RISING MULTI-ARMED BANDITS WITH KNOWN HORIZONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Rising Multi-Armed Bandits (RMABs) model sequential decision problems where each arm's expected reward improves with repeated pulls. In such problems, the value of investing in an arm depends on how much time remains, making knowledge of the horizon useful side information, yet its benefit remains underexplored. We investigate this benefit through CURE-UCB, a horizon-aware algorithm that estimates each arm's cumulative reward over the remaining horizon. Theoretically, under structured assumptions, we prove that CURE-UCB uniformly dominates a representative horizon-agnostic algorithm and show that the advantage of horizon awareness can be substantial: on some instances, CURE-UCB incurs only $O(1)$ regret whereas the horizon-agnostic algorithm suffers $\Omega(T)$. Furthermore, we establish a regret upper bound for the general concave rising bandit setting whose growth-dependent term matches the known lower bound in its dependence on T. Empirically, across synthetic benchmarks and real-world model selection tasks, CURE-UCB achieves lower regret than both rising and non-stationary baselines over a wide range of horizons.

1 INTRODUCTION

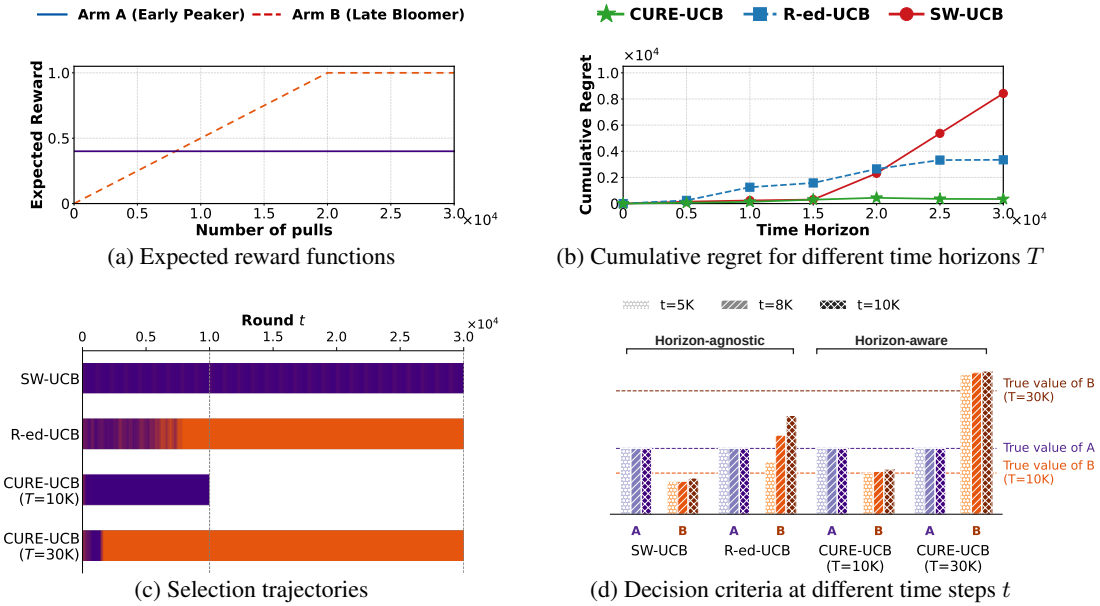
Rising Multi-Armed Bandits (RMABs) model environments where an arm's expected reward improves with repeated trials, capturing learning phenomena in applications such as hyperparameter optimization (Jamieson & Talwalkar, 2016; Li et al., 2018; 2020; Falkner et al., 2018), robot skill acquisition (Song et al., 2026; Yoon et al., 2024; Mehta et al., 2024; Portelas et al., 2020), and online crowdsourcing tasks (Sawwan & Wu, 2024; Song & Jin, 2021). In such practical scenarios, a learner optimizes the performance under a fixed and known horizon T, such as a number of training epochs. This known horizon is the extra information the learner can bring to each decision.

Figure 1 illustrates why knowledge of the horizon can be beneficial. Even with the same reward functions (Figure 1a), the value of investing in an arm depends on the horizon over which its improving rewards can be realized; i.e., an early peaker is preferable for a short horizon, whereas a late bloomer is preferable for a long horizon. This horizon dependence results in distinct regret patterns across horizons: as shown in Figure 1b, R-ed-UCB incurs large regret at short horizons ($T = 10K$), whereas SW-UCB performs poorly at longer horizons ($T = 30K$), while CURE-UCB maintains low regret across the considered horizons.

Figures 1c and 1d explain why these regret patterns arise. Figure 1c shows the different selection trajectories. SW-UCB continues to select the early peaker, whereas R-ed-UCB switches from the early peaker to the late bloomer and then commits to it. These trajectories remain unchanged across the horizons as the decision rules of both algorithms are independent of T. In contrast, CURE-UCB follows different trajectories for $T = 10K$ and $T = 30K$, adapting its selections to the given horizon. Figure 1d illustrates the difference in their decision criteria. SW-UCB focuses on current rewards and therefore tends to select the early peaker, whereas R-ed-UCB extrapolates each arm's reward to a specific future point and switches to the late bloomer once its projected reward becomes larger. Although these criteria capture different aspects of an arm's reward trajectory, each can lead to suboptimal selections depending on T.

Knowing T in advance could help to avoid such horizon-dependent suboptimal selections. CURE-UCB, our proposed CUmulative Reward Estimation UCB, is designed to put this information to use. At each round it uses the remaining horizon to estimate how much cumulative reward each

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.



(a) Expected reward functions

(b) Cumulative regret for different time horizons T

(c) Selection trajectories

(d) Decision criteria at different time steps t

Figure 1: **Demonstration of Horizon-Adaptiveness.** (a) Expected reward functions. **Arm A** represents an *Early Peaker* (high initial reward, limited growth), and **Arm B** represents a *Late Bloomer* (low initial reward, high potential). (b) Cumulative regret for different time horizons T, showing different regret trends among **CURE-UCB**, **R-ed-UCB**, and **SW-UCB**. (c) Selection trajectories: Horizon-agnostic baselines (R-ed-UCB and SW-UCB) follow a fixed selection trajectory regardless of the horizon T. In contrast, CURE-UCB adapts its trajectory to the horizon. (d) Decision criteria at different time steps t: SW-UCB focuses on current rewards, whereas R-ed-UCB focuses on a specific future point. CURE-UCB instead uses a horizon-aware criterion, evaluating the same arms differently depending on remaining time.

arm could deliver if committed to for the rest of the horizon. As illustrated in Figure 1d, this allows CURE-UCB to evaluate the same arms differently depending on T, leading to horizon-dependent decisions. We note that our algorithm aligns with the optimality criterion of RMAB (Heidari et al., 2016), presented in Proposition 3.2.

To characterize the advantage of horizon awareness, we first study Linear-Then-Flat (LTF) reward instances, a structured setting that captures the increasing-then-converging shape commonly observed in practical tasks such as model selection and hyperparameter tuning (Xia et al., 2024; Li et al., 2018). In this setting, we show that CURE-UCB uniformly dominates R-ed-UCB, a representative horizon-agnostic rising bandit algorithm (Metelli et al., 2022) for every finite horizon (Theorem 5.2). This result provides a controlled comparison showing that exploiting horizon information can consistently improve performance in this structured setting. Moreover, the advantage can be substantial rather than marginal: on some instances, CURE-UCB incurs only $O(1)$ regret while R-ed-UCB suffers $\Omega(T)$ (Proposition 5.3). Beyond the deterministic LTF setting, we establish a regret bound for CURE-UCB under general concave rising rewards (Theorem 5.5) and show that its growth-dependent term matches the known lower bound (Fiandri et al., 2024).

Empirically, we evaluate CURE-UCB across a broad range of synthetic and real-world environments. Our synthetic experiments cover both LTF dynamics and diverse concave reward functions, allowing us to examine its behavior under different reward-growth patterns and horizon lengths. Furthermore, we consider real-world model-selection tasks, including online model selection on IMDB (Maas et al., 2011) and LLM selection Xia et al. (2024), where model performance improves with additional training. Across these settings, CURE-UCB consistently achieves lower regret than both rising and non-stationary baselines over a wide range of horizons.

Our main contributions are summarized as follows:

- We propose CURE-UCB, a horizon-aware algorithm that directly incorporates the remaining horizon into arm evaluation. (Section 4).

- We theoretically characterize the benefit of horizon awareness, establishing uniform dominance in a structured setting, an order-level regret gap on some instances, and a general regret bound (Section 5).
- We demonstrate the practical benefit of CURE-UCB across diverse synthetic and real-world model selection tasks, consistently outperforming rising and non-stationary baselines over a wide range of horizons. (Section 6).

2 Related work

The RMAB setting is a special case of rested non-stationary bandits (Tekin & Liu, 2012; Levine et al., 2017), where the reward of an arm shifts only with the number of times it is selected. Standard non-stationary methods based on forgetting (Garivier & Moulines, 2011; Trovo et al., 2020) or restarts (Allesiardo et al., 2017; Besbes et al., 2014) are designed to track a drifting current mean, and do not exploit the monotone rested structure of RMAB.

Rising bandit algorithms (Metelli et al., 2022; Fiandri et al., 2025; Xia et al., 2024; Patil et al., 2023; Fiandri et al., 2024; Blum & Ravichandran, 2025) instead capture this rising trend. Most are horizon-agnostic, targeting a surrogate such as the instantaneous reward at a specific future time, which is a sound objective as the horizon grows large but is no longer aligned with the cumulative reward attainable in the time left once T is finite.

A separate line of work explicitly incorporates the horizon T, and our contribution is situated here. Heidari et al. (2016) characterize the horizon-dependent optimal policy, but only consider deterministic rewards and provide a sublinear policy-regret guarantee, but not the stochastic finite-sample regret guarantee considered here. Amichay & Mansour (2025) provide a tight regret bound with R-ed-AE, but confine it only to linear drift. In both studies, horizon awareness is a means to an end rather than an object of study, and the quantitative benefit of knowing T is left open. CURE-UCB is built to answer exactly this: through both theoretical and empirical analysis, it isolates the effect of horizon awareness that prior work leaves implicit.

3 Problem Formulation

We consider the RMAB problem, where the expected reward of each arm increases with the number of times it has been selected. Let K be the number of arms, indexed by $[K] := \{1, \dots, K\}$. At each round $t \in [T]$, the learner selects an arm $I_t \in [K]$ and observes a reward X_t. This reward is drawn independently from a distribution $D_{I_t}(N_{I_t,t-1}+1)$, where $N_{i,t-1}$ denotes the number of times arm i has been selected up to time $t-1$.

We assume that the reward distribution $D_i(n)$ is σ^2-subgaussian with a known variance parameter σ^2. Let $\mu_i(n) := \mathbb{E}_{X\sim D_i(n)}[X]$ denote the expected reward of arm i on its n-th pull, assuming $\mu_i(n) \in [0,1]$ for all $i \in [K]$ and $n \in [T]$. We denote the problem instance by the vector of mean reward functions $\boldsymbol{\mu} := (\mu_1, \mu_2, \dots, \mu_K)$. A defining characteristic of the rising bandit setting is that, for all arms $i \in [K]$, the mean rewards are non-decreasing in the number of plays n:

$$\gamma_i(n) := \mu_i(n+1) - \mu_i(n) \geq 0, \quad n \geq 1\ . \tag{1}$$

The learner's history up to time t is denoted by $\mathcal{H}_t := \{(I_\tau, X_\tau) : \tau \in [t]\}$. A policy π maps this history $\mathcal{H}_{t-1}$ to an arm $I_t \in [K]$ (or a distribution on arms) to be selected at round t.

For analytical tractability, consistent with the rising bandit literature (Heidari et al., 2016; Metelli et al., 2022; Patil et al., 2023), we assume the concavity of the mean reward function μ_i. This implies that while the expected rewards increase, the marginal gain from each additional pull is non-increasing.

Assumption 3.1. (Concavity) For all arms $i \in [K]$ and number of plays $n \geq 1$, the increments are non-increasing:

$$\gamma_i(n) \geq \gamma_i(n+1). \tag{2}$$

Regret Minimization The objective of the learner is to maximize the cumulative reward over a finite horizon T. Let $V(\pi, T) := \mathbb{E}_\pi\left[\sum_{t=1}^{T} X_t\right]$ be the expected cumulative reward of a policy π.

We define the optimal policy π^* as the policy that maximizes the expected cumulative reward:

$$\pi^* := \arg\max_{\pi} V(\pi, T). \tag{3}$$

The performance of a policy π is measured by its regret, defined as the difference between the expected cumulative reward of the optimal policy and that of π:

$$Reg_{\boldsymbol{\mu}}(\pi, T) := V(\pi^*, T) - V(\pi, T). \tag{4}$$

It is crucial to distinguish the regret definition employed here from that used in standard adversarial bandits. Standard adversarial bandit algorithms typically optimize for *external regret* (Auer et al., 2002; Blum & Mansour, 2007), which measures the learner against the best policy on the trajectory it actually ran. However, this metric is conceptually inadequate when rewards depend on past actions, as in the RMAB setting, because it ignores higher rewards attainable on other trajectories. For this reason, *policy regret* (Arora et al., 2012) has been established as the standard metric in rising bandits for regret minimization (Metelli et al., 2022; Song et al., 2026; Genalti et al., 2024; Heidari et al., 2016). *Policy regret* instead compares the learner against the optimal policy π^* over all possible trajectories, i.e., the highest expected cumulative reward attainable over the horizon T by any policy, independent of the trajectory the learner actually ran. While identifying π^* generally involves dynamic planning, the non-decreasing reward structure simplifies the optimal policy. Specifically, the optimal policy for a given T reduces to selecting a single arm that maximizes cumulative reward throughout the horizon (Heidari et al., 2016):

Proposition 3.2 (Horizon-dependent optimality, Heidari et al. (2016))**.** *For an RMAB problem with a finite horizon T, the optimal policy π^* consists of selecting the same arm $i^* \in [K]$ at all $t \in [T]$, where i^* is the arm maximizing the cumulative reward over the horizon T:*

$$i^* = \arg\max_{i \in [K]} \sum_{t=1}^{T} \mu_i(t) \ . \tag{5}$$

This proposition shows that the optimal policy depends on T. More precisely, it selects a single arm i^* throughout the horizon, and this arm depends on T, so i^* can differ across horizons.

4 ALGORITHM

Motivated by the theoretical insight that the optimal policy commits to the single arm with the largest cumulative reward over the horizon, we propose the CUmulative Reward Estimation UCB (CURE-UCB) algorithm. By leveraging the additional information of the horizon T, CURE-UCB aligns its selection with the criterion of Proposition 3.2. The complete procedure is outlined in Algorithm 1.

Algorithm 1 CURE-UCB $\pi_\epsilon^{\text{cure}}$

1: Input: Horizon T, Number of arms K, sliding window parameter ϵ
2: Initialize: Select each arm $i \in [K]$ for $\lceil \frac{2}{\epsilon} \rceil$ times.
3: **for** $t = \lceil \frac{2}{\epsilon} \rceil K + 1, \ldots, T$ **do**
4: Compute index $B_i(t)$ for each arm, where $B_i(t)$ is defined in equation 6
5: Select arm $I_t = \arg\max_{i \in [K]} B_i(t)$ and observe X_t.
6: Update estimates and counter $N_{I_t} \leftarrow N_{I_t} + 1$.
7: **end for**

4.1 THE HORIZON-ADAPTIVE INDEX

We now describe the construction of the horizon-adaptive index $B_i(t)$ in detail. Let $X_i(l)$ denote the reward observed when arm i is selected for the l-th time. The index is formally defined as the

sum of three distinct components:

$$B_i(t) := \underbrace{\frac{1}{h_i}\sum_{l=N_{i,t-1}-h_i+1}^{N_{i,t-1}} X_i(l)}_{\text{(i) Recent average}} + \underbrace{\frac{(T-t+2)}{2}\cdot\frac{1}{h_i}\sum_{l=N_{i,t-1}-h_i+1}^{N_{i,t-1}}\frac{X_i(l)-X_i(l-h_i)}{h_i}}_{\text{(ii) Expected future } \textit{cumulative} \text{ gain}}$$
$$+\underbrace{\sigma\sqrt{\frac{2\left[3(T-t+2)^2+8h_i^2\right]\log t^3}{4h_i^3}}}_{\text{(iii) Exploration bonus}} \tag{6}$$

where σ^2 is the known variance parameter and $h_i := \max\{\lfloor \epsilon N_{i,t-1}\rfloor, 1\}$. The sliding window parameter ϵ controls a bias-variance tradeoff in estimation.

The components of $B_i(t)$ are interpreted as follows:

(i) *Recent average* estimates the current expected reward of arm i.

(ii) *Expected future cumulative gain* captures the future potential from the arm's growth over the remaining horizon $(T-t+1)$.

(iii) *Exploration bonus* accounts for estimation uncertainty, following the UCB principle (Lattimore & Szepesvári, 2020).

Remark: Comparison between CURE-UCB and R-ed-UCB

To clarify the benefit of horizon awareness in the index design, it is instructive to contrast CURE-UCB with R-ed-UCB (Metelli et al., 2022), which is a representative horizon-agnostic rising algorithm. Among existing baselines, R-ed-UCB is algorithmically closest to CURE-UCB: both are UCB-based algorithms and utilize estimates of reward growth rates. Therefore, comparing these two algorithms isolates the central effect of horizon awareness in the index. The index of R-ed-UCB is given as follows:

$$B_i^{red}(t) := \underbrace{\frac{1}{h_i}\sum_{l=N_{i,t-1}-h_i+1}^{N_{i,t-1}} X_i(l)}_{\text{(i) Recent average}} + \underbrace{\frac{1}{h_i}\sum_{l=N_{i,t-1}-h_i+1}^{N_{i,t-1}}(t-l)\frac{X_i(l)-X_i(l-h_i)}{h_i}}_{\text{(ii) Expected future gain (}\textit{agnostic of horizon } T\text{)}}$$
$$+\underbrace{\sigma(t-N_{i,t-1}+h_i-1)\sqrt{\frac{10\log(1/\delta_t)}{h_i^3}}}_{\text{(iii) Exploration bonus}}, \tag{7}$$

where δ_t is the confidence parameter. While both algorithms share the same overall structure, the sums of terms (i) and (ii) estimate fundamentally different quantities. Firstly, R-ed-UCB targets the expected reward at a single future point $\mu_i(t)$:

$$\mu_i(t) = \mu_i(n) + \sum_{l=n}^{t-1}\gamma_i(l) \le \mu_i(n) + (t-n)\gamma_i(n)\ , \tag{8}$$

where equation 8 follows from Assumption 3.1. Plugging in sample estimates of $\mu_i(n)$ and $\gamma_i(n)$ and adding an exploration bonus yields the R-ed-UCB index given in equation 7. We note that $\mu_i(t)$ is a surrogate for the optimality criterion of Proposition 3.2, adopted in the absence of horizon knowledge. This surrogate is sound as $T\to\infty$, where the optimal policy commits to the arm with the highest asymptotic reward. At a finite T, however, it is not exactly aligned with the optimality criterion, which can lead to a suboptimal choice in cases such as Figure 1.

In contrast, CURE-UCB targets the cumulative reward over the remaining horizon:

$$\sum_{l=1}^{T-t+1}\mu_i(N_{i,t-1}+l) \le \sum_{l=1}^{T-t+1}\left(\mu_i(N_{i,t-1}) + l\gamma_i(N_{i,t-1})\right) \tag{9}$$
$$= (T-t+1)\mu_i(N_{i,t-1}) + \frac{(T-t+1)(T-t+2)}{2}\gamma_i(N_{i,t-1})\ , \tag{10}$$

where equation 9 follows from Assumption 3.1. Dividing by $(T - t + 1)$ and plugging in sample estimates with an exploration bonus recovers the CURE-UCB index (equation 6). This targets the remaining-horizon cumulative reward that governs optimality in Proposition 3.2, via a concave upper bound that knowledge of T makes available.

5 Analysis

In this section, we provide several theoretical analyses of the proposed CURE-UCB algorithm. We denote CURE-UCB with sliding-window parameter ϵ by $\pi_\epsilon^{\text{cure}}$. We first analyze CURE-UCB in the deterministic LTF setting (Section 5.1) and then derive a regret bound for general concave rising rewards (Section 5.2).

5.1 Horizon Awareness in the Linear-Then-Flat (LTF) setting

To substantiate the theoretical advantage of horizon awareness, we analyze the performance in a concrete structured environment, i.e., the "Linear-Then-Flat" (LTF) setting, which serves as a canonical example where the optimal strategy shifts sharply based on the horizon. We note that this LTF setting is an illustrative example: a simplified abstraction of the increasing-then-converging training curves observed in practice (Xia et al., 2024; Li et al., 2018). Our regression analysis in Appendix C further shows that LTF closely approximates these real-world training curves. We first formally define the LTF instance class:

Definition 5.1. (Linear-Then-Flat Instance) Let $\mathfrak{U}_K$ denote the set of K-armed Linear-Then-Flat problem instances. For any instance $\boldsymbol{\mu} \in \mathfrak{U}_K$, the expected reward of arm i at its n-th play is:

$$\mu_i(n) = \min\{b_i, a_i \cdot n\} \ , \tag{11}$$

where $a_i \in (0, 1)$ is the growth rate (slope) and $b_i \in (0, 1]$ is the maximum attainable reward. We define the saturation time of arm i as $c_i := b_i/a_i$ and assume that $c_i \in \mathbb{N}$. Thus, arm i grows linearly at rate a_i until it reaches b_i at $n = c_i$, after which it yields the constant reward b_i.

For every deterministic LTF instance $\boldsymbol{\mu} \in \mathfrak{U}_K$, we establish the uniform dominance of CURE-UCB over R-ed-UCB (Metelli et al., 2022). For comparison, let $\pi_\epsilon^{\text{red}}$ denote R-ed-UCB with sliding-window parameter ϵ. Since each observed reward equals its exact mean, we can use a unit sliding window by setting $\epsilon = 0$ and select every arm only once during initialization.

Theorem 5.2 (Uniform Dominance in deterministic LTF)**.** *Consider a deterministic LTF instance $\mu \in \mathcal{U}_K$ ($\sigma = 0$) and π_0^{cure} and π_0^{red} with sliding-window parameter $\epsilon = 0$. Then, for any finite horizon T,*

$$Reg_{\boldsymbol{\mu}}(\pi_0^{\text{cure}}, T) \le Reg_{\boldsymbol{\mu}}\big(\pi_0^{\text{red}}, T\big) \ . \tag{12}$$

The intuition behind Theorem 5.2 is that CURE-UCB evaluates whether further exploration is worthwhile based on the remaining time. Thus, when R-ed-UCB explores an additional growing arm despite insufficient time for that exploration to pay off, CURE-UCB avoids such unnecessary exploration, leading to the uniform regret dominance. The detailed proof is provided in Appendix D.

Furthermore, we show that the regret difference between CURE-UCB and R-ed-UCB can be critically large:

Proposition 5.3 (The dominance gap can be linear)**.** *For every fixed $K \ge 2$ and sufficiently large T, there exists a deterministic LTF instance $\boldsymbol{\mu} \in \mathcal{U}_K$ such that:*

$$Reg_{\boldsymbol{\mu}}(\pi_0^{\text{red}}, T) = \Omega(T), \quad Reg_{\boldsymbol{\mu}}(\pi_0^{\text{cure}}, T) = O(1) \ . \tag{13}$$

Proposition 5.3 shows that the dominance in Theorem 5.2 can be far from marginal. This shows that the benefit of horizon awareness can shape performance at the leading order. The detailed proof is provided in Appendix E.

5.2 General Regret Upper Bound

To quantify the complexity of general RMAB instances beyond the LTF setting, we adapt the definition of *cumulative increment* from Metelli et al. (2022), which characterizes the total magnitude of reward growth:

Definition 5.4. (Cumulative increment (Metelli et al., 2022)) Let $\gamma_{\max}(n) := \max_{i\in[K]} \gamma_i(n)$ be the maximum instantaneous increment across all arms at the n-th play. For any $M \in [T]$ and a parameter $q \in [0, 1]$, we define the *cumulative increment* $\Upsilon_{\boldsymbol{\mu}}(M, q)$ as follows:

$$\Upsilon_{\boldsymbol{\mu}}(M, q) := \sum_{n=1}^{M} (\gamma_{\max}(n))^q \,. \tag{14}$$

Leveraging this complexity measure, we establish the following regret upper bound, which guarantees the efficient performance of CURE-UCB in general concave rising environments.

Theorem 5.5 (General regret upper bound)**.** *Consider a general rising concave instance $\boldsymbol{\mu}$ satisfies Assumption 3.1 and $\pi_\epsilon^{\text{cure}}$. For $T > K\lceil 2/\epsilon \rceil$, $q \in [0, 1]$, $\sigma > 0$, and $\epsilon \in (0, 0.5)$, we have :*

$$Reg_{\boldsymbol{\mu}}(\pi_\epsilon^{\text{cure}}, T) \le \frac{K(1 + 2^{-q})T^q}{1 - 2\epsilon} \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil (1 - 2\epsilon)\frac{T}{K} \right\rceil, q \right) + O\left(\frac{K}{\epsilon} (\sigma T)^{2/3} (\log T)^{1/3} \right) \tag{15}$$

This result provides a regret guarantee for CURE-UCB beyond the structured LTF setting, covering general concave rising rewards. Moreover, the growth-dependent $\Upsilon_{\boldsymbol{\mu}}$-term is optimal in its dependence on T: Fiandri et al. (2024) establish a lower bound of $\Omega(T^q \Upsilon_{\boldsymbol{\mu}}(T, q))$, which also applies to the known-horizon setting. The proof of Theorem 5.5 and the extension of the lower bound to the known-horizon setting are provided in Appendices F and G, respectively.

6 Experiments

We evaluate CURE-UCB in two synthetic environments and two real-world model-selection tasks. The synthetic experiments consider the LTF setting (Section 6.1) and diverse concave reward dynamics (Section 6.2), while the real-world experiments cover online model selection on IMDB (Section 6.3) and LLM selection (Section 6.4). While real training curves need not be exactly monotone or concave at every step, they exhibit the overall increasing-then-converging trend motivating the RMAB model. We use these tasks to evaluate CURE-UCB beyond the idealized assumptions used in our analysis. Detailed experimental specifications are provided in Appendix B, and additional robustness analysis on window size and reward noise are reported in Appendix K.

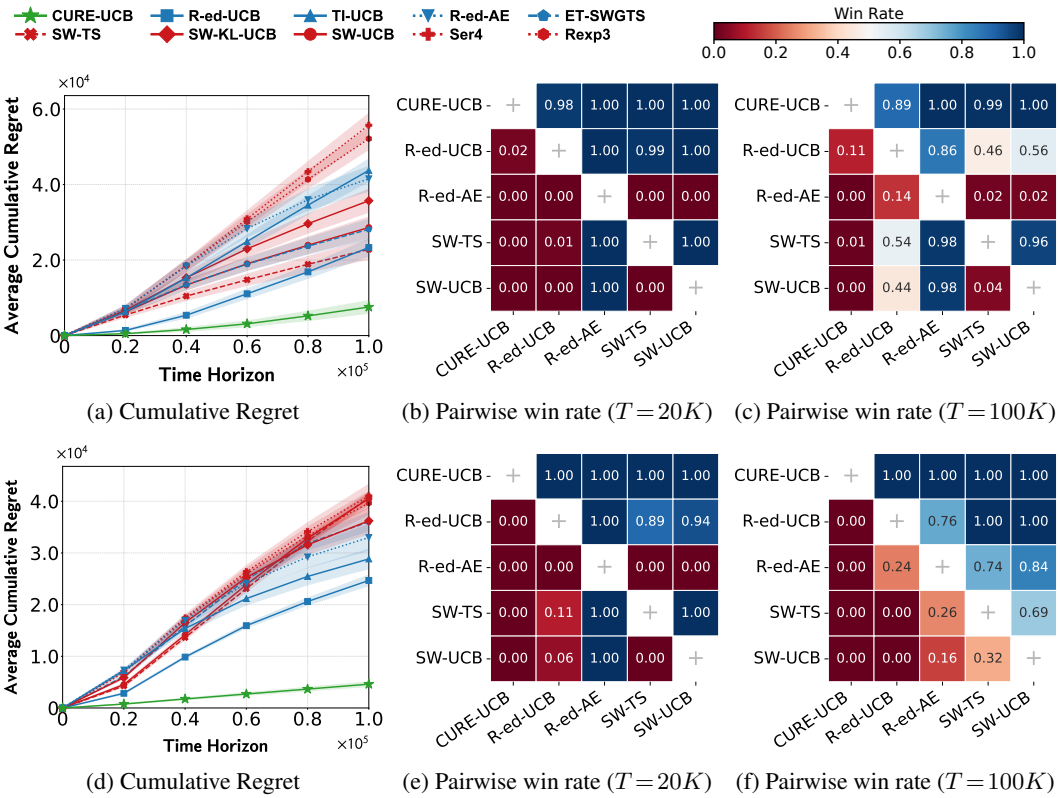
Baselines We compare CURE-UCB against two categories of algorithms. Detailed descriptions for all baseline algorithms are provided in Appendix I.

- **Rising bandit algorithms**: We adopt state-of-the art algorithms designed for the rising bandit framework, including R-ed-UCB (Metelli et al., 2022), ET-SWGTS (Fiandri et al., 2025), TI-UCB (Xia et al., 2024), and R-ed-AE (Amichay & Mansour, 2025). Among these, R-ed-AE is horizon-aware, whereas the others are horizon-agnostic.
- **Non-stationary bandit algorithms**: As the RMAB framework is a subclass of non-stationary environments, we include representative algorithms handling non-stationary rewards. These comprise sliding-window approaches (SW-UCB (Garivier & Moulines, 2011), SW-KL-UCB (Garivier & Cappé, 2011), SW-TS (Trovo et al., 2020)), as well as the restarting-based Rexp3 (Besbes et al., 2014) and the sequential elimination-based Ser4 (Allesiardo et al., 2017).

Evaluation Metrics We report final cumulative regret across separately evaluated horizons T. We also report the pairwise win rate of the representative algorithms, calculated over 100 independent problem instances at fixed horizons ($T = 20K$ and $100K$), where each instance is run with 40 random seeds.

6.1 Linear-Then-Flat Setting

To construct a rigorous evaluation environment for the Linear-Then-Flat (LTF) setting, we generated problem instances strictly following the formulation in Definition 5.1. We enforce structural diversity by randomizing the parameters rather than fixing them (detailed specifications in Appendix B). CURE-UCB performs well across the LTF instances: it achieves lower regret than every baseline on most instances at both horizons (Figures 2b and c) and attains the lowest cumulative regret overall (Figure 2a). At the long horizon, CURE-UCB achieves a pairwise win rate of 89% against R-ed-UCB, demonstrating a substantial empirical advantage under stochastic LTF rewards.



(a) Cumulative Regret (b) Pairwise win rate ($T = 20K$) (c) Pairwise win rate ($T = 100K$)

(d) Cumulative Regret (e) Pairwise win rate ($T = 20K$) (f) Pairwise win rate ($T = 100K$)

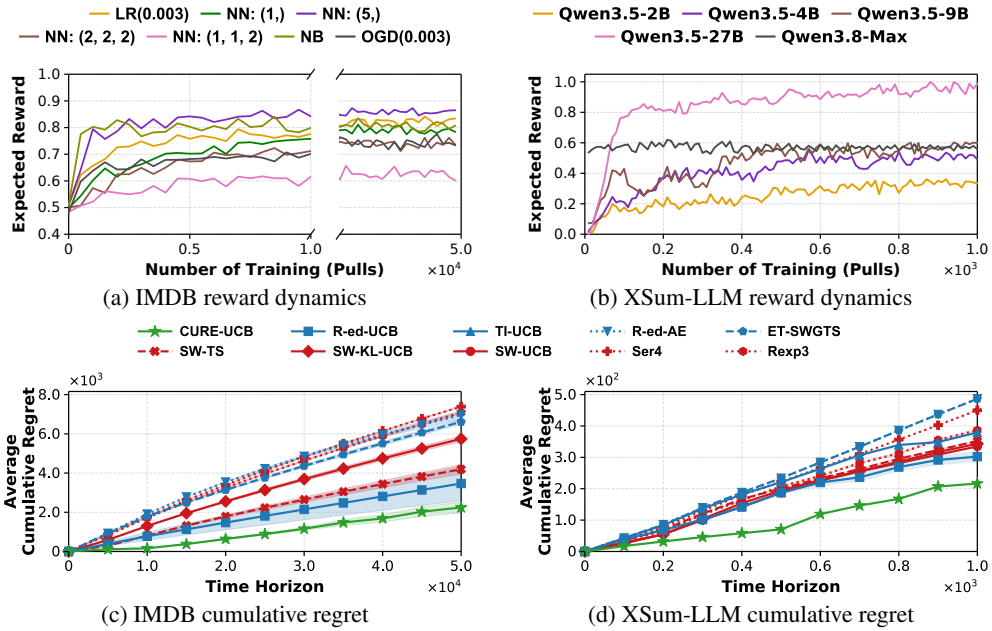
Figure 2: **Performance Analysis in LTF (top row) and Concave (bottom row) Settings.** (a, d) Cumulative regret as a function of the time horizon T. (b, e) Pairwise win rate at $T = 20K$ (short horizon). (c, f) Pairwise win rate at $T = 100K$ (long horizon). Each cell of the win-rate matrices shows how often the row algorithm attains lower regret than the column algorithm. Shaded regions and error bars denote 95% confidence intervals over 40 seeds, with the full results over all baselines in Appendix J.

6.2 Concave Setting

We next extend our evaluation to the concave setting, constructed using diverse concave functions to simulate diverse non-linear dynamics (detailed specifications in Appendix B). Each instance places an early peaker against a late bloomer, so that which arm is best depends on the horizon T. As shown in Figure 2d, CURE-UCB attains the lowest average cumulative regret across all horizons, and it achieves lower regret than every baseline on essentially all instances at both the short ($T = 20K$) and long ($T = 100K$) horizons (Figures 2e and f). This consistency across horizons highlights robustness of CURE-UCB.

6.3 Online Model Selection (IMDB)

To demonstrate the practical applicability of our framework, we investigate the problem of online model selection for sentiment analysis using the IMDB dataset (Maas et al., 2011). In this task, each arm represents a distinct learning model (e.g., neural networks with different hyperparameters). The objective is to maximize cumulative validation accuracy through sequential model selection within a finite training budget T. We follow the IMDB online model-selection setup of Metelli et al. (2022), including the candidate models and training protocol. As illustrated in Figure 3a, real-world training curves exhibit persistent fluctuations and noise, presenting a significantly more challenging environment compared to synthetic settings.



(a) IMDB reward dynamics

(b) XSum-LLM reward dynamics

(c) IMDB cumulative regret

(d) XSum-LLM cumulative regret

Figure 3: **Online model selection on IMDB and XSum-LLM.** The top row shows reward dynamics, and the bottom row shows cumulative regret. The left column corresponds to IMDB sentiment analysis for horizons up to $T = 50K$, and the right column corresponds to LLM fine-tuning on XSum summarization for horizons up to $T = 1K$. The legends above (a) and (b) indicate model configurations, and the shared legend between the two rows indicates the algorithms used in the cumulative-regret plots. Shaded regions denote 95% confidence intervals over 40 seeds.

6.4 LLM SELECTION

We further evaluate CURE-UCB on the LLM fine-tuning task introduced by Xia et al. (2024), where each arm corresponds to a candidate language model and the per-step reward reflects model performance penalized by accumulated fine-tuning cost. We use five arms: four Qwen3.5 (Qwen Team, 2026a) models (2B, 4B, 9B, 27B), each fine-tuned with LoRA; their per-step reward rises as training proceeds and is penalized by an accumulated, size-dependent cost, so larger models improve more but pay a higher price. The fifth arm is Qwen3.8-Max (Qwen Team, 2026b), a 2.4T-parameter mixture-of-experts model with 95B active parameters. At this scale fine-tuning is practically infeasible, so we use Qwen3.8-Max off-the-shelf with in-context examples only. Because it is never fine-tuned, its reward stays high but flat, whereas the smaller fine-tuned arms start lower and rise with training. Detailed setup is in Appendix B.2. As shown in Figure 3d, CURE-UCB consistently achieves the lowest cumulative regret across all horizons, outperforming other baselines. This result demonstrates the robustness of CURE-UCB across diverse horizons in LLM-scale model selection.

7 CONCLUSION

In this paper, we highlighted the role of horizon awareness in RMAB, showing that knowledge of the horizon T enables direct alignment with the optimality structure of RMAB. As a concrete instantiation of this principle, we proposed CURE-UCB and established uniform dominance over R-ed-UCB in the deterministic LTF settings, a regret upper bound for general concave rising rewards, and consistent empirical gains across synthetic and various real-world tasks such as LLM model selection. One limitation of our work is that the uniform dominance result is restricted to the deterministic LTF setting. Extending this guarantee to broader stochastic or non-LTF settings is a natural direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to implement algorithms and experiments, assist with translation, provide critical ingredients for proofs, and check the correctness of the proofs of our theoretical results. We did not use generative AI tools to develop the problem formulation or theoretical framework, formulate mathematical claims, design the research methodology or experiments, generate synthetic datasets, clean or reformat datasets, or interpret results; qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to draft parts of the paper and edit the text for readability. We reviewed all AI-assisted work. All theorems and proofs were formulated and written by the authors; LLM-based checking served only as an additional verification step, and every issue it raised was examined and resolved by the authors. LLM-generated code was reviewed and tested by the authors for consistency with the algorithm descriptions in the paper. All AI-assisted text was revised by the authors, and every technical statement was checked against the corresponding results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283, 2017.

Omer Amichay and Yishay Mansour. Rising rested mab with linear drift. *arXiv preprint arXiv:2501.04403*, 2025.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.

Avrim Blum and Kavya Ravichandran. Nearly-tight approximation guarantees for the improving multi-armed bandits problem. In *36th International Conference on Algorithmic Learning Theory*, 2025. URL https://openreview.net/forum?id=oytwNAlEfY.

Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In *International conference on machine learning*, pp. 1437–1446. PMLR, 2018.

Marco Fiandri, Alberto Maria Metelli, and Francesco Trovo. Rising rested bandits: Lower bounds and efficient algorithms. *arXiv preprint arXiv:2411.14446*, 2024.

Marco Fiandri, Alberto Maria Metelli, and Francesco Trovò. Thompson sampling-like algorithms for stochastic rising bandits. *arXiv preprint arXiv:2505.12092*, 2025.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International conference on algorithmic learning theory*, pp. 174–188. Springer, 2011.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto Maria Metelli. Graph-triggered rising bandits. In *Forty-first International Conference on Machine Learning*, 2024.

Hoda Heidari, Michael J Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *IJCAI*, pp. 1562–1570, 2016.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics*, pp. 240–248. PMLR, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. *Advances in neural information processing systems*, 30, 2017.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Yang Li, Jiawei Jiang, Jinyang Gao, Yingxia Shao, Ce Zhang, and Bin Cui. Efficient automatic cash via rising bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:4763–4771, 2020.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Shaunak A Mehta, Soheil Habibian, and Dylan P Losey. Waypoint-based reinforcement learning for robot manipulation tasks. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 541–548. IEEE, 2024.

Alberto Maria Metelli, Francesco Trovo, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *International Conference on Machine Learning*, pp. 15421–15457. PMLR, 2022.

Vishakha Patil, Vineet Nair, Ganesh Ghalme, and Arindam Khan. Mitigating disparity while maximizing reward: Tight anytime guarantee for improving bandits. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 4100–4108. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/456. URL https://doi.org/10.24963/ijcai.2023/456. Main Track.

Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments. In *Conference on Robot Learning*, pp. 835–853. PMLR, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Abdalaziz Sawwan and Jie Wu. Diversity-based recruitment in crowdsensing by combinatorial multi-armed bandits. *Tsinghua Science and Technology*, 30(2):732–747, 2024.

Seokbin Song, Youngsik Yoon, Siwei Wang, Wei Chen, and Jungseul Ok. Combinatorial rising bandits. In *International Conference on Learning Representations*, volume 2026, pp. 102490–102527, 2026.

Yiwen Song and Haiming Jin. Minimizing entropy for crowdsourcing with combinatorial multi-armed bandit. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2021.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58:5588–5611, 2012.

Francesco Trovo, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

Yu Xia, Fang Kong, Tong Yu, Liya Guo, Ryan A Rossi, Sungchul Kim, and Shuai Li. Which llm to play? convergence-aware online model selection with time-increasing bandits. In *Proceedings of the ACM on Web Conference 2024*, pp. 4059–4070, 2024.

Youngsik Yoon, Gangbok Lee, Sungsoo Ahn, and Jungseul Ok. Breadth-first exploration on adaptive grid for reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024.

A Appendix

This material provides proof of theorems, details of environments and baselines, and additional experimental results:

- **Appendix B:** Detailed explanation for experiment
- **Appendix C:** Additional analysis for LTF setting
- **Appendix D:** Proof of Theorem 5.2
- **Appendix E:** Proof of Proposition 5.3
- **Appendix F:** Proof of Theorem 5.5
- **Appendix G:** Known-horizon extension of Thm 4.6 in (Fiandri et al., 2024)
- **Appendix H:** Proof of Lemmas
- **Appendix I:** Detailed description of the baselines in Section 6.
- **Appendix J:** Additional experimental results
- **Appendix K:** Additional experiments for robustness

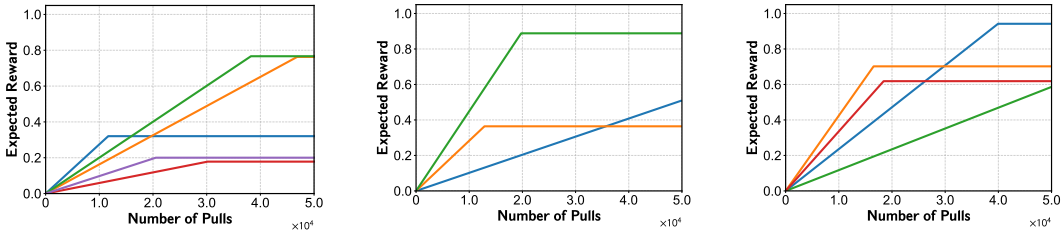
B Detailed explanation for experiment

B.1 Synthetic Environment Generation

In this section, we explain how we randomly generate the problem instances.

B.1.1 Linear-Then-Flat (LTF) Setting

We generated 100 independent problem instances for the LTF setting. Each instance consists of $K \in \{2, 3, 4, 5\}$ arms, where each arm's reward function follows the form $\mu_i(n) = \min(b_i, a_i \cdot n)$. The saturation level b_i is sampled uniformly from $[0.1, 1.0]$, and the saturation time t_{sat} is randomly selected from $[0.05T, T]$ for $T = 100K$. The slope is then determined as $a = b/t_{\text{sat}}$.



(a) Example 1 (b) Example 2 (c) Example 3

Figure 4: **Example instances of LTF Setting.**

B.1.2 Concave Setting

We generate 100 independent problem instances for the concave setting. The number of arms in each instance is sampled uniformly from $K \in \{3, \ldots, 15\}$, and the arm order is randomly permuted

within each instance so that the optimal arm is not identified by its index. The purpose of this synthetic benchmark is not to model the empirical distribution of real training curves, but to create diverse and nontrivial concave instances in which the preferred arm can genuinely depend on the horizon. Fully unconstrained random generation often produces trivial instances in which one arm dominates across all horizons, making horizon awareness largely irrelevant. We therefore use a role-based construction with an early peaker, a late bloomer, near-optimal challengers, and distractors, while randomizing the number of arms, arm ordering, growth families, and curve parameters to preserve substantial instance diversity. Each instance contains four types of arms: one optimal late bloomer, one early peaker, $\lfloor K/2 \rfloor$ near-optimal challengers, and $K - 2 - \lfloor K/2 \rfloor$ distractors. The optimal late bloomer is a low-start, high-potential; the early peaker starts high but saturates low, making it attractive early and inferior later; the near-optimal challengers closely track the optimal late bloomer but stay strictly below it; and the distractors saturate at clearly lower levels.

The expected reward of an arm at its n-th pull is a bounded concave curve

$$\mu(n) = s + (L - s)\, g(n),$$

where $s = \mu(0)$ is the initial reward, $L = \mu(H)$ is the reward reached at the design horizon $H = 100{,}000$, and $g(n) \in [0, 1]$ is a normalized growth function with $g(0) = 0$ and $g(H) = 1$. For each arm, we randomly draw g from three families:

$$g_{\mathrm{rat}}(n; k) = \min\left\{1, \frac{n/(n+k)}{H/(H+k)}\right\},$$

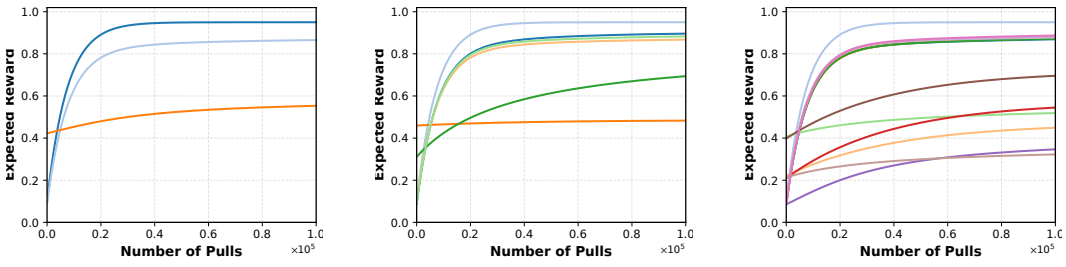
$$g_{\mathrm{exp}}(n; k) = \min\left\{1, \frac{1 - e^{-kn}}{1 - e^{-kH}}\right\},$$

$$g_{\mathrm{atan}}(n; k) = \min\left\{1, \frac{\arctan(kn)}{\arctan(kH)}\right\}.$$

each normalized to reach one at $n = H$ and held constant afterward by the $\min\{1, \cdot\}$ cap.

Each arm role is specified by its initial reward s and its horizon value $L = \mu(H)$. The optimal late bloomer has the lowest start and highest ceiling, $s = 0.10$ and $L = 0.95$, using the exponential family with $k = 4/30{,}000$ so that it completes most of its growth well before H. The early peaker has $s \in [0.40, 0.58]$ and $L \leq 0.60$. Each near-optimal challenger has $s \sim \mathcal{U}[0.08, 0.12]$ and $L \sim \mathcal{U}[0.85, 0.90]$, and its growth blends the optimal late bloomer's fast exponential shape with a slower sampled curve, so that it shadows the optimal late bloomer while keeping a small tail that improves up to H. Each distractor has $s \sim \mathcal{U}[0.05, 0.50]$ and $L \leq 0.70$.

Although every normalized curve formally reaches its limit L only at $n = H$, the shape parameter k governs how early an arm effectively saturates, i.e., how quickly it approaches L. We control this per arm through a target checkpoint $n_i^\star$: for each arm i, we set k so that about 85% of the rise from s to L is completed by $n_i^\star$, solving for k in closed form for the rational family and by binary search for the exponential and arctangent families, with a uniform $\pm 5\%$ multiplicative perturbation applied to the solved k. The optimal late bloomer instead uses a directly fixed rate $k = 4/30{,}000$, which front-loads its growth so that it rises faster than its competitors.



(a) Example 1 (b) Example 2 (c) Example 3

Figure 5: **Example instances of Concave Setting.**

B.2 LLM MODEL SELECTION

We largely follow the LLM fine-tuning setup of Xia et al. (2024), adapting their cost-penalized reward formulation to the Qwen3.5 model family and Qwen 3.8 max. To enable bandit simulation without re-training LLMs at every trial, we collect reward trajectories offline and replay them stochastically during simulation.

Arms and fine-tuning setup. We use five arms. Four correspond to the Qwen3.5 Qwen Team (2026a) family at sizes 2B, 4B, 9B, and 27B, each fine-tuned independently for abstractive summarization on XSum (sub-sampled to 16000 train / 1000 eval documents with seed 42). We use a plain-text prompt template

```
Summarize the following document:\n\n{document}\n\nSummary:
```

without chat formatting or reasoning mode. Each arm is adapted with LoRA (r=16, α=32, dropout 0.05) on all linear projections trained with AdamW (learning rate 5×10^{-5}, weight decay 0.01, cosine schedule with 100 warm-up steps), gradient accumulation to a constant effective batch size of 16, mixed-precision bfloat16, and a fixed budget of 1000 optimizer steps. ROUGE-2 F1 is evaluated on a 200-document held-out subset every 10 steps, giving one checkpoint per evaluation. The 27B arm additionally uses QLoRA with NF4 4-bit quantization on four 48 GB GPUs. The fifth arm, Qwen3.8-Max, is not fine-tuned; we evaluate it off-the-shelf with two in-context examples on the same subset, yielding a constant ROUGE-2 score across checkpoints.

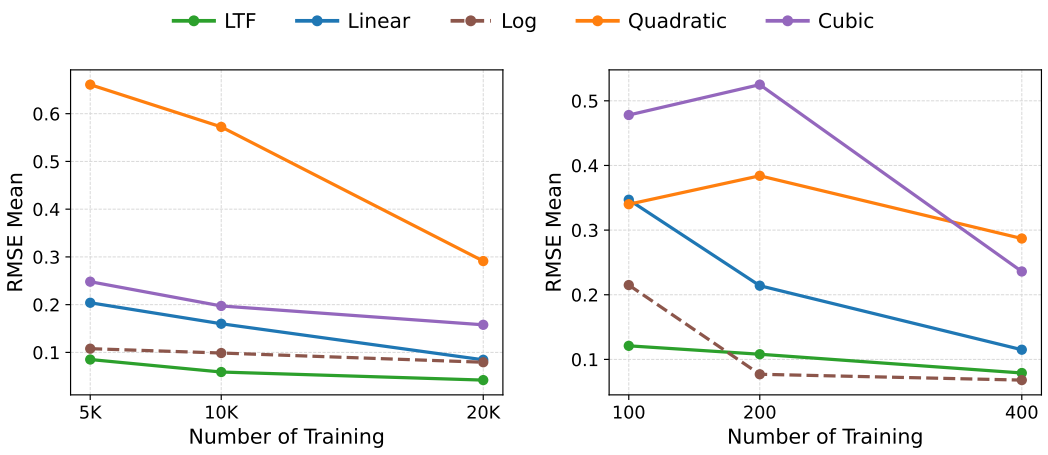
Reward construction. The raw ROUGE-2 checkpoint trajectories $\{\rho_t^{(k)}\}$ are converted into a fixed reward function via the following steps:

1. **Smoothing:** a 5-checkpoint uniform moving average yields $\tilde{\rho}_t^{(k)}$. This suppresses checkpoint-to-checkpoint evaluation noise so that the underlying learning trend, rather than measurement jitter, drives the reward.
2. **Cost penalty:** a size-dependent penalty gives the net reward $\nu_t^{(k)} = \tilde{\rho}_t^{(k)} - m_k \cdot t$ with cost coefficients $m_k \in \{5\times10^{-5}, 7.5\times10^{-5}, 1.05\times10^{-4}, 1.6\times10^{-4}\}$ for k = 2B, 4B, 9B, 27B. This charges each arm for the compute it spends, so that larger models, which improve more per step but are costlier to train, face a proportionally higher price. Since Qwen3.8-Max is not fine-tuned, it incurs no accumulated fine-tuning cost, so $m_k = 0$ for this arm.
3. **Peak-capping:** we enforce monotonicity by setting $\nu_t^{(k)} \leftarrow \nu_{t_k^\star}^{(k)}$ for all $t \geq t_k^\star = \arg\max_t \nu_t^{(k)}$, satisfying the rising-bandit assumption. We note that this is not just assumption to satisfy rising assumption. Instead, it indicates stopping the training after sufficient learning in real-world scenarios to avoid overfitting.
4. **Perturbation and normalization:** we add a fixed Gaussian perturbation $\eta_t^{(k)} \sim \mathcal{N}(0, 0.005^2)$, drawn once and held fixed across all simulations, zero-clip negative values, and apply a global affine rescaling $r_t^{(k)} = (\nu_t'^{(k)} - \min)/(\max - \min)$, where $\min$ and $\max$ are taken over all (k, t) pairs in the perturbed net-reward arrays. The perturbation keeps the curves from being perfectly smooth, mimicking the small run-to-run variation of real training, while the shared rescaling maps all arms onto a common $[0, 1]$ scale without distorting their relative order. This yields the fixed reward function μ_i.

Simulation. The environment is rested: μ_i is evaluated at each arm's local pull count, so at round t the selected arm returns $\mu_i(N_{i,t-1})$, obtained by linear interpolation (`np.interp`) over the precomputed length-1000 array, holding the first and last checkpoint values beyond the trajectory range. The algorithm observes $Y_{i,t} = \mu_i(N_{i,t-1}) + \varepsilon_{i,t}$ with $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_{\text{obs}}^2)$, so it queries an offline-measured reward rather than re-running the model at every trial.

B.3 EXPERIMENTAL ENVIRONMENT

All experiments were conducted on a server equipped with two AMD EPYC 7543 32-Core processors (64 cores in total) and four NVIDIA RTX A6000 GPUs (48GB memory each), running Ubuntu



(a) IMDB (b) XSum-LLM

Figure 6: Extrapolation accuracy of parametric families on real reward curves. Each family is fit to an initial prefix of each arm's training curve and evaluated by its RMSE against the full mean reward curve, averaged over arms. Most evaluated points lie beyond the prefix, so the RMSE reflects extrapolation of the remaining reward. Across both tasks, the two-parameter LTF family is consistently among the most accurate and clearly outperforms the higher-degree polynomial fits.

24.04 LTS. Synthetic experiments (Sections 6.1, 6.2) and IMDB experiment (Section 6.3) were run on CPU, while the LLM fine-tuning experiments (Section 6.4) utilized the GPUs.

C Additional analysis for LTF setting

We test whether the Linear-Then-Flat (LTF) form predicts real learning curves well from an early observation window. We use the mean reward curves $\mu_i(n)$ from the IMDB model-selection experiment (Figure 3a) and the XSum-LLM fine-tuning experiment (Figure 3b). For each arm, we fit each candidate family by least squares to a prefix of the first B pulls and evaluate its RMSE against the full curve, using $B \in \{5K, 10K, 20K\}$ for IMDB and $B \in \{100, 200, 400\}$ for XSum-LLM. Since most evaluated points are unseen during fitting, this RMSE measures extrapolation of the remaining reward rather than fit to the prefix.

We compare five families. The two-parameter LTF model matches Definition 5.1,

$$\widehat{\mu}_{\mathrm{LTF}}(n) = \min\{c_0,\, c_1 \cdot n\},$$

with growth rate c_1 and maximum attainable reward c_0. The other models are parameterized as follows:

$$\begin{aligned}
\text{Linear:} \quad & \widehat{\mu}(n) = c_0 + c_1 z, \\
\text{Quadratic:} \quad & \widehat{\mu}(n) = c_0 + c_1 z + c_2 z^2, \\
\text{Cubic:} \quad & \widehat{\mu}(n) = c_0 + c_1 z + c_2 z^2 + c_3 z^3, \\
\text{Log:} \quad & \widehat{\mu}(n) = c_0 \log(1 + c_1 n/B), \quad B = \text{number of observed data points},
\end{aligned}$$

where $z = 2n/B - 1$ is a time coordinate normalized to the fitting window, using 2, 3, 4, and 2 parameters, respectively. All predictions are clipped to $[0, 1]$ before evaluation.

As shown in Figure 6, the two-parameter LTF family is consistently among the most accurate on both tasks and clearly outperforms the higher-degree polynomial fits. The polynomial families fit the prefix but extrapolate poorly, since unconstrained polynomial growth does not encode saturation; the linear family cannot represent a transition to a plateau. LTF instead directly captures the two phases seen during training, improvement followed by a plateau once learning has effectively converged, which motivates it as a controlled test bed for reasoning about finite horizon.

D PROOF OF THEOREM 5.2

In this section, we provide the proof for Theorem 5.2. Since we consider deterministic reward setting (e.g., σ=0), CURE-UCB and R-ed-UCB (Metelli et al., 2022), we do not need to consider large sliding-window and can just set sliding window parameter $\epsilon = 0$ and initialize with selecting all arms once. For the reader's convenience, we recall index of both algorithms. Firstly, we recall index of R-ed-UCB in deterministic setting, which is given in Section 4.2 of (Metelli et al., 2022)

$$B_i^{red}(t) := \mu_i(N_{i,t-1}) + (t - N_{i,t-1})\gamma_i(N_{i,t-1} - 1) \ . \tag{16}$$

The index of CURE-UCB in deterministic setting is given as follows:

$$B_i^{cure}(t) := \mu_i(N_{i,t-1}) + \frac{(T-t+2)}{2}\gamma_i(N_{i,t-1} - 1) \ . \tag{17}$$

In LTF setting, the expected reward of every arm $i \in [K]$ is given as follows:

$$\mu_i(n) = \min(b_i, a_i n) \ . \tag{18}$$

For every arm, we define $c_i := \frac{b_i}{a_i}$, which means the time steps required to make arm i saturated. For convenience, we fix some T and assume that the arms are ordered in descending order with respect to a_i. That is, $a_1 \geq a_2 \geq \cdots \geq a_K$. For convenience, we assume throughout this proof that the initialization step has already been completed and that the slope a_i of each arm has been estimated exactly. Let $I_{cure}(T)$ and $I_{red}(T)$ denote the set of arms that are played at least once until time T after initialization by CURE-UCB and by R-ed-UCB respectively. Then, we assert the following claim:

Claim 1. Both $I_{cure}(T)$ and $I_{red}(T)$ grows monotonically with T, and its element appear in ascending order $\{1, 2, \cdots, K\}$.

This claim trivially holds because we assume that all arms increase linearly starting from zero. Consequently, both algorithms are compelled to select the arms in the order of decreasing order with respect to a_i.

Then, we now characterize the behavior of the two algorithms. Firstly, we consider CURE-UCB:

Claim 2. Let i^*_{cure} be defined as follows:

$$i^*_{cure} := \underset{i \in I_{cure}(T)}{\arg\max}\ b_i \ . \tag{19}$$

Then, CURE-UCB selects each arm $i \in I_{cure}(T) \setminus i^*_{cure}$ exactly $(c_i + 1)$ times, and select arm i^*_{cure} for the remaining $T - (K - |I_{cure}(T)|) - \sum_{i \in I_{cure} \setminus i^*_{cure}} (c_i + 1)$ times.

We begin by an observation that if CURE-UCB selects some arm i for the first time at time t_i after the initialization, then it must continue to select arm i for c_i consecutive rounds. The fact that CURE-UCB selects arm i at time t_i implies that the index of arm i, $B_i^{cure}(t_i)$, is the largest among all arms. Then, at time $t_i + 1$, we observe that $B_i^{cure}(t_i + 1) = B_i^{cure}(t_i) + a_i/2$, while for any other arm $j \in [K]$, we have $B_j^{cure}(t_i + 1) \leq B_j^{cure}(t_i)$. Therefore, CURE-UCB also selects arm i at time $t_i + 1$. This continues until arm i becomes flat, namely for c_i additional selections after initialization.

Also, we observe that if some arm i, after being selected $c_i + 1$ times, is selected again at some time t'_i, then arm i will continue to be selected for the remaining time. We consider two cases for another arm j: either by time t'_i it has already been selected $c_j + 1$ times, or it has not. If arm j has already been selected $c_j + 1$ times, then the fact that arm i is selected at time t'_i implies that $b_i > b_j$. It is then immediate that, for the remaining time, the index of arm i remains greater than that of arm j. If arm j has been selected fewer than $c_j + 1$ times, then by the previous observation, arm j has not been selected after initialization yet. At time $t'_i + 1$, comparing the indices yields

$$B_i^{cure}(t'_i + 1) = B_i^{cure}(t'_i) \tag{20}$$

for arm i, and

$$B_j^{cure}(t'_i + 1) = B_j^{cure}(t'_i) - \frac{a_j}{2} \tag{21}$$

for arm j. Then, since $B_j^{cure}(t'_i + 1) \leq B_j^{cure}(t'_i) \leq B_i^{cure}(t'_i) = B_i^{cure}(t'_i + 1)$, which means that arm i will be selected at time $t'_i + 1$. The same reasoning can be applied repeatedly for every round from $t = t'_i + 1$ up to T.

Based on the two preceding observations, we prove the claim. First, if every arm in $I_{cure}(T)$ has been selected at least $c_i + 1$ times, then by preceding observations, our claim is immediate. Now, suppose that there exists an arm in $I_{cure}(T)$ that has been selected at most c_i times. By the first observation, such an arm must be unique. Let us denote this unique arm by i'. Let t' denote the time at which arm i' is firstly selected. Then, by the first observation, we have $(T - t' + 1) < c_{i'}$. Define b' as the maximum value of b_i among the arms in $i \in I_{cure}(T) \setminus i'$, with $b' = 0$ when this set is empty. Then, at time t', we have

$$a_{i'} + \frac{a_{i'}}{2}(T - t' + 2) > b'. \tag{22}$$

Rearranging this inequality gives

$$b' < a_{i'}\frac{(T - t' + 4)}{2} \leq a_{i'}(T - t' + 2) \leq a_{i'}c_{i'} = b_{i'} \tag{23}$$

Hence, i' corresponds to i^*_{cure}, and the proof of the claim is complete.

Now, we make a claim for R-ed-UCB.

Claim 3. Let i'_{red} be defined as follows:

$$i'_{red} := \underset{i \in I_{red}(T)}{\arg\min}\, a_i. \tag{24}$$

Then, R-ed-UCB selects each arm $i \in I_{red}(T) \setminus \{i'_{red}\}$ at least $(c_i + 1)$ times.

Similarly, we begin by an observation that if R-ed-UCB selects some arm i for the first time at time t_i, then it must continue to select arm i for the next c_i rounds. By the design of R-ed-UCB, at time t_i, among the arms that are not selected, arm i is the one with the largest slope. Moreover, all arms with larger slopes than arm i must already have been pulled at least $c_i + 1$ times. Otherwise, R-ed-UCB would select another arm instead of arm i. Therefore, it is clear that arm i will be selected from time t_i to $\min\{T, t_i + c_i - 1\}$. However, unlike CURE-UCB, note that R-ed-UCB cannot guarantee that these arms are pulled exactly $(c_i + 1)$ times.

Then, we make following claim for relationship between $I_{cure}(T)$ and $I_{red}(T)$.

Claim 4. For any fixed T,

$$I_{cure}(T) \subseteq I_{red}(T)\ . \tag{25}$$

Let T_i^{red} be the first total time when i is selected after initialization by R-ed-UCB. Then, we can write T_i^{red} as follows:

$$\begin{aligned}
T_1^{red} &= K + 1 \\
T_2^{red} &\leq \max\left\{T_1^{red} + c_1, \frac{b_1}{a_2} + 1\right\} \\
T_3^{red} &\leq \max\left\{T_2^{red} + c_2, \frac{\max(b_1, b_2)}{a_3} + 1\right\} \\
&\cdots \\
T_k^{red} &\leq \max\left\{T_{k-1}^{red} + c_{k-1}, \frac{\max(b_1, b_2, \cdots, b_{k-1})}{a_k} + 1\right\}\ .
\end{aligned}$$

Similarly, let T_i^{cure} be the first time when i is selected after initialization by CURE-UCB. Then, we can write T_i^{cure} as follows:

$$\begin{aligned}
T_1^{cure} &= K + 1 \\
T_2^{cure} &= K + 1 + c_1 \\
T_3^{cure} &= K + 1 + c_1 + c_2 \\
&\cdots \\
T_k^{cure} &= K + 1 + \sum_{i=1}^{k-1} c_i\ .
\end{aligned}$$

Then, it suffices to show that $T_i^{red} \leq T$ for all arms $i \in I_{cure}(T)$. To show it, we utilize induction for $D_i := T_i^{red} - T_i^{cure} \leq \max\left\{\frac{\max(b_1,b_2,\ldots,b_{i-1})}{a_i} - K, 0\right\}$. For base case ($i = 1$), $D_i = 0$. Then, from previous results, we have:

$$D_k \leq \max\left\{T_{k-1}^{red} + c_{k-1} - T_k^{cure}, \frac{\max(b_1, b_2, \cdots, b_{k-1})}{a_k} + 1 - T_k^{cure}\right\} \tag{26}$$

$$= \max\left\{T_{k-1}^{red} + c_{k-1} - T_{k-1}^{cure} - c_{k-1}, \frac{\max(b_1, b_2, \cdots, b_{k-1})}{a_k} + 1 - 1 - K - \sum_{i=1}^{k-1} c_i\right\} \tag{27}$$

$$= \max\left\{D_{k-1}, \frac{\max(b_1, b_2, \cdots, b_{k-1})}{a_k} - K - \sum_{i=1}^{k-1} c_i\right\} \tag{28}$$

$$\leq \max\left\{D_{k-1}, \frac{\max(b_1, b_2, \ldots, b_{k-1})}{a_k} - K\right\} \tag{29}$$

$$\leq \max\left\{0, \frac{\max(b_1, b_2, \ldots, b_{k-2})}{a_{k-1}} - K, \frac{\max(b_1, b_2, \ldots, b_{k-1})}{a_k} - K\right\} \tag{30}$$

$$= \max\left\{\frac{\max(b_1, b_2, \ldots, b_{k-1})}{a_k} - K, 0\right\} . \tag{31}$$

At time T_k^{cure}, CURE-UCB selects arm k and it implies that:

$$\max(b_1, b_2, \ldots, b_{k-1}) \leq a_k + \frac{T - T_k^{cure} + 2}{2} a_k \tag{32}$$

$$\frac{\max(b_1, b_2, \ldots, b_{k-1})}{a_k} \leq \frac{T - T_k^{cure} + 4}{2} \tag{33}$$

Combining every result, we have:

$$D_k \leq \max\left\{\frac{T - T_k^{cure} + 4}{2} - K, 0\right\} \leq T - T_k^{cure} . \tag{34}$$

Therefore, $T_k^{red} \leq T$, which completes proof of Claim 4. Now, we proceed the proof of main theorem. Fix any horizon T and define $m := \max I^{cure}(T)$, which is last explored by CURE-UCB and $S^{cure}(T) := I^{cure}(T) \setminus \{m\}$. Then, by Claim 2, for every arm $i \in S^{cure}(T)$, it is guaranteed that i is selected by CURE-UCB at least c_i times. Also, by Claim 4, every arm $i \in I^{cure}(T)$ are selected by both R-ed-UCB and except for m, it is guaranteed that every arm $i \in S^{cure}(T)$ is selected at least c_i times by Claim 3. Therefore, we can cancel common selections: for every arm $i \in S^{cure}(T)$, both algorithm select it for c_i times. Then, we can consider remaining time $R = T - K - \sum_{i=1}^{m-1} c_i$. We consider two cases.

Case 1: R-ed-UCB selects arm m at most c_m times.
Let $b' := \max_{i \leq m-1} b_i$ and $D := T_m^{red} - T_m^{cure}$, $N_{m,T}^{\text{red}}$ be the number of selecting arm m by R-ed-UCB until T. In this case, R-ed-UCB selects arm m for $R - D$ times and previous plateaus for D times. Then, the cumulative reward of R-ed-UCB is upper bounded by $\sum_{t=2}^{R-D+1} \mu_m(t) + Db'$. In contrast to this, CURE-UCB selects arm m until its first zero increment or the horizon. If the block finishes, each subsequent selection yields $\max\{b', b_m\} \geq b_m$. Thus, its cumulative reward is at least $\sum_{t=2}^{R+1} \mu_m(t)$. If $D = 0$, the preceding upper and lower bounds already give the desired comparison. Then, the cumulative reward of CURE-UCB is lower bound by $\sum_{t=2}^{R+1} \mu_m(t)$. Then, equation 31 and equation 33 gives:

$$D \leq \frac{b'}{a_m} - K, \frac{b'}{a_m} \leq \frac{R+3}{2} . \tag{35}$$

Suppose henceforth that $D > 0$. Consequently, we have:

$$N_{m,T}^{\text{red}} = 1 + R - D \geq 1 + R - \frac{b'}{a_m} + K \tag{36}$$

$$\geq \frac{b'}{a_m} + K - 2 \tag{37}$$

$$\geq \frac{b'}{a_m} . \tag{38}$$

Since $N_{m,T}^{\text{red}} \leq c_m$, we have $\mu_m(R - D + 1) = a_m(R - D + 1) \geq b'$. Then, by monotonicity,

$$\sum_{t=2}^{R+1} \mu_m(t) - \sum_{t=2}^{R-D+1} \mu_m(t) - Db' = \sum_{t=R-D+2}^{R+1} \mu_m(t) - Db' \tag{39}$$

$$\geq Db' - Db' = 0 \tag{40}$$

Therefore, CURE-UCB has cumulative reward no smaller than R-ed-UCB.

Case 2: R-ed-UCB selects arm m at least $c_m + 1$ times.
In this case, we can also cancel arm m for c_m times. Let $b^\star := \max_{i \leq m} b_i$. For the remaining time $L = T - K - \sum_{i=1}^{m} c_i$, CURE-UCB selects arm with $b^\star$ for L times and R-ed-UCB may select arm $m + 1$. However, the fact that CURE-UCB does not select arm $m + 1$ implies that

$$b^\star \geq a_{m+1} \frac{L + 3}{2} . \tag{41}$$

It implies that for remaining time, the optimal policy is selecting arm with $b^\star$ continuously and therefore CURE-UCB is optimal.

Therefore, for both cases, it is guaranteed that

$$Reg_\mu(\pi_0^{\text{cure}}, T) \leq Reg_\mu\big(\pi_0^{\text{red}}, T\big) . \tag{42}$$

E Proof of Proposition 5.3

Proof. We first prove the claim for $K = 2$ and then extend it to arbitrary K.

Construction ($K = 2$). Let arm 1 be a constant early peaker and arm 2 a late bloomer:

$$\mu_1(n) = \tfrac{1}{2}, \qquad \mu_2(n) = \min\{1,\, a_2 n\}, \quad a_2 = \frac{1}{\lfloor \alpha T \rfloor}, \;\; \alpha \in (0, 1/2), \tag{43}$$

so that arm 2 grows linearly at rate a_2 and saturates at pull $c_2 := \lfloor \alpha T \rfloor$.

Optimal policy. By Proposition 3.2 the optimal policy plays a single arm for every time steps. The two candidates have cumulative rewards as follows:

$$\sum_{n=1}^{T} \mu_1(n) = \frac{T}{2}, \qquad \sum_{n=1}^{T} \mu_2(n) = a_2 \frac{c_2(c_2 + 1)}{2} + (T - c_2) = T - \frac{c_2 - 1}{2}. \tag{44}$$

Since $c_2 = \lfloor \alpha T \rfloor \leq T$, we have $T - \frac{c_2 - 1}{2} > \frac{T}{2}$, so arm 2 is the unique optimal arm and $\text{OPT} = \sum_{n=1}^{T} \mu_2(n)$.

R-ed-UCB. In the deterministic setting its index is $B_i^{\text{red}}(t) = \mu_i(N_{i,t-1}) + (t - N_{i,t-1})\, \gamma_i(N_{i,t-1} - 1)$. Arm 1 is constant, so $B_1^{\text{red}}(t) = \frac{1}{2}$. While arm 2 is in its linear phase ($N_{2,t-1} < c_2$),

$$B_2^{\text{red}}(t) = \mu_2(N_{2,t-1}) + (t - N_{2,t-1})\, a_2 = a_2 N_{2,t-1} + (t - N_{2,t-1})\, a_2 = a_2 t, \tag{45}$$

which depends only on t. Hence R-ed-UCB prefers arm 1 while $a_2 t \leq \frac{1}{2}$ and switches to arm 2 permanently once $t > \frac{1}{2a_2} = \frac{c_2}{2}$. It thus pulls the suboptimal arm 1 exactly $m := \left\lfloor \frac{c_2}{2} \right\rfloor$ times, then

plays arm 2 for the remaining $T - m$ rounds. Its cumulative reward is

$$V^{\mathrm{red}} = \frac{m}{2} + \sum_{n=1}^{T-m} \mu_2(n), \tag{46}$$

so, using $\mathrm{OPT} = \sum_{n=1}^{T} \mu_2(n)$,

$$\mathrm{Reg}_\mu(\pi_{\mathrm{det}}^{\mathrm{red}}, T) = \sum_{n=T-m+1}^{T} \mu_2(n) - \frac{m}{2} = m - \frac{m}{2} = \frac{m}{2} = \left\lfloor \frac{\alpha T}{4} \right\rfloor \geq \frac{\alpha T}{4} - 1, \tag{47}$$

where the middle step uses that the top m pulls of arm 2 are all at the saturated level 1 (since $T - m > c_2$ for $\alpha < 1/2$). Thus $\mathrm{Reg}_\mu(\pi_{\mathrm{det}}^{\mathrm{red}}, T) \geq \frac{\alpha T}{4} - 1$.

CURE-UCB. Its index carries the horizon factor, $B_i^{\mathrm{cure}}(t) = \mu_i(N_{i,t-1}) + \frac{T-t+2}{2}\gamma_i(N_{i,t-1} - 1)$. Right after initialization, the future-gain term of arm 2 is $\frac{T-t+2}{2}a_2 = \frac{T-t+2}{2c_2} = \frac{T-t+2}{2\lfloor \alpha T \rfloor} \geq \frac{T-t+2}{T}$, which exceeds $\frac{1}{2}$. Therefore, $B_2^{\mathrm{cure}} > B_1^{\mathrm{cure}}$ from the first index-based round. Then, as shown in the proof of Theorem 5.2, CURE-UCB commits to arm 2 until it saturates. After then, it is clear that arm 2 has larger value than arm 1, which leading to committing arm 2 for remaining time steps. Therefore, the CURE-UCB commits arm 2 for every time step except for initialization steps and the cumulative regret is:

$$\mathrm{Reg}_\mu(\pi_{\mathrm{det}}^{\mathrm{cure}}, T) = \tfrac{1}{2}. \tag{48}$$

Combining the two,

$$\mathrm{Reg}_\mu(\pi_{\mathrm{det}}^{\mathrm{red}}, T) - \mathrm{Reg}_\mu(\pi_{\mathrm{det}}^{\mathrm{cure}}, T) \geq \frac{\alpha}{4}T - \frac{3}{2} = \Omega(T), \tag{49}$$

which establishes the $K = 2$ case with any $\alpha \in (0, 1/2)$.

Extension to $K \geq 2$. Augment the two-armed instance with $K-2$ dummy arms that saturate after a single pull at a constant level below $\frac{1}{2}$, e.g. $\mu_i(n) = \frac{1}{4}$ for $i \geq 3$. These arms have zero growth ($\gamma_i = 0$) and reward strictly below both arm 1 and the saturated arm 2, so after the one initialization pull each, their index equals $\frac{1}{4}$ under both algorithms and is never maximal again. Consequently neither algorithm selects a dummy arm after initialization, the selection dynamics of arms 1 and 2 are unchanged, and each dummy contributes only an $O(1)$ additive term to both regrets, which cancels in the difference. Therefore $\mathrm{Reg}_\mu(\pi_{\mathrm{det}}^{\mathrm{red}}, T) - \mathrm{Reg}_\mu(\pi_{\mathrm{det}}^{\mathrm{cure}}, T) = \Omega(T)$ holds for every $K \geq 2$. □

F Proof of Theorem 5.5

Proof. Let $\tau := K\lceil 2/\epsilon \rceil$ be the number of initialization time steps. For well-estimated event, we define $\bar{\mu}_i(t)$ and $\widetilde{\mu}_i(t)$ as follows:

$$\bar{\mu}_i(t) := \frac{1}{h_i} \sum_{l=N_{i,t-1}-h_i+1}^{N_{i,t-1}} \left(X_i(l) + \frac{(T-t+2)}{2} \frac{X_i(l) - X_i(l-h_i)}{h_i} \right) , \tag{50}$$

$$\widetilde{\mu}_i(t) := \frac{1}{h_i} \sum_{l=N_{i,t-1}-h_i+1}^{N_{i,t-1}} \left(\mu_i(l) + \frac{(T-t+2)}{2} \frac{\mu_i(l) - \mu_i(l-h_i)}{h_i} \right) , \tag{51}$$

$$\beta_i(t) := \sigma \sqrt{\frac{2[3(T-t+2)^2 + 8h_i^2] \log t^3}{4h_i^3}} , \tag{52}$$

$$B_i(t) := \bar{\mu}_i(t) + \beta_i(t) . \tag{53}$$

We define well-estimated event $\mathcal{E}_t$ as follows:

$$\mathcal{E}_{i,t} := \{ |\widetilde{\mu}_i(t) - \bar{\mu}_i(t)| \leq \beta_i(t) \} , \tag{54}$$

$$\mathcal{E}_t := \bigcap_{i \in [K]} \mathcal{E}_{i,t} . \tag{55}$$

Then, we rewrite the regret as follows:

$$Reg_{\boldsymbol{\mu}}(\pi, T) = \sum_{t=1}^{T} \mathbb{E}\left[\mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t})\right] \tag{56}$$

$$\leq \tau + \sum_{t=\tau+1}^{T} \mathbb{E}\left[(\mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t}))\mathbb{1}\left\{\mathcal{E}_t\right\}\right]$$

$$+ \sum_{t=\tau+1}^{T} \mathbb{E}\left[(\mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t}))\mathbb{1}\left\{\mathcal{E}_t^c\right\}\right] \tag{57}$$

$$\leq \tau + \underbrace{\sum_{t=\tau+1}^{T} \mathbb{E}\left[(\mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t}))\mathbb{1}\left\{\mathcal{E}_t\right\}\right]}_{(A)} + \underbrace{\sum_{t=\tau+1}^{T} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t^c\right\}\right]}_{(B)} . \tag{58}$$

Then, we bound (B) term.

$$(B) = \sum_{t=\tau+1}^{T} \mathbb{P}\left(\mathcal{E}_t^c\right) \tag{59}$$

$$= \sum_{t=\tau+1}^{T} \mathbb{P}\left(\bigcup_{i\in[K]} \mathcal{E}_{i,t}^c\right) \tag{60}$$

$$\leq \sum_{t=\tau+1}^{T} \sum_{i\in[K]} \mathbb{P}(\mathcal{E}_{i,t}^c) , \tag{61}$$

where equation 60 holds due to De Morgan's Law and equation 61 follows from the application of the union bound. From Lemma H.1, we have:

$$\sum_{t=\tau+1}^{T} \sum_{i\in[K]} \mathbb{P}(\mathcal{E}_{i,t}^c) \leq \sum_{t=\tau+1}^{T} \sum_{i\in[K]} 2t^{-2} \leq \frac{2K}{\tau} . \tag{62}$$

Now, we bound (A) term. Under $\mathcal{E}$, we can decompose (A) term as follows:

$$\mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t}) \leq \mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t}) + B_{i_t}(t) - B_{i^*}(t) \tag{63}$$

$$= \mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t}) + (\bar{\mu}_{i_t}(t) + \beta_{i_t}(t)) - (\bar{\mu}_{i^*}(t) + \beta_{i^*}(t)) \tag{64}$$

$$\leq \mu_{i^*}(t) - \mu_{i_t}(N_{i_t,t}) + (\widetilde{\mu}_{i_t}(t) + 2\beta_{i_t}(t)) - \widetilde{\mu}_{i^*}(t) \tag{65}$$

$$= \underbrace{\mu_{i^*}(t) - \widetilde{\mu}_{i^*}(t)}_{(A1)} + \underbrace{\widetilde{\mu}_{i_t}(t) - \mu_{i_t}(N_{i_t,t})}_{(A2)} + \underbrace{2\beta_{i_t}(t)}_{(A3)} , \tag{66}$$

where equation 65 holds by the definition of $\mathcal{E}$. Write $[x]_+ := \max\{x, 0\}$. Then equation 66 implies that:

$$(A) \leq \mathbb{E}\left[\sum_{t=\tau+1}^{T} \left(\min\left\{1, [(A1)]_+\right\}\right) + \min\left\{1, [(A2)]_+\right\} + \min\left\{1, [(A3)]_+\right\}\right] . \tag{67}$$

Then, we bound (A1) term. We can rewrite (A1) term as follows:

$$(A1) = \mu_{i^*}(t) - \frac{1}{h_i^*} \sum_{l=N_{i^*,t-1}-h_{i^*}+1}^{N_{i^*,t-1}} \left(\mu_i(l) + \frac{(T-t+2)}{2} \frac{\mu_{i^*}(l) - \mu_{i^*}(l-h_{i^*})}{h_{i^*}}\right) \tag{68}$$

$$= \frac{1}{h_{i^*}} \sum_{l=N_{i^*,t-1}-h_{i^*}+1}^{N_{i^*,t-1}} \left(\mu_{i^*}(t) - \mu_{i^*}(l) - \frac{(T-t+2)}{2} \frac{\mu_{i^*}(l) - \mu_{i^*}(l-h_{i^*})}{h_{i^*}}\right) \tag{69}$$

$$= \frac{1}{h_{i^*}} \sum_{l=N_{i^*,t-1}-h_{i^*}+1}^{N_{i^*,t-1}} \left(\mu_{i^*}(t) - \mu_{i^*}(l) - \frac{(T-t+2)}{2h_{i^*}} \sum_{k=l-h_{i^*}}^{l-1} \gamma_{i^*}(k) \right) \tag{70}$$

$$\leq \frac{1}{h_{i^*}} \sum_{l=N_{i^*,t-1}-h_{i^*}+1}^{N_{i^*,t-1}} \left(\mu_{i^*}(t) - \mu_{i^*}(l) - \frac{(T-t+2)}{2} \gamma_{i^*}(l-1) \right) \tag{71}$$

$$\leq \frac{1}{h_{i^*}} \sum_{l=N_{i^*,t-1}-h_{i^*}+1}^{N_{i^*,t-1}} \left[t - l - \left\lfloor \frac{T-t+2}{2} \right\rfloor \right]_+ \gamma_{i^*} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right) \tag{72}$$

$$\leq \frac{1}{h_{i^*}} \sum_{l=N_{i^*,t-1}-h_{i^*}+1}^{N_{i^*,t-1}} T\gamma_{i^*} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right) \tag{73}$$

$$= T\gamma_{i^*} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right) , \tag{74}$$

where equation 70 holds by the definition of γ_i and equation 71 holds by concavity assumption, (i.e., $\gamma_i(n)$ is decreasing over n) and equation 72 holds by Lemma H.2.

With summation, we have:

$$\sum_{t=\tau+1}^{T} (A1) \leq \sum_{t=\tau+1}^{T} T\gamma_{i^*} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right) \tag{75}$$

$$\leq \sum_{t=\tau+1}^{T} T\gamma_{\max} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right) \tag{76}$$

$$\leq \sum_{t=2}^{\lceil (T-\tau)/2 \rceil + 1} 2T\gamma_{\max}(t) \tag{77}$$

$$\leq 2T \sum_{t=1}^{T} \gamma_{\max}(t) , \tag{78}$$

where equation 77 holds because each integer increment index occurs at most twice. Then, we have:

$$\sum_{t=\tau+1}^{T} \min\{1, [(A1)]_+\} \leq \sum_{t=\tau+1}^{T} \min\left\{1, T\gamma_{i^*} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right)\right\} \tag{79}$$

$$\leq \sum_{t=\tau+1}^{T} \min\left\{1, T\gamma_{max} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right)\right\} \tag{80}$$

$$\leq T^q \sum_{t=\tau+1}^{T} \gamma_{max} \left(\left\lfloor \frac{T-t+2}{2} \right\rfloor + 1 \right)^q \tag{81}$$

$$\leq \frac{KT^q}{1-2\epsilon} \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q \right) , \tag{82}$$

where equation 81 holds since when $\min\{1, x\} \leq \min\{1, x\}^q \leq x^q$ for any $x \geq 0$ and equation 82 holds by Lemma C.2. from (Metelli et al., 2022).

Then, we bound (A2) term.

$$(A2) = \frac{1}{h_{i_t}} \sum_{l=N_{i_t,t-1}-h_{i_t}+1}^{N_{i_t,t-1}} \left(\mu_{i_t}(l) + \frac{(T-t+2)}{2} \frac{\mu_{i_t}(l) - \mu_{i_t}(l-h_{i_t})}{h_{i_t}} \right) - \mu_{i_t}(N_{i_t,t}) \tag{83}$$

$$= \frac{1}{h_{i_t}} \sum_{l=N_{i_t,t-1}-h_{i_t}+1}^{N_{i_t,t-1}} \left(\mu_{i_t}(l) + \frac{(T-t+2)}{2} \frac{\mu_{i_t}(l) - \mu_{i_t}(l-h_{i_t})}{h_{i_t}} - \mu_{i_t}(N_{i_t,t}) \right) \tag{84}$$

$$\leq \frac{1}{h_{i_t}} \sum_{l=N_{i_t,t-1}-h_{i_t}+1}^{N_{i_t,t-1}} \frac{(T-t+2)}{2} \frac{\mu_{i_t}(l) - \mu_{i_t}(l-h_{i_t})}{h_{i_t}} \tag{85}$$

$$= \frac{(T-t+2)}{2h_{i_t}} \sum_{l=N_{i_t,t-1}-h_{i_t}+1}^{N_{i_t,t-1}} \sum_{k=l-h_{i_t}}^{l-1} \frac{1}{h_{i_t}} \gamma_{i_t}(k) \tag{86}$$

$$\leq \frac{(T-t+2)}{2h_{i_t}} \sum_{l=N_{i_t,t-1}-h_{i_t}+1}^{N_{i_t,t-1}} \gamma_{i_t}(l-h_{i_t}) \tag{87}$$

$$\leq \frac{(T-t+2)}{2} \gamma_{i_t}(N_{i_t,t-1} - 2h_{i_t} + 1) \ , \tag{88}$$

where equation 85 holds since $\mu_i(n)$ is non-decreasing and $l \leq N_{i_t,t-1} \leq N_{i_t,t}$ and equation 87 holds by concavity assumption.

Then, we have:

$$\sum_{t=\tau+1}^{T} \min\{1, [(A2)]_+\} \leq \sum_{t=\tau+1}^{T} \min\left\{1, \frac{(T-t+2)}{2} \gamma_{i_t}(N_{i_t,t-1} - 2h_{i_t} + 1)\right\} \tag{89}$$

$$\leq \sum_{t=\tau+1}^{T} \min\left\{1, \frac{T}{2} \gamma_{i_t}(N_{i_t,t-1} - 2h_{i_t} + 1)\right\} \tag{90}$$

$$\leq \sum_{i=1}^{K} \sum_{n=\lceil 2/\epsilon \rceil}^{N_{i,T}-1} \min\left\{1, \frac{T}{2} \gamma_i(n - 2h_i + 1)\right\} \tag{91}$$

$$\leq \frac{KT^q}{2^q(1-2\epsilon)} \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q\right), \tag{92}$$

where equation 92 holds by Lemma C.2. from (Metelli et al., 2022).

Then, we bound (A3) term. We also utilize the fact that the regret for one time step cannot precede 1.

$$\sum_{t=\tau+1}^{T} \min\{1, [(A3)]_+\} \leq \sum_{i=1}^{K} \sum_{j=1}^{N_{i,T}} \min\left\{1, 2\sigma\sqrt{\frac{2[3T^2 + 8(\epsilon j)^2]3\log T}{4(\epsilon j)^3}}\right\}$$

$$\leq \sum_{i=1}^{K} \sum_{j=1}^{N_{i,T}} \min\left\{1, \frac{\sigma T\sqrt{\log T}}{\epsilon^{3/2} j^{3/2}} \sqrt{18 + 48\epsilon^2}\right\}$$

$$\leq K \sum_{j=1}^{\infty} \min\left\{1, \frac{C}{j^{3/2}}\right\} \quad \text{where } C = \sigma T\sqrt{\log T}\epsilon^{-3/2}\sqrt{18+48\epsilon^2}$$

$$\leq K \int_0^{\infty} \min\left\{1, Cx^{-3/2}\right\} dx \tag{93}$$

$$\leq K\left(j^* + \int_{j^*}^{\infty} Cx^{-3/2} dx\right) \quad \text{with } j^* = C^{2/3}$$

$$\leq 3KC^{2/3}$$

$$= 3K\left(\sigma T\sqrt{\log T}\epsilon^{-3/2}\sqrt{18+48\epsilon^2}\right)^{2/3}$$

$$= \frac{3K}{\epsilon}(\sigma T)^{2/3}(\log T)^{1/3}(18+48\epsilon^2)^{1/3} \ . \tag{94}$$

Combining all results, we have:

$$Reg_{\mu}(\pi_{\epsilon}^{\text{cure}}, T) \leq 2K\left\lceil \frac{2}{\epsilon} \right\rceil + \frac{K(1+2^{-q})T^q}{1-2\epsilon} \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q\right) \tag{95}$$

$$+ \frac{3K}{\epsilon}(\sigma T)^{2/3}(\log T)^{1/3}(18 + 48\epsilon^2)^{1/3}$$

□

G KNOWN-HORIZON EXTENSION OF THM 4.6 IN (FIANDRI ET AL., 2024)

We first provide formal statement for lower bound.

Theorem G.1 (Known-horizon extension of the lower bound, (Fiandri et al., 2024)). *Let π denote arbitrary policy (including horizon-aware). For any $T \geq 4$ and $q \in [0, 1]$, there exists a deterministic ($\sigma = 0$) two-armed instance $\boldsymbol{\mu}^\star$, with non-decreasing and concave reward functions bounded in $[0, 1]$, such that*

$$Reg_{\boldsymbol{\mu}^\star}(\pi, T) \;\geq\; \frac{T^q}{64}\,\Upsilon_{\boldsymbol{\mu}^\star}(T, q). \tag{96}$$

Proof. Let $m := \lceil T/2 \rceil$ and $r := T - m = \lfloor T/2 \rfloor$. We construct two deterministic instances $\boldsymbol{\mu}^L$ and $\boldsymbol{\mu}^P$. The expected reward of arm 2 in each instance is given as follows:

$$\mu_2^L(n) := \min\Big\{\tfrac{n}{T},\, 1\Big\}, \qquad \mu_2^P(n) := \min\Big\{\tfrac{n}{T},\, \tfrac{m}{T}\Big\}. \tag{97}$$

Thus arm 2 grows linearly up to level 1 in $\boldsymbol{\mu}^L$, whereas it saturates after its m-th pull in $\boldsymbol{\mu}^P$. Let $\bar{\mu}_2^L$ and $\bar{\mu}_2^P$ denote the horizon-T average rewards. A direct calculation gives

$$\bar{\mu}_2^L = \frac{T+1}{2T}, \qquad \bar{\mu}_2^P = \frac{1}{T^2}\Big[\tfrac{m(m+1)}{2} + m(T-m)\Big], \qquad \bar{\mu}_2^L - \bar{\mu}_2^P = \frac{r(r+1)}{2T^2}. \tag{98}$$

We then define

$$c := \frac{\bar{\mu}_2^L + \bar{\mu}_2^P}{2}, \qquad \Delta := \frac{\bar{\mu}_2^L - \bar{\mu}_2^P}{2} = \frac{r(r+1)}{4T^2}, \tag{99}$$

and set arm 1 to be constant in both instances, $\mu_1^L(n) = \mu_1^P(n) := c$. Since $0 < \bar{\mu}_2^P < c < \bar{\mu}_2^L \leq 1$, all rewards lie in $[0, 1]$, and every reward function is non-decreasing with non-increasing increments, hence concave (Assumption 3.1).

We first bound the regret on each instance. Let $N_{i,T}$ be the number of pulls of arm i up to time T, so that $N_{1,T} + N_{2,T} = T$. For any non-decreasing sequence, the average of its first $N_{2,T}$ terms is at most its horizon-T average, so pathwise $\sum_{n=1}^{N_{2,T}} \mu_2(n) \leq N_{2,T}\,\bar{\mu}_2$. In $\boldsymbol{\mu}^L$, the optimal value is at least that of always playing arm 2, and thus

$$\mathrm{Reg}_{\boldsymbol{\mu}^L}(\pi, T) \geq T\bar{\mu}_2^L - \mathbb{E}_L\Big[c\,N_{1,T} + \textstyle\sum_{n=1}^{N_{2,T}} \mu_2^L(n)\Big] \tag{100}$$

$$\geq T(c+\Delta) - c\,\mathbb{E}_L[N_{1,T}] - (c+\Delta)\,\mathbb{E}_L[N_{2,T}] = \Delta\,\mathbb{E}_L[N_{1,T}], \tag{101}$$

where the last equality uses $N_{1,T} + N_{2,T} = T$. Symmetrically, in $\boldsymbol{\mu}^P$ the optimal value is at least that of always playing arm 1, which yields

$$\mathrm{Reg}_{\boldsymbol{\mu}^P}(\pi, T) \geq \Delta\,\mathbb{E}_P[N_{2,T}]. \tag{102}$$

Next, we couple the two instances by running π with the same internal randomness in both. We claim that the two histories coincide pathwise through round m. Suppose they agree up to time $t-1 < m$. Then π receives the same horizon T, the same history, and the same internal randomness in both instances, so it selects the same arm at time t; a pull of arm 1 yields reward c in both instances, and the k-th pull of arm 2 satisfies $k \leq t \leq m$, so $\mu_2^L(k) = \mu_2^P(k) = k/T$. By induction the histories agree for all $t \leq m$, and in particular $N_{i,m}$ is identical across the two instances. Note that this argument applies directly to a horizon-aware policy: the known horizon T is identical in both instances and is therefore one of the coinciding inputs through round m, so knowledge of T cannot help π distinguish $\boldsymbol{\mu}^L$ from $\boldsymbol{\mu}^P$ before round m. Using $N_{i,T} \geq N_{i,m}$, we obtain

$$\mathrm{Reg}_{\boldsymbol{\mu}^L}(\pi, T) + \mathrm{Reg}_{\boldsymbol{\mu}^P}(\pi, T) \geq \Delta\big(\mathbb{E}_L[N_{1,m}] + \mathbb{E}_P[N_{2,m}]\big) \tag{103}$$

$$= \Delta\,\mathbb{E}_L[N_{1,m} + N_{2,m}] = \Delta m, \tag{104}$$

where the equality uses that $N_{i,m}$ coincides across the two instances. Consequently,

$$\max\{\mathrm{Reg}_{\boldsymbol{\mu}^L}(\pi,T),\, \mathrm{Reg}_{\boldsymbol{\mu}^P}(\pi,T)\} \geq \frac{\Delta m}{2}. \tag{105}$$

Finally, we normalize by the cumulative increment. Arm 1 is constant, so $\gamma_{\max}$ is governed by arm 2. In $\boldsymbol{\mu}^L$ we have $\gamma_2^L(\ell) = 1/T$ for $1 \leq \ell \leq T-1$, and in $\boldsymbol{\mu}^P$ we have $\gamma_2^P(\ell) = 1/T$ for $\ell \leq m-1$ and 0 afterward, so

$$\Upsilon_{\boldsymbol{\mu}^L}(T,q) = (T-1)\,T^{-q}, \qquad \Upsilon_{\boldsymbol{\mu}^P}(T,q) = (m-1)\,T^{-q} \leq (T-1)\,T^{-q}. \tag{106}$$

Let $\boldsymbol{\mu}^\star \in \{\boldsymbol{\mu}^L, \boldsymbol{\mu}^P\}$ attain the larger regret. Then

$$\frac{\mathrm{Reg}_{\boldsymbol{\mu}^\star}(\pi,T)}{\Upsilon_{\boldsymbol{\mu}^\star}(T,q)} \geq \frac{\Delta m/2}{(T-1)\,T^{-q}} = \frac{m\,r(r+1)}{8\,T^2(T-1)}\,T^q, \tag{107}$$

where the inequality uses $\Upsilon_{\boldsymbol{\mu}^\star}(T,q) \leq (T-1)T^{-q}$. Since $m \geq T/2$, $r \geq (T-1)/2$, and $r+1 \geq T/2$, we have $m\,r(r+1) \geq (T-1)T^2/8$, and therefore

$$\frac{\mathrm{Reg}_{\boldsymbol{\mu}^\star}(\pi,T)}{\Upsilon_{\boldsymbol{\mu}^\star}(T,q)} \geq \frac{T^q}{64}, \tag{108}$$

which is the claimed bound. □

H Proof of Lemmas

Lemma H.1. *For every $\epsilon \in (0,1/2)$, arm $i \in [K]$, time $\tau < t \leq T$, where $\tau = K\lceil 2/\epsilon \rceil$ sliding window $h_i = \lfloor \epsilon N_{i,t-1} \rfloor$, we have:*

$$\mathbb{P}(|\widetilde{\mu}_i(t) - \bar{\mu}_i(t)| > \beta_i(t)) \leq 2t^{-2} \tag{109}$$

Proof. Since $N_{i,t-1} \geq \lceil 2/\epsilon \rceil$ after initialization,

$$\mathbb{P}(|\widetilde{\mu}_i(t) - \bar{\mu}_i(t)| > \beta_i(t)) = \mathbb{P}(\exists n \in \{\lceil 2/\epsilon \rceil, \cdots, t-1\} : |\widetilde{\mu}_i(t) - \bar{\mu}_i(t)| > \beta_i(t), N_{i,t-1} = n) \tag{110}$$

$$\leq \sum_{n=\lceil 2/\epsilon \rceil}^{t-1} \mathbb{P}(|\widetilde{\mu}_i(t) - \bar{\mu}_i(t)| > \beta_i(t), N_{i,t-1} = n)\,. \tag{111}$$

Fixing n, we have:

$$h_i(t)(\widetilde{\mu}_i(t) - \bar{\mu}_i(t)) \tag{112}$$

$$= \sum_{l=n-h_i+1}^{n} \left((\mu_i(l) - X_i(l)) + \frac{(T-t+2)}{2} \frac{(\mu_i(l) - X_i(l)) - (\mu_i(l-h_i) - X_i(l-h_i))}{h_i} \right) \tag{113}$$

$$= \sum_{l=n-h_i+1}^{n} \left(1 + \frac{(T-t+2)}{2h_i} \right) (\mu_i(l) - X_i(l)) - \sum_{l=n-h_i+1}^{n} \frac{(T-t+2)}{2h_i} (\mu_i(l-h_i) - X_i(l-h_i))\,. \tag{114}$$

□

Since $X_i(l)$ is only drawn from $D_i(l)$, which only depends on l, all random variables $X_i(l)$ and $X_i(l-h_i)$ are independent. Also, since $\mathbb{E}[X_i(l)] = \mu_i(l)$, we can define $Z_i(l) := X_i(l) - \mu_i(l)$ and apply Azuma-Hoeffding inequality (Metelli et al., 2022). For that, we compute the sum of square

constants:

$$\sum_{l=n-h_i+1}^{n}\left[\left(1+\frac{(T-t+2)}{2h_i}\right)^2+\left(\frac{(T-t+2)}{2h_i}\right)^2\right] \tag{115}$$

$$\leq\frac{h_i}{4h_i^2}\left[(2h_i+(T-t+2))^2+(T-t+2)^2\right] \tag{116}$$

$$\leq\frac{1}{4h_i}\left[3(T-t+2)^2+8h_i^2\right]\ , \tag{117}$$

where equation 117 holds since $(a+b)^2\leq 2a^2+2b^2$. Thus, we have:

$$\mathbb{P}(|\widetilde{\mu}_i(t)-\bar{\mu}_i(t)|>\beta_i(t)) \tag{118}$$

$$=\mathbb{P}\left(\left|\sum_{l=n-h_i+1}^{n}\left[\left(1+\frac{(T-t+2)}{2h_i}\right)Z_i(l)-\frac{(T-t+2)}{2h_i}Z_i(l-h_i)\right]\right|>h_i\beta_i(t)\right) \tag{119}$$

$$\leq 2\exp\left(-\frac{(h_i\beta_i(t))^2}{2\sigma^2\left(\frac{3(T-t+2)^2+8h_i^2}{4h_i}\right)}\right) \tag{120}$$

$$\leq 2t^{-3}\ . \tag{121}$$

Substituting this result into equation 111, we have:

$$\sum_{n=\lceil 2/\epsilon\rceil}^{t-1}\mathbb{P}(|\tilde{\mu}_i(t)-\bar{\mu}_i(t)|>\beta_i(t),N_{i,t-1}=n)\leq\sum_{n=1}^{t-1}2t^{-3}\leq 2t^{-2}\ , \tag{122}$$

which concludes the proof.

Lemma H.2. *For every arm $i\in[K]$, time $t\in[T]$ and any number of pulls $2\leq n\leq t$, we have:*

$$\mu_i(t)-\mu_i(n)-\frac{T-t+2}{2}\gamma_i(n-1)\leq\max\left\{0,\left(t-n-\left\lfloor\frac{T-t+2}{2}\right\rfloor\right)\gamma_i\left(\left\lfloor\frac{T-t+2}{2}\right\rfloor+1\right)\right\} \tag{123}$$

Proof. Firstly, we consider $t\leq n+\lfloor(T-t+2)/2\rfloor$. Then:

$$\mu_i(t)\leq\mu_i(n)+\sum_{l=n}^{t-1}\gamma_i(l) \tag{124}$$

$$\leq\mu_i(n)+(t-n)\gamma_i(n-1) \tag{125}$$

$$\leq\mu_i(n)+\frac{T-t+2}{2}\gamma_i(n-1)\ , \tag{126}$$

where equation 125 holds by concavity assumption (i.e., γ_i is non-increasing). Since $t-n-\lfloor(T-t+2)/2\rfloor\leq 0$, we can easily see that our claim holds. Then, we consider the case when $t>n+\lfloor(T-t+2)/2\rfloor$. We decompose $\mu_i(t)-\mu_i(n)$ as follows:

$$\mu_i(t)-\mu_i(n)=\sum_{l=n}^{t-1}\gamma_i(l) \tag{127}$$

$$=\sum_{l=n}^{n+\lfloor(T-t+2)/2\rfloor-1}\gamma_i(l)+\sum_{l=n+\lfloor(T-t+2)/2\rfloor)}^{t-1}\gamma_i(l) \tag{128}$$

$$\leq\left\lfloor\frac{(T-t+2)}{2}\right\rfloor\gamma_i(n-1)+\left(t-n-\left\lfloor\frac{(T-t+2)}{2}\right\rfloor\right)\gamma_i\left(n+\left\lfloor\frac{(T-t+2)}{2}\right\rfloor\right) \tag{129}$$

$$\leq\frac{(T-t+2)}{2}\gamma_i(n-1)+\left(t-n-\left\lfloor\frac{(T-t+2)}{2}\right\rfloor\right)\gamma_i\left(\left\lfloor\frac{(T-t+2)}{2}\right\rfloor+1\right)\ , \tag{130}$$

where equation 129 and equation 130 hold by concavity assumption, which completes proof. □

I BASELINES

I.1 DESCRIPTION

- R-ed-UCB (Metelli et al., 2022) is a representative algorithm for the RMAB framework. It utilizes a derivative-based estimator to capture the growth rate of the reward function. By estimating this slope, the algorithm constructs an index designed to predict the instantaneous expected reward at the specific future time step.

- ET-SWGTS (Fiandri et al., 2025) is a Thompson Sampling-based algorithm adapted for RMAB framework. It addresses the non-stationary nature of reward growth by combining an initial exploration phase with a sliding-window mechanism. The algorithm first forces exploration to gather sufficient preliminary data for all arms. Subsequently, it updates the Gaussian posterior distributions using only the most recent observations within a fixed window, thereby forgetting outdated low rewards to track the current rising trend.

- TI-UCB (Xia et al., 2024) is tailored for the RMAB framework, specifically focusing on the increasing-then-converging reward trend. The core mechanism of the algorithm revolves around detecting the convergence point, where the critical moment when an arm's growth saturates and performance stabilizes. By focusing on identifying this transition to the asymptotic limit, TI-UCB aims to select the arm with the highest ultimate capacity.

- R-ed-AE (Amichay & Mansour, 2025) addresses the RMAB problem by specifically targeting environments where rewards exhibit linear drifts. The algorithm operates on a successive elimination principle, maintaining a set of active candidate arms. By estimating the growth parameters and future potential of each arm, R-ed-AE progressively discards arms that are statistically inferred to be suboptimal, thereby concentrating computational resources on identifying the most promising candidate.

- SW-UCB (Garivier & Moulines, 2011) is a modification of the standard UCB policy designed to handle non-stationary environments. It relies on a sliding window of fixed size to estimate the expected reward of each arm. By exclusively utilizing the most recent observations and discarding outdated history, the algorithm effectively forgets past behavior to adapt to dynamic changes in the reward distribution.

- SW-KL-UCB (Garivier & Cappé, 2011) enhances the sliding-window approach by replacing the standard Euclidean confidence interval with one based on KL divergence. While retaining the fixed-size window to track non-stationary changes, this modification allows for tighter confidence bounds, particularly in bounded reward settings, leading to more efficient exploration compared to the standard SW-UCB.

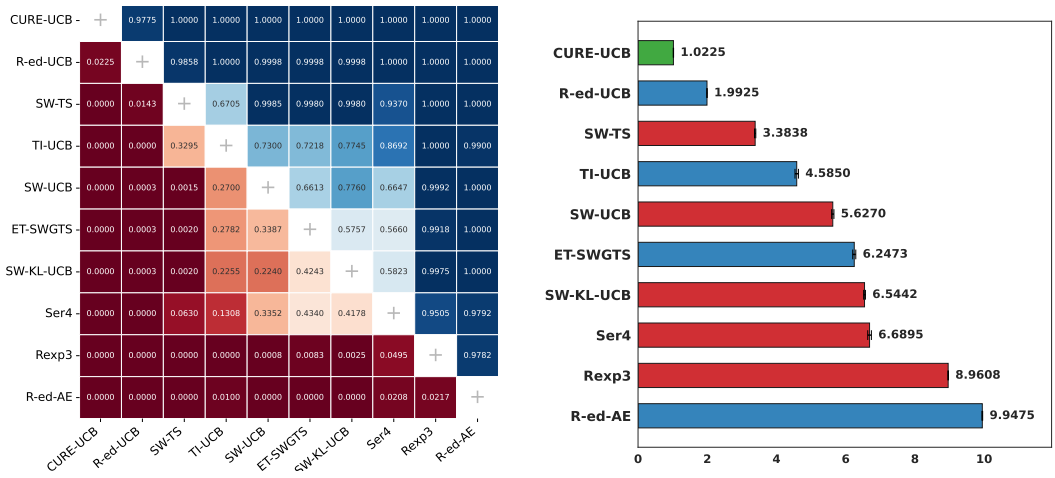
- SW-TS (Trovo et al., 2020) adapts the standard Thompson Sampling framework to non-stationary environments by incorporating a sliding-window mechanism. Instead of aggregating the entire history, the algorithm updates its Bayesian posterior distributions using exclusively the most recent observations contained within the window. This approach enables the probabilistic policy to track dynamic changes in the underlying reward distributions by effectively discarding outdated information.

- Rexp3 (Besbes et al., 2014) adapts the adversarial Exp3 algorithm to stochastic non-stationary settings by introducing a periodic restart mechanism. The algorithm divides the time horizon into fixed-size epochs and resets the Exp3 weights at the beginning of each epoch. This approach allows the policy to handle distribution changes by effectively limiting its memory to the current epoch, with the restart frequency typically tuned based on the total variation budget of the reward functions.

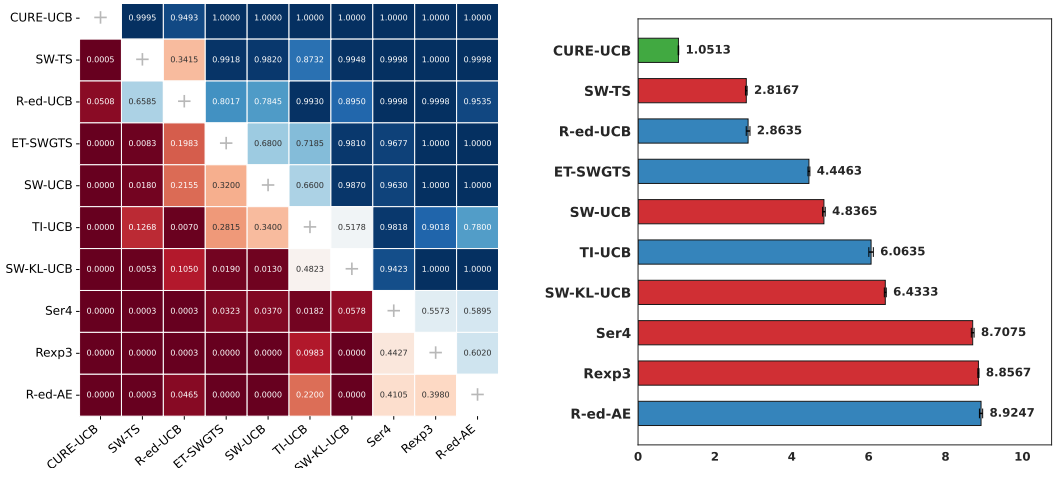
- Ser4 (Allesiardo et al., 2017) handles non-stationary environments by combining a successive elimination strategy with a periodic reset mechanism. Within each epoch, the algorithm progressively discards arms identified as suboptimal based on their estimated confidence bounds. To adapt to dynamic reward changes, Ser4 periodically resets the set of active arms, thereby reviewing previously eliminated candidates to check if they have become optimal in the new environment.

J Additional Experimental Results

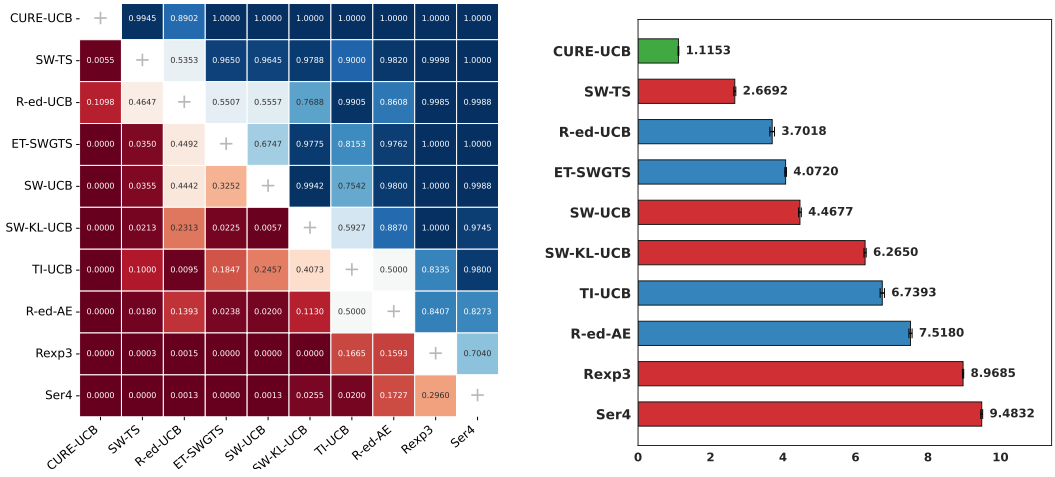
In this section, we present supplementary experimental results for a fuller picture of the algorithms' behavior. We report the pairwise win rates and average rankings over all baselines, computed across 100 independent instances with 40 random seeds each, separately for the Linear-Then-Flat (LTF) and Concave settings.



(a) $T = 20K$

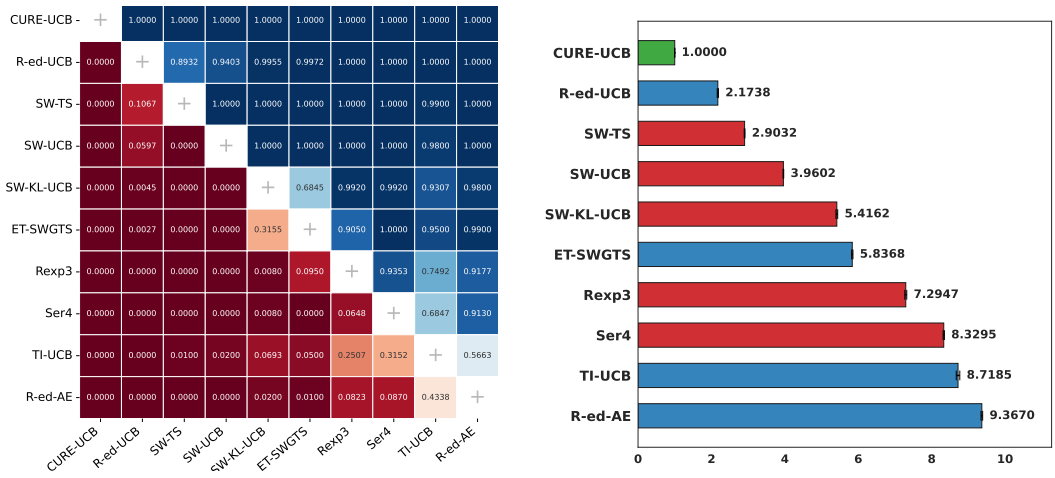


(b) $T = 60K$

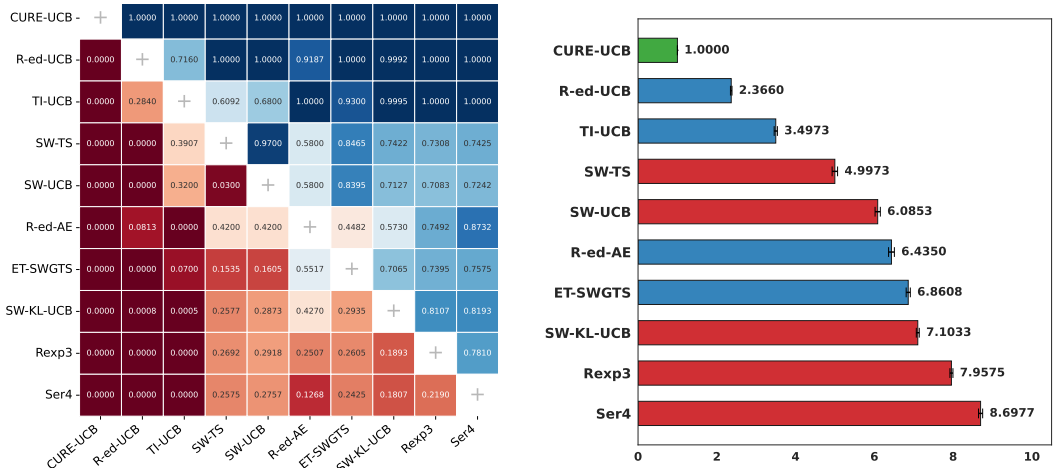


(c) $T = 100K$

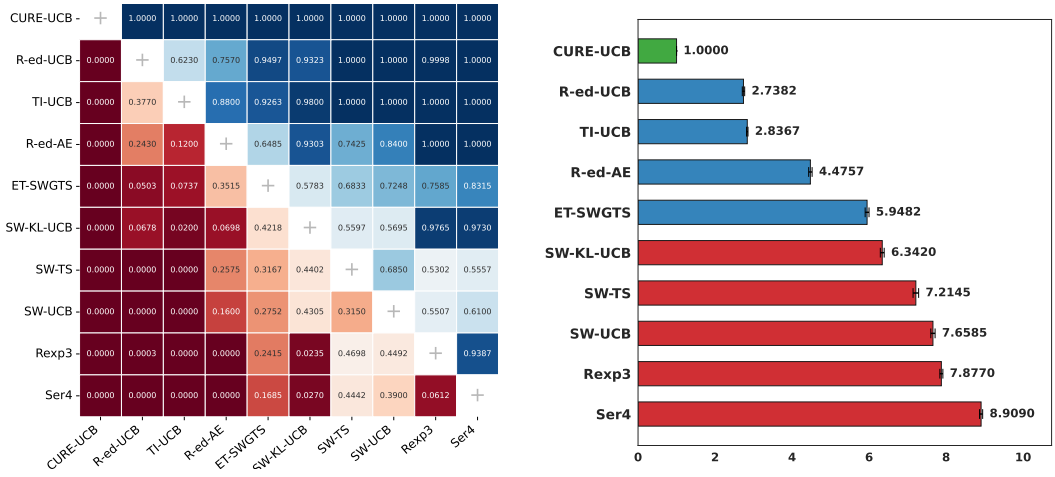
Figure 7: **Full results for the LTF setting.** (Left) Pairwise win-rate heatmap over all baselines. (Right) Average ranking over all baselines. Both are computed over 100 independent instances at $T = 20K, 60K$, and $100K$.



(a) $T = 20K$



(b) $T = 60K$



(c) $T = 100K$

Figure 8: **Full results for the Concave setting.** (Left) Pairwise win-rate heatmap over all baselines. (Right) Average ranking over all baselines. Both are computed over 100 independent instances at $T = 20K, 60K$, and $100K$.

K ADDITIONAL EXPERIMENTS FOR ROBUSTNESS

First, we examine sensitivity to the window-size parameter ε, which governs the bias-variance tradeoff in growth-rate estimation (Appendix K.1). Second, we test robustness to varying levels of reward noise (Appendix K.2). Across all studies, CURE-UCB consistently maintains its advantage over horizon-agnostic baselines.

K.1 SENSITIVITY TO WINDOW SIZE h_i

The CURE-UCB index uses a sliding window of size h_i to estimate the recent reward and growth rate. We follow the standard choice $h_i := \max\{\lfloor \epsilon N_{i,t-1} \rfloor, 1\}$ with $\varepsilon \in [0, 0.5]$, controlling the bias-variance tradeoff: larger windows reduce variance but introduce more bias from older samples. To assess sensitivity to this hyperparameter, we vary $\varepsilon \in \{1/2, 1/4, 1/8, 1/16, 1/32\}$ at $T = 50{,}000$.

Table 1 reports the mean cumulative regret in both LTF and Concave settings. Performance degrades gradually as the window shrinks, with no abrupt failure at any particular value. Larger windows yield better performance because they provide more samples for cumulative reward estimation, reducing variance.

Table 1: Mean cumulative regret of CURE-UCB under varying window-size parameter ε at $T = 50{,}000$, where $h_i = \max\{\lfloor \varepsilon N_{i,t-1} \rfloor, 1\}$.

ε	Concave	LTF
1/2	2,056	3,230
1/4	2,818	4,576
1/8	3,691	6,574
1/16	4,766	9,158
1/32	6,069	12,254

K.2 ROBUSTNESS TO NOISY REWARDS

Rising bandit algorithms estimate growth rates from differences between past samples, making the choice of noise level relevant. Since these estimates are aggregated over multiple observations within the sliding window, the effect of individual noisy samples is averaged out, allowing the algorithm to capture the underlying growth pattern despite per-sample noise. To empirically validate this, we evaluate CURE-UCB across a range of noise levels at $T = 50{,}000$.

Table 2 reports the mean cumulative regret in both LTF and Concave settings, with noise standard deviation $\sigma \in \{0.001, 0.01, 0.05, 0.1\}$. While CURE-UCB and R-ed-UCB both incorporate σ in their index design, causing their regret to vary with the noise level, CURE-UCB consistently maintains a significant advantage over all baselines across all tested noise levels, confirming its robustness under noisy reward dynamics.

Table 2: Mean cumulative regret under varying noise levels σ at $T = 50{,}000$.

	Concave				LTF			
Algorithm	σ=0.001	σ=0.01	σ=0.05	σ=0.1	σ=0.001	σ=0.01	σ=0.05	σ=0.1
CURE-UCB	**783**	**2,057**	**3,439**	**4,145**	**1,974**	**3,239**	**5,450**	**6,981**
SW-UCB	10,212	10,209	10,209	10,176	7,855	7,837	7,896	7,890
R-ed-UCB	6,667	7,886	8,059	8,530	8,761	10,268	13,805	15,585