof. The anchored likelihood-ratio bound (53) in Appendix D gives $\mathbb{P}\{\mathcal{C}_i(t) \cap \mathcal{B}_i = \varnothing\} \le m_i t^{-\alpha}$. The first two slots retain the full grid, so $\mathbb{E}K_i(\infty) \le m_i z_\alpha$. The square-root anchored likelihood ratio divided by the product of its predictable factors $a_{i,h}(p)$ is a mean-one nonnegative martingale, also in the off-grid case. At a retained exterior point the ratio is at least $t^{-\alpha}$. Consequently, for every real u,

$$\mathbb{P}\{p \in \mathcal{C}_i(t), N_i(t) \ge u\} \le t^{\alpha/2} r_i^u.$$

By (56), $K_i(\infty) \le t/(4\kappa_i)$ implies $N_i(t) \ge t/(2\widehat{D}_i) - A_i/\widehat{D}_i$. Let $F_{i,t}$ be the event that some exterior point belongs to $\mathcal{C}_i(t)$ and $N_i(t)$ satisfies this lower bound. Then

$$\{\mathcal{C}_i(t) \not\subseteq \mathcal{B}_i\} \subseteq \{t < 4\kappa_i K_i(\infty)\} \cup F_{i,t}, \qquad \mathbb{P}(F_{i,t}) \le m_i r_i^{-A_i/\widehat{D}_i} t^{\alpha/2} a_i^t.$$

The number of integers $t \ge 2$ satisfying the first event is at most $4\kappa_i K_i(\infty)$. Summing proves (57). To bound τ_i, put $J_i = \sup(\{t \ge 2 : F_{i,t}\} \cup \{0\})$. Pathwise $\tau_i \le 2 + 4\kappa_i K_i(\infty) + J_i$, and $\mathbb{E}J_i \le \sum_{t \ge 2} t\mathbb{P}(F_{i,t})$, proving (58). For an on-grid source, a nonempty set without exterior points is the true singleton. When this holds for every source, the fixed tie rule makes the learner's action equal to the action of known-rate WIRS at the same ages. A union bound over sources completes the proof. □

E.2 A BOUNDED POLICY BIAS UNDER A COMMON CYCLE MEAN

Here all true rates are on-grid and all sources are certified. This restriction ensures that each true competing index is bounded by its full-grid envelope. Set

$$\bar{D}_i = H_i + \left\lfloor \frac{M_i - \Lambda_i}{B - \Lambda_i} \right\rfloor, \qquad L = 1 + \max_i \bar{D}_i, \qquad Q = \prod_i \{1, \dots, \bar{D}_i\}.$$

Let F be the deterministic known-rate WIRS state map, with visited sources having age one at the next decision. From any initial ages, source i is visited within $\max\{0, H_i - h_i\} + \bar{D}_i - H_i \le \bar{D}_i$ decision offsets, and all its subsequent intervals are at most $\bar{D}_i$. Thus $F^t(x) \in Q$ for every initial state x and every $t \ge L$. The added one in L accounts for the decision just after the latest possible first visit.

Transient paths from Q can leave it, so we use the finite closed set

$$X = \bigcup_{r=0}^{L-1} F^r(Q), \qquad F(X) \subseteq X. \tag{59}$$

Closure follows from $F^L(Q) \subseteq Q$. Every recurrent cycle lies entirely in Q.

Write $\Gamma = \sum_i g_i$ and $\ell(x, a) = \sum_{i:a_i=0} g_i[1 - (1 - p_i)^{x_i}] \in [0, \Gamma]$.

Theorem 11 (Multi-source cost under a verifiable cycle condition)**.** *Suppose every recurrent cycle of F has the same average stage cost ρ_W. Then there is a bounded function v on all age states satisfying*

$$\rho_W + v(x) = \ell(x, \pi_W(x)) + v(F(x)).$$

If E_T is the learner's number of actions differing from π_W at its own state and $V = \sup v - \inf v$, then

$$|\mathbb{E}C_T - T\rho_W| \le (\Gamma + V)\mathbb{E}E_T + V = O(1). \tag{60}$$

The same bound applies to the absolute expected cost difference from an oracle run starting at the same initial state. If additionally ρ_W attains a Lagrangian lower bound, it is the global hard-budget optimal average cost.

Proof. On a finite deterministic graph, the condition that every cycle has the same average stage cost is necessary and sufficient for a scalar policy Poisson equation. Necessity follows by summing the equation around each cycle. For sufficiency, choose one bias value in each cycle and assign subsequent values by $v(Fx) = v(x)+\rho_W-\ell(x, \pi_W(x))$. Because each recurrent cycle has average stage cost ρ_W, this assignment returns to the chosen value after one pass around the cycle. Extend backward along each transient path. This gives v_X on X.

For arbitrary x, including learner states with arbitrarily large ages, define

$$v(x) = \sum_{r=0}^{L-1} [\ell(F^r(x), \pi_W(F^r(x))) - \rho_W] + v_X(F^L(x)).$$

This agrees with v_X on X. Comparing consecutive states and using the Poisson equation at $F^L(x) \in X$ proves the one-step equation everywhere. Moreover $\min_X v_X - L\rho_W \le v(x) \le \max_X v_X + L(\Gamma - \rho_W)$, so

$$V \le \operatorname{span}(v_X) + L\Gamma < \infty. \tag{61}$$

The residual $\delta(x, a) = \ell(x, a)-\rho_W+v(F(x, a))-v(x)$ is zero for the WIRS action and has absolute value at most $\Gamma + V$ otherwise. Telescoping proves (60). Theorem 10 bounds $\mathbb{E}E_T$ uniformly. For a separately run oracle, initial bias values cancel and terminal bias values differ by at most V. The argument applies pathwise to the age-based cost. The conditional observation model and the tower property then identify its expectation with the expected realized staleness cost.

For the last assertion, if a common price $\nu \ge 0$ makes a recurrent cycle attain $\sum_i f_i^*(\nu) - \nu B$, the hard-budget lower bound proves optimality of ρ_W. The single-source bounded biases also imply that every feasible finite-horizon policy costs at least $T\rho_W - N\nu$. Thus the learner has bounded excess over the finite-horizon optimum as well. □

With two recurrent cycles of different means, an early learning error can move the learner to a cycle whose average cost differs from that of a separately initialized oracle. The cycle condition excludes this case.

F Noisy refresh as a belief Markov decision process

If retrieval can return outdated content, a refresh no longer resets staleness to zero with certainty. Age alone then does not suffice for scheduling. In a binary-state model with a specified observation channel, one can instead track a belief $\omega \in [0, 1]$. Waiting updates the belief as $\omega \mapsto \omega + (1 - \omega)p$. A refresh requires a Bayesian update determined by the returned observation and the retrieval-error model. Partially observed restless bandits provide a framework for such states (Liu & Zhao, 2010; Meshram et al., 2018; Chen et al., 2022). Threshold structure and indexability then require explicit conditions on that channel.

G Relation to earlier refresh learners

The closest methods differ in what they can observe and in their budget constraints. The BLLCε algorithm of Kolobov et al. (2019, Supplement Section 9.3) reserves at least $\varepsilon R/|W|$ crawl rate

Table 1: Cumulative excess over known-rate WIRS, in thousands. Known-rate WIRS costs 12.56, 5.23, and 2.42 per slot at $B = 1$, 3, and 6. Means and 95% intervals across 12 instances.

Policy	$B = 1$	$B = 3$	$B = 6$
UCB-WIRS	7.19 ± 1.68	1.81 ± 0.22	0.76 ± 0.20
Rate optimism	286.31 ± 28.91	61.73 ± 8.66	1.14 ± 0.16
MLE + initial pass	20.10 ± 6.83	1.96 ± 0.26	1.07 ± 0.15
MLE + decreasing exploration	59.05 ± 10.15	2.65 ± 0.63	0.95 ± 0.17
MWA	22.04 ± 2.09	34.67 ± 2.72	28.43 ± 3.04

per source before optimizing the remaining allocation. Its learning version uses this floor to keep gathering observations. In our slot model, a fixed fraction of cyclic, budget-charged visits plays a similar exploration role. Kolobov et al. (2020) also permit probes that produce cost observations without resetting the source. Our feedback is instead the version-change bit returned by a reset. Our cost and information guarantees hold under this feedback and a hard slot budget.

H Experimental details

The unknown-rate studies evaluate log-likelihoods in closed form with numerically stable expressions. The main learning study uses 12 instances. The known-rate budget sweep and model-error tests each use one instance with 20 trajectory seeds. Scaling uses eight independently drawn populations per size. Unless stated otherwise, costs are realized stochastic losses. We average four paired traces within each instance and compute 95% Student-t intervals over independent instances. Code and saved numerical results accompany the experiment directories.

Computing the lower bound. We evaluate (11) by golden-section search over ν. For each source and price, we minimize (8) over $\tau \leq 4000$ and include the never-refresh value g_i. The truncation error is negligible for the rates and weights used here.

H.1 Learning comparisons

The main learning study uses instance seeds 6200–6211. In instance $i = 0, \ldots, 11$ and replicate $s = 0, \ldots, 3$, the uniform draws that generate source changes come from NumPy's `default_rng` with seed $720000 + 100i + s$. All policies share the same $T \times N$ array for each instance and replicate. Every grid point retains its cumulative likelihood even when excluded. Fresh observations contribute $h \log(1 - p)$. Changed observations contribute $\log[-\operatorname{expm1}(h \log(1 - p))]$. Ties in the MLE go to the smallest grid point, and ties in priority go to the smallest source ID. All copies start fresh, and the learners start without observations. No source is visited at the initial age of zero.

Exploration. In these comparisons, B divides N. MLE + initial pass uses cyclic batches for slots $1, \ldots, N/B$ and then uses the MLE for every decision. MLE + decreasing exploration keeps an accumulator that starts at zero and adds $e_t = \min\{1, \sqrt{N/(Bt)}\}$ at each $t \geq 1$. Whenever the accumulator reaches one, the policy schedules a cyclic batch and subtracts one. Its first N/B slots form the same initial pass. Exploration visits count against the budget and return the same binary feedback as other visits. Table 1 reports the policies of Figure 3.

H.2 Known-rate experiments

Known-rate instances. The budget sweep (Figure 2a) and the model-error tests (Figures 4 and 5) each draw one base instance with $N = 50$ and vary the trajectory seeds. All policies use the same seeds. These figures show means and bands of one standard deviation across 20 trajectories. Horizons are $T = 4 \times 10^4$, with the first 4×10^3 slots excluded from steady-state averages. The scaling experiment (Figure 2b) draws eight independent populations at each $N \in \{10, 50, 100, 200, 400\}$ with $B = N/10$ and runs four trajectory seeds per population. It reports means and 95% intervals across populations. Sampling variation can put a finite-horizon estimate slightly below the bound.

In the budget sweep, round robin and random incur 7–32% and 25–111% excess cost over the bound, respectively.

Misspecification. In the burstiness test (Figure 5), each source has a hidden regime $M_i(t) \in \{\text{lo}, \text{hi}\}$ that starts uniformly at random and flips with probability $s = 0.02$ per slot, so each regime lasts 50 slots on average. In the two regimes, the change probability of source i is $p_i^{\text{lo}} = 2p_i/(1+\varphi)$ and $p_i^{\text{hi}} = 2p_i\varphi/(1+\varphi)$, which preserves the stationary mean p_i. We test $\varphi \in \{1, 3, 10, 30\}$, and $\varphi = 1$ recovers Assumption 1. Model-based policies use the constant rate p_i. WIRS keeps the lowest cost at every tested φ. As φ grows, the excess of MWA over WIRS falls from 12% to 6%, while myopic stays about 33% above WIRS. Figure 5 reports absolute costs because the lower bound (11) requires Assumption 1.

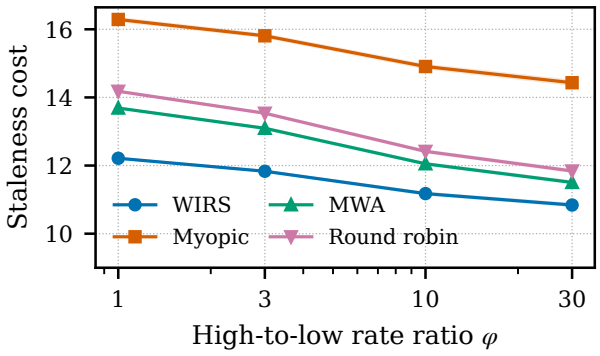


Figure 5: Staleness cost vs. the high-to-low rate ratio φ of a hidden two-regime change process with fixed stationary mean rates. Model-based policies use the mean rates.

In the rate-error test (Figure 4), dynamics follow Assumption 1. The estimated rates used by WIRS and myopic are $\hat{p}_i = \text{clip}(p_i e^{\sigma\xi_i}, 10^{-4}, 0.5)$. We draw $\xi_i \sim \mathcal{N}(0,1)$ per seed and test $\sigma \in \{0, 0.5, 1, 1.5, 2\}$. Before clipping, $\sigma = 1$ corresponds to a median multiplicative error $\max\{\hat{p}_i/p_i, p_i/\hat{p}_i\}$ of about two. MWA and round robin do not use rate estimates and appear as horizontal lines. The bound uses the true rates. Policies share trajectory seeds, and each rate perturbation stays fixed over its trajectory.

I LLM AGENT TASK

Environment. The task reuses the first 10^4 slots of the replicate-0 change trajectories of the 12 main learning instances, which have seeds 6200–6211 and $N = 30$. Each source is a supplier record with a unit price and a stock level. Supplier i has a base price drawn uniformly from the integers 5–30. Each change draws a new record whose price is the base plus a uniform integer 0–10 and whose stock is a uniform integer 0–20. A new record coincides with the old one with probability $1/231$. All copies start fresh, and the learners start without observations. Policies use the grid, cutoff, and tie rules of Appendix H.1. Myopic uses the true rates. The full-refresh reference reads all 30 records in every slot and ignores the budget.

Questions. Questions arrive at slots $25, 75, \ldots, 9975$, giving 200 per instance. Each asks about supplier i with probability $g_i/\sum_j g_j$ and requests a quantity k drawn uniformly from 1–8. The correct answer is k times the current unit price if the current stock is at least k, and `OUT OF STOCK` otherwise. Because suppliers are sampled in proportion to their weights, the probability that a question concerns a stale copy equals the normalized staleness cost $\sum_i g_i Z_i(t^+)/\sum_j g_j$. In the scheduling model, every copy is used in every slot, and the questions sample this use. All policies see the same questions and change trajectories.

Agent. We use `smolagents` 1.26.0 `ToolCallingAgent`, MLX LM 0.31.3, and `mlx-community/Qwen2.5-7B-Instruct-4bit` at revision

Table 2: LLM agent task. "Agent" is the accuracy of the agent's answers, "Copy" the accuracy of the answer implied by the retained copy, and "Stale" the share of questions about stale copies, all in percent. Agent accuracies show means and 95% intervals across 12 instances with 200 questions each. Full refresh is unbudgeted and identical at both budgets.

	$B = 3$			$B = 1$		
Policy	Agent	Copy	Stale	Agent	Copy	Stale
Known-rate WIRS	85.7 ± 1.7	88.7	12.7	67.7 ± 3.4	70.0	33.9
UCB-WIRS	85.1 ± 1.8	88.0	13.5	67.8 ± 2.6	70.0	33.9
Rate optimism	80.0 ± 2.3	82.9	19.5	49.0 ± 2.7	50.5	56.6
MWA	83.0 ± 1.7	86.0	15.9	65.8 ± 3.5	67.9	36.1
Myopic	82.4 ± 2.7	85.2	17.3	59.1 ± 3.8	61.0	44.3
Full refresh	96.3 ± 1.0	100	0	96.3 ± 1.0	100	0

`c26a38f6a37d0a51b4e9a1eb3026530fa35d9fed`, with greedy decoding, at most 300 new tokens per call, and at most three steps. Each question starts a new conversation. The tools are `lookup_supplier`, which takes a supplier name and returns the unit price and stock of that supplier's copy, and `final_answer`. The agent sees no ages, rates, histories, or policy names. The prompt reads "A customer wants to order [k] units from [name]. Use lookup_supplier to read the record for [name] in working memory. If its stock is at least [k] units, the answer is the total price in dollars, which is [k] times the unit price, written as a whole number. Otherwise the answer is OUT OF STOCK. Use final_answer with ONLY the whole number or OUT OF STOCK."

Scoring. An answer is correct if the agent looked up the supplier named in the question and its final answer equals the correct answer. Greedy responses are reused for token-identical model inputs. Means and 95% Student-t intervals are computed across the 12 instances, and paired differences use identical questions.

Results. Table 2 reports all policies. Index optimism answers more questions correctly than rate optimism in all 12 instances at both budgets, by about 5 points at $B = 3$ and 19 points at $B = 1$. Its accuracy is within one point of known-rate WIRS at both budgets. In paired comparisons, known-rate WIRS is more accurate than myopic at both budgets and than MWA at $B = 3$. With every copy fresh, most of the agent's errors are `OUT OF STOCK` answers for a supplier with enough stock. On stale copies, the agent answers 11–15% of questions correctly, namely those for which the new record gives the same answer. For every budgeted policy, agent accuracy is one to three points below copy accuracy.