THE PRICE OF WAITING: BUDGETED ONLINE ROUTING FOR LATENCY-TIERED LLM SERVICES

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive large language model (LLM) applications must trade off response quality, latency, and the cost of premium compute. A routing platform with a limited budget must decide which LLM to query and when to pay for faster delivery, all while learning from feedback that may be arbitrarily delayed or censored. We formulate this challenge as an online routing problem with immediate and delayed service modes under a hard premium budget. For a given LLM, both modes share the same response-quality distribution, but delayed service postpones feedback and discounts the response value. To address this, we bound the remaining value of any pending request using its current age and a known discount kernel. By combining these envelopes with completed observations, we construct confidence intervals that accommodate arbitrary reward–delay dependence without requiring known delay distributions or bounded delays. We propose `DC-LADA` for cost-penalized objectives and `EQF-LADA` for hard-budget constraints. The latter integrates optimistic allocation with quantized updates and deterministic local credit. Under stationary outcomes and deterministic costs, `EQF-LADA` strictly satisfies the budget on every realized sequence and achieves a regret bound that disentangles standard learning error, uncertainty from premium observations, and uncertainty in pending values. For a finite action space, summable pending-value envelopes guarantee sublinear regret without relying on raw-delay bounds. Trace-replay experiments on MMLU and HellaSwag demonstrate that `EQF-LADA` consistently outperforms evaluated baselines in mean discounted reward across various budget and discount settings while strictly adhering to spending constraints.

1 INTRODUCTION

For interactive applications such as coding assistants and conversational agents, LLM responses are only useful if they arrive within an actionable time frame. High-quality suggestions rapidly lose value if delayed, rendering latency a core component of overall utility Gnewuch et al. (2022); Maslych et al. (2025). Platforms must therefore navigate the trade-off between response quality, latency, and serving cost. This trade-off is governed by both model selection and system-level serving decisions, as scheduling and resource allocation directly dictate how quickly a chosen model delivers its response Fu et al. (2024); Oh et al. (2024); Wu et al. (2026).

Consider a platform that routes requests across multiple LLM services, with the option to purchase priority access for each. Standard processing conserves the premium budget but subjects requests to uncertain queuing delays; priority processing guarantees prompt delivery at an additional cost. Always requesting priority service quickly depletes the budget, whereas defaulting to standard service risks severe value degradation due to latency. Thus, the platform must dynamically determine which LLM to query and when the latency reduction justifies the premium cost Jaiswal et al. (2025). We formalize this decision-making process using a two-mode service model under a hard premium budget. We assume fresh, stationary outcomes and mode-invariant response quality, meaning that standard service incurs zero premium cost but subjects the intrinsic reward to delay-induced discounting.

Optimizing these routing decisions online is challenging because the platform lacks prior knowledge of intrinsic service quality and the joint distribution of reward and delay. Feedback from standard requests often arrives out of order, and rapidly completed requests may not constitute an unbiased sample. Consequently, policies relying solely on completed responses are vulnerable to severe

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.

estimation bias. Conversely, waiting for all responses is infeasible, particularly when delays are heavy-tailed or some requests are entirely censored. Meanwhile, the platform must dynamically adapt its allocations based on this censored feedback without violating strict budget constraints. The core challenge thus lies in learning from incomplete feedback while judiciously pacing premium expenditure to support future routing decisions.

Existing approaches address isolated facets of this problem. LLM routing methods optimize quality–cost trade-offs Feng et al. (2025); Wang et al. (2025); Li et al. (2025), whereas classical delayed bandits focus on delayed observations of undiscounted payoffs Joulani et al. (2013); Thune et al. (2019). Bandits with knapsacks (BwK) offer a framework for resource-constrained learning Badanidiyuru et al. (2018); Agrawal & Devanur (2014), but our setting introduces a fundamental coupling: a request's delay simultaneously postpones feedback revelation and degrades the realized reward. Although framing service–mode pairs as budgeted actions formalizes the allocation space, the robust estimation of pending discounted values and the budget-safe execution of dynamically updating policies remain critical open challenges.

Our approach is grounded in a key structural observation: even before a request completes, its elapsed waiting time strictly limits the maximum value it can eventually deliver. Given a known non-increasing discount kernel, we can analytically bound this remaining value; for strictly decaying kernels, this envelope naturally tightens as waiting time increases. Building upon all-submission confidence bounds Lancewicki et al. (2021), we leverage these envelopes to robustly estimate discounted service values without erroneously treating completed requests as an unbiased sample. We first instantiate this estimator in Discount-Censored and Latency-Aware Dual Action algorithm (`DC-LADA`) to solve a cost-penalized objective. To address the hard-budget constraint, we propose Epoch-Local Quantized Frontier `LADA` (`EQF-LADA`), which integrates optimistic allocation with a deterministic spending rule. By updating target allocations in quantized epochs and utilizing a local credit mechanism to schedule premium actions, `EQF-LADA` strictly confines realized expenditure within the budget even as value estimates fluctuate. Our main contributions are as follows:

(1) We formalize the problem of online LLM routing with immediate and delayed service modes, unknown outcomes, and a strict premium budget. Our objective explicitly models the degradation of response utility via a known delay-discount kernel.

(2) We develop confidence bounds that incorporate all submitted requests, including those that are still pending. This approach naturally accommodates reward–delay dependence, out-of-order feedback, and infinite delays (censorship), culminating in a rigorous regret guarantee for `DC-LADA` under a cost-penalized objective.

(3) We prove that `EQF-LADA` strictly satisfies the budget constraint on every realized sequence. It achieves a regret of $\widetilde{O}(\sqrt{KT} + \sqrt{KM_B} + K\Psi_g^{\mathrm{cert}}(T))$, where M_B is the maximum affordable immediate selections and $\Psi_g^{\mathrm{cert}}(T)$ quantifies the cumulative uncertainty from pending values. Crucially, summable value envelopes guarantee sublinear regret for fixed K without requiring bounded raw delays. We complement this with a zero-budget lower bound that isolates the inherent cost of delayed revelation, matching the exponential-envelope scale up to constants.

(4) We evaluate calibrated practical variants using real-world traces from MMLU and HellaSwag. Empirical results demonstrate that `EQF-LADA` consistently achieves the highest mean discounted reward among evaluated baselines across various budget and discount scenarios, while strictly honoring the spending constraint. These replay experiments validate the practical efficacy of our algorithm design independently of theoretical parameter prescriptions.

Related Work *Delayed and partial feedback.* Classical delayed bandits investigate how delayed observations impact learning when the underlying payoff remains constant Joulani et al. (2013); Thune et al. (2019); Masoudian et al. (2024). Reward-dependent models accommodate informative censorship: it constructs all-submission confidence bounds under unrestricted delays Lancewicki et al. (2021), whereas other approaches rely on specific parametric payoff–delay relationships Tang et al. (2024); Zhang et al. (2025). Alternatively, conversion and progressive-feedback models leverage known delay distributions or intermediate signals Vernade et al. (2017); Zhang et al. (2026). These methods address distinct mechanisms of incomplete information. In contrast, we optimize a discounted target, where the same delay simultaneously postpones feedback revelation and degrades the intrinsic value. Building upon all-submission estimation, we bound pending values using solely a

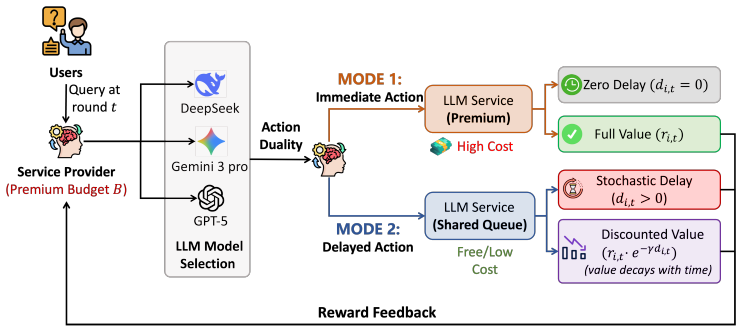

Figure 1: Latency-tiered LLM routing under a hard premium budget.

known discount kernel (obviating the need for prior delay distributions or restrictive reward–delay structural assumptions) and integrate this estimation challenge with a hard premium budget constraint. While prior work has explored global-time discounted rewards Mersereau et al. (2009) and payoffs conditioned on the time elapsed since an arm's last selection Cella & Cesa-Bianchi (2020), our relevant clock is the per-request service delay. Under a fresh stationary model, this delay uniquely determines both the final realized value and the observation time.

Budget-constrained learning and allocation. Bandits with Knapsacks (BwK) provide a foundation for learning under resource constraints, encompassing stochastic consumption and contextual variants Badanidiyuru et al. (2018); Agrawal & Devanur (2014; 2016). Deterministic-cost policies and UCB-Simplex bridge optimistic allocation with knapsack optimization and mixture execution Tran-Thanh et al. (2012); Flajolet & Jaillet (2015); SUAK enforces strict cumulative cost constraints at every round Elumar et al. (2025). Other related formulations constrain rewards relative to a baseline policy Wu et al. (2016), bound expected per-round costs Pacchiano et al. (2021), or allow for resource replenishment Bernasconi et al. (2024). Although these works establish robust tools for resource control, their allocation mechanisms do not prescribe how to estimate the pending discounted rewards unique to our setting. We consider known deterministic costs alongside a zero-premium standard action that retains residual value. By coupling age-dependent confidence bounds with quantized target allocations and a local credit system, we strictly enforce the budget on every realized sequence.

LLM routing and serving. Model and prompt routing frameworks optimize the trade-off between quality and cost across diverse LLM services Feng et al. (2025); Dai et al. (2024); Wang et al. (2025); Li et al. (2025); Atalar (2026). StageRoute extends this to streaming deployments and dynamic routing mixtures under expected average-cost constraints Li & Li (2026), while serving schedulers like FastServe mitigate latency via low-level execution control Wu et al. (2026). Our work studies a fixed set of services equipped with immediate and delayed modes, concurrently learning their values and strictly enforcing a realized premium budget. Complex stateful deployments and detailed queueing dynamics remain outside the scope of our stationary model. A comprehensive comparison is provided in Appendix E.

2 Problem Formulation

As illustrated in Fig. 1, a learner submits one request per round over $T \geq 2$ rounds to one of K LLM services, called arms and indexed by $i \in [K] := \{1, \ldots, K\}$. Each arm offers immediate service (`Imm`) with prompt feedback at a known deterministic premium $c_i > 0$, and delayed service (`Del`) with a random delay and no premium charge. The learner has a total premium budget $B \geq 0$. At round $t \in \{1, \ldots, T\}$, it selects an action $A_t = (i, o)$, where i identifies the arm and o the service mode. The action set is $\mathcal{S} = [K] \times \{\texttt{Imm}, \texttt{Del}\}$.

Reward and delay model. For a request to arm i, let $R_i \in [0, 1]$ denote its raw reward and $D_i \in [0, \infty]$ be its delay under delayed service, measured in round units. Infinite delay means that feedback never arrives. The unknown joint distribution of (R_i, D_i) is denoted by P_i. For a request to arm i in round t, write its outcome as $(R_{i,t}, D_{i,t})$; the notation (R_i, D_i) denotes a generic draw.

Assumption 1 (Fresh, stationary, and mode-invariant outcomes)**.** Conditional on the observed history and the selected arm i, each request generates a fresh pair $(R_i, D_i) \sim \mathsf{P}_i$. This distribution is fixed across rounds, independent of prior request outcomes, and unaffected by the selected mode. Within a request, R_i and D_i may be arbitrarily dependent.

Mode invariance gives the two modes a common raw-reward distribution: immediate service changes when feedback arrives, while preserving response quality. For LLM services, this describes priority scheduling or reserved capacity with the same model, context, and generation parameters. Stationarity excludes load-dependent outcomes and congestion effects caused by the routing policy. Costs measure premium charges only, so delayed service can consume computational resources while having zero cost in the budget model.

Action values. Waiting reduces response value through a known, non-increasing discount kernel $g : [0, \infty] \to [0, 1]$, with $g(\infty) := \lim_{d\to\infty} g(d)$. Immediate service yields the reward R_i, whereas delayed service yields the discounted reward $R_i g(D_i)$. With expectations taken under P_i, define

$$Y_i = R_i g(D_i), \mu_i = \mathbb{E}[R_i], q_i = \mathbb{E}[Y_i], \ell_i = \mu_i - q_i = \mathbb{E}[R_i(1 - g(D_i))]. \tag{1}$$

Thus, μ_i and q_i are the expected values of the immediate and delayed service modes, respectively. The latency loss ℓ_i captures the expected value uplift achieved by serving arm i immediately rather than incurring the delay discount. Importantly, because the reward and delay may be correlated, q_i does not necessarily equal $\mu_i \mathbb{E}[g(D_i)]$. The expected value $V_{i,o}$ and the deterministic premium cost $C_{i,o}$ for each compound action are given by

$$V_{i,\texttt{Imm}} = \mu_i, C_{i,\texttt{Imm}} = c_i, V_{i,\texttt{Del}} = q_i, C_{i,\texttt{Del}} = 0. \tag{2}$$

For any action $a = (i, o)$, let $V_a := V_{i,o}$ and $C_a := C_{i,o}$, so that C_{A_t} represents the cost of the action selected in round t. We denote the minimum and maximum premiums as $c_{\min} := \min_i c_i$ and $c_{\max} := \max_i c_i$, respectively. While these definitions and our structural results apply to any known non-increasing kernel, we primarily focus on exponential discounting, defined as $g_\gamma(d) = e^{-\gamma d}$ for finite delay d and discount rate $\gamma \geq 0$. At infinite delay, the limit yields $g_\gamma(\infty) = 0$ for $\gamma > 0$ and $g_0(\infty) = 1$. Thus, the $\gamma = 0$ case corresponds to an undiscounted setting, where the reward is fully preserved even if feedback never arrives.

Decision and feedback timeline. For a request submitted to arm i in round s, we abbreviate $(R_{i,s}, D_{i,s})$ as (R_s, D_s). Feedback processed at the end of a round becomes available for the subsequent decision. Immediate service reveals R_s by the end of round s, but the counterfactual delay under the delayed mode remains unobserved. Conversely, under delayed service, finite-delay feedback (R_s, D_s) arrives at the end of round $s + A(D_s)$. Here, $A(d) = \lceil d \rceil$ defines the integer feedback offset for a continuous delay d, with $A(\infty) = \infty$. Consequently, feedback is available prior to decision t if and only if $s + A(D_s) \leq t - 1$. The learner merely observes that a request is pending until its completion. Crucially, the realized value is governed by the exact continuous delay D_s, whereas the offset $A(D_s)$ strictly dictates the feedback timing.

Information from pending requests. If a request submitted in round $s < t$ remains pending prior to decision t, its continuous delay must satisfy $D_s > t - s - 1$. Let $u_{s,t}$ denote the induced upper bound on its discount factor. Because the raw reward is bounded as $R_s \in [0, 1]$, we obtain the following deterministic envelope for the pending discounted value:

$$u_{s,t} = \sup_{d > t-s-1} g(d), \qquad 0 \leq R_s g(D_s) \leq u_{s,t}. \tag{3}$$

This computable envelope enables the algorithm to learn from censored observations before feedback completion. For the exponential discount kernel, this supremum simplifies to $u_{s,t} = e^{-\gamma(t-s-1)}$. Further details on envelope constructions and kernel-specific properties are provided in Appendix B.8.

Objective and benchmark. Let $X_t(A_t)$ denote the eventual realized value of action A_t, where $X_t(A_t) = R_{i,t}$ in the immediate mode and $X_t(A_t) = R_{i,t} g(D_{i,t})$ in the delayed mode. The cumulative objective incorporates the values of all T submissions, including requests whose feedback arrives after the horizon T or is entirely censored. A valid policy π must be non-anticipating, i.e., making decisions based only on the available history, and must strictly satisfy the pathwise budget constraint $\sum_{t=1}^{T} C_{A_t} \leq B$. Let $\mathbb{E}_\pi$ denote the expectation over both the request outcomes and the internal randomization of π. An adaptive oracle, equipped with full knowledge of distributions

$\{\mathsf{P}_i\}_{i\in[K]}$, operates under identical feedback and budgets. Its optimal expected total reward is:

$$\mathrm{OPT}_B^{\mathrm{ad}}(T) = \sup_{\substack{\pi \text{ is non-anticipating:} \\ \sum_{t=1}^T C_{A_t} \le B \text{ a.s.}}} \mathbb{E}_\pi\left[\sum_{t=1}^T X_t(A_t)\right]. \tag{4}$$

For any feasible learning policy π, the expected budgeted regret is then given by:

$$R_B^\pi(T) = \mathrm{OPT}_B^{\mathrm{ad}}(T) - \mathbb{E}_\pi\left[\sum_{t=1}^T X_t(A_t)\right]. \tag{5}$$

3 Structural Properties of the Budgeted Oracle

With known action values, the oracle reduces to a fixed allocation of requests. Let n_a be the number assigned to action $a \in \mathcal{S}$, and define the integer count oracle

$$\mathrm{OPT}_B^{\mathrm{int}}(T) = \max_{n\in\mathbb{Z}_+^{|\mathcal{S}|}} \left\{\sum_{a\in\mathcal{S}} n_a V_a \sum_a n_a = T,\ \sum_a n_a C_a \le B\right\}. \tag{6}$$

All delayed actions have zero cost, so only the largest delayed value $q_* = \max_i q_i$ is needed by the oracle. The resulting cost–value points are

$$P_0 = (c_0, v_0) = (0, q_*), \qquad P_i = (c_i, v_i) = (c_i, \mu_i), \quad i \in [K]. \tag{7}$$

For a per-round budget $b \ge 0$, their fractional allocation frontier is

$$F(b) = \max_{x\in\Delta_{K+1}} \left\{\sum_{j=0}^K x_j v_j : \sum_{j=0}^K x_j c_j \le b\right\}, \qquad \Delta_{K+1} = \{x \ge 0 : \textstyle\sum_j x_j = 1\}. \tag{8}$$

Write $\mathrm{LP}_B(T) := TF(B/T)$ for the corresponding fractional count relaxation. At each duplicated cost, retain the highest-value point; remove dominated points and collinear interior points, and choose the cheapest maximizer on a flat face. Appendix A gives the precise construction.

Theorem 1 (Structural Budget Frontier)**.** *Under Assumption 1 and the expected eventual-value objective: (1) the known-distribution adaptive hard-budget oracle equals the integer count oracle in Eq. (6); (2) the zero-cost Delayed actions reduce to the single point $(0, q_*)$, so the fractional oracle is the upper concave frontier of the $K+1$ points in Eq. (7); (3) $\mathrm{LP}_B(T) = TF(B/T)$, an optimal LP allocation uses at most two adjacent efficient vertices, and any dual optimizer is a supporting shadow price; and (4) $0 \le \mathrm{LP}_B(T) - \mathrm{OPT}_B^{\mathrm{ad}}(T) \le 1$.*

The proof is given in Appendix A.3. The theorem suggests constructing an optimistic frontier and allocating requests across at most two adjacent efficient vertices. The next section shows how to execute these allocations under a hard budget as value estimates change.

4 Algorithm Design

We first introduce `DC-LADA` for a cost-penalized objective, then extend its confidence estimates to the hard-budget setting with `EQF-LADA`. Before decision t, let $n_i(t)$ denote the total number of delayed submissions to arm i. We partition the indices of these submissions into a set of arrived (completed) requests $\mathcal{A}_i(t)$ and a set of pending requests $\mathcal{P}_i(t)$. For a submission j made in round $s_{i,j}$, its pending-value upper bound is $u_{i,j}(t) = u_{s_{i,j},t}$. Let $Y_{i,j} = R_{i,j} g(D_{i,j})$ denote the realized discounted reward for each completed request. For $n_i(t) > 0$, we define the empirical mean of completed requests and the all-submission optimistic estimate as:

$$\widehat{q_i^-}(t) = \frac{1}{n_i(t)} \sum_{j\in\mathcal{A}_i(t)} Y_{i,j}, \quad \widehat{q_i^+}(t) = \frac{1}{n_i(t)} \left(\sum_{j\in\mathcal{A}_i(t)} Y_{i,j} + \sum_{j\in\mathcal{P}_i(t)} u_{i,j}(t)\right). \tag{9}$$

Crucially, these estimators incorporate all submissions by imputing missing values with zero (for the lower bound) or their certified upper bounds (for the upper bound). At each step, the algorithm incrementally updates sums for newly arrived feedback and tightens the envelopes for pending requests based on their elapsed delays. Arms without delayed submissions receive a default upper confidence bound (UCB) of one. Arm and action ties are broken deterministically by index and the fixed hierarchy $(1, \texttt{Imm}), (1, \texttt{Del}), \ldots, (K, \texttt{Del})$.

Algorithm 1 Discount-Censored LADA (DC−LADA)

1: **Input:** $K, T, g, \{c_i\}_{i=1}^K$
2: **for** $t = 1, \ldots, T$ **do**
3: Process all feedback available before decision t
4: Form the Immediate and all-submission Delayed indices in Eq. (11)–equation 13
5: Select an action with maximal index using fixed tie breaking
6: Register the selected submission; at round end store Immediate R or any delivered Delayed (R, D) for use from the next decision

4.1 Discount-Censored LADA

DC−LADA (Algorithm 1) optimizes a cost-penalized objective, aiming to maximize the net reward minus the premium cost. The expected net value for each mode is defined as:

$$\widetilde{V}_{i,\texttt{Imm}} = \mu_i - c_i, \qquad \widetilde{V}_{i,\texttt{Del}} = q_i. \tag{10}$$

Unlike the hard-budget formulation, this objective inherently penalizes spending. Letting $\widetilde{V}^* = \max_a \widetilde{V}_a$ and $\Delta_a = \widetilde{V}^* - \widetilde{V}_a$, the cost-penalized regret is $R_u(T) = \mathbb{E}[\sum_{t=1}^T \Delta_{A_t}]$. We configure the confidence radius using $L_T = \log(4KT^3)$ and $\beta_n^T = \sqrt{L_T/(2n)}$ for $n \geq 1$. Let $m_i(t)$ denote the number of immediate observations for arm i prior to decision t, with corresponding empirical mean $\widehat{\mu}_i^{\mathrm{I}}(t)$. For streams with at least one observation, we construct the immediate and delayed UCBs as:

$$\overline{\mu}_i^{\mathrm{I}}(t) = \min\{1, \widehat{\mu}_i^{\mathrm{I}}(t) + \beta_{m_i(t)}^T\}, \quad m_i(t) > 0, \tag{11}$$

$$\overline{q}_i^{\mathrm{DC}}(t) = \min\{1, \widehat{q}_i^{+}(t) + \beta_{n_i(t)}^T\}, \quad n_i(t) > 0, \tag{12}$$

where $\widehat{q}_i^{+}(t)$ is defined in Eq. (9). If a stream has zero observations, its UCB defaults to one. By subtracting the immediate premium from the immediate UCB, we define the action indices:

$$U_i^{\texttt{Imm}}(t) = \overline{\mu}_i^{\mathrm{I}}(t) - c_i, \qquad U_i^{\texttt{Del}}(t) = \overline{q}_i^{\mathrm{DC}}(t). \tag{13}$$

DC−LADA greedily selects the action with the highest index in each round. Note that the immediate estimates rely exclusively on immediate feedback and we discard delayed feedback for these estimates to avoid the inherent selection bias caused by arrival-censored samples.

4.2 Epoch-Local Quantized Frontier LADA

EQF−LADA (Algorithm 2) integrates optimistic value estimation with a deterministic scheduling mechanism to strictly enforce the hard budget constraint. The algorithmic design comprises two core phases: first, stabilizing the estimated frontier via quantization; second, executing the resulting fractional allocations using a local credit system. For boundary cases, the algorithm defaults to greedy selection: if $B = 0$, it selects the action with the largest delayed UCB; if $B \geq Tc_{\max}$, it selects the one with the largest immediate UCB. In the following, we assume the non-trivial regime $0 < B < Tc_{\max}$ and define the per-round budget rate as $b = B/T$.

Optimistic frontier and epochs. We construct the frontier using the raw-value bounds $U_i^{\mathrm{I}}(t) = \overline{\mu}_i^{\mathrm{I}}(t)$ and $U_i^{\mathrm{D}}(t) = \overline{q}_i^{\mathrm{DC}}(t)$ without subtracting costs. For each action a, we maintain a running minimum, initialized to one: $\overline{U}_a(t) = \min_{1 \leq s \leq t} U_a(s)$. Using a quantization parameter $\varepsilon \in (0, 1]$, we define:

$$Q_\varepsilon(x) = \min\left\{1, \varepsilon \left\lceil \frac{x}{\varepsilon} \right\rceil\right\}, \qquad \widetilde{U}_a(t) = Q_\varepsilon(\overline{U}_a(t)). \tag{14}$$

This quantization strategically limits the frequency of scheduler restarts, and we use $\varepsilon = \min\{1, \sqrt{(K+1)/T}\}$ as the default choice. Let $\widetilde{U}_0(t) = \max_i \widetilde{U}_i^{\mathrm{D}}(t)$ be the maximum delayed estimate, and let $i_0(t)$ be its realizing arm. We define the corresponding cost-value points as $(c_0, \widetilde{v}_0) = (0, \widetilde{U}_0)$ and $\widetilde{v}_i = \widetilde{U}_i^{\mathrm{I}}$ for $i \in [K]$. The optimistic fractional frontier is then given by:

$$\widetilde{F}_t(b) = \max_{x \in \Delta_{K+1} : \sum_j x_j c_j \leq b} \sum_j x_j \widetilde{v}_j(t). \tag{15}$$

At each cost, retain the highest-value point with deterministic tie-breaking. Remove dominated and collinear interior points, and choose the cheapest maximizer on flat faces (Appendix B.5). An *epoch*

Algorithm 2 Epoch-Local Quantized Frontier LADA (EQF−LADA)

Input: $K, T, B, g, \{c_i\}_{i=1}^K, \varepsilon \in (0, 1]$
Initialize empty histories;
if $B = 0$ **then**
  Each round, update UCBs and select $A_t = (\arg\max_i \overline{q}_i^{\mathrm{DC}}(t), \texttt{Del})$
else if $B \geq Tc_{\max}$ **then**
  Each round, update UCBs and select $A_t = (\arg\max_i \overline{\mu}_i^{\mathrm{I}}(t), \texttt{Imm})$
else
  Set $b \leftarrow B/T$, $B_{\mathrm{rem}} \leftarrow B$, $\overline{U}_a \leftarrow 1$ for all a; no active epoch
  for $t = 1, \ldots, T$ **do**
    Process available feedback; update raw UCBs, running minima, and quantized bounds via Eq. (14); Set $i_0(t) \leftarrow \arg\max_i \widetilde{U}_i^{\mathrm{D}}(t)$ and $z_t \leftarrow (\widetilde{U}_0(t), \widetilde{U}_1^{\mathrm{I}}(t), \ldots, \widetilde{U}_K^{\mathrm{I}}(t))$
    if no active epoch, or $z_t \neq z_{t-1}$ **then**
      Start epoch e; construct and classify its frontier by Lemma 8; set $r \leftarrow 0$, $D_{e,0} \leftarrow 0$
    Set $r \leftarrow r + 1$; select the S/V vertex or an E endpoint by Eq. (16)
    Map vertex 0 to $A_t = (i_0(t), \texttt{Del})$, or vertex $i > 0$ to $A_t = (i, \texttt{Imm})$
    For type E, update $D_{e,r} \leftarrow D_{e,r-1} + b - C_{A_t}$; Set $B_{\mathrm{rem}} \leftarrow B_{\mathrm{rem}} - C_{A_t}$

is a maximal interval over which $(\widetilde{U}_0, \widetilde{U}_1^{\mathrm{I}}, \ldots, \widetilde{U}_K^{\mathrm{I}})$ is constant. Any coordinate change starts a new epoch, even if the frontier is unchanged. Each round, the zero-cost point maps to $(i_0(t), \texttt{Del})$ and positive-cost points to their immediate arms. Changes in $i_0(t)$ alone do not restart epochs; estimates always use the corresponding arm's own counts.

Budget-feasible scheduling. To execute the fractional allocation, we map it to discrete actions. According to Lemma 8, the frontier optimum is either a single vertex with cost $c \leq b$ (type S or V), which the algorithm executes directly; or it lies on an edge between adjacent vertices L_e and R_e with $c_{L_e} < b < c_{R_e}$ (type E). At the start of each epoch e, we initialize a virtual credit $D_{e,0} = 0$. For the r-th local decision, the algorithm selects the higher-cost vertex R_e if and only if the accumulated credit covers the excess cost:

$$R_e \quad \Longleftrightarrow \quad D_{e,r-1} \geq c_{R_e} - b, \quad \text{with update} \quad D_{e,r} = D_{e,r-1} + b - C_{A_t}, \tag{16}$$

and selects L_e otherwise. This virtual credit tracks the unused budget allowance within the epoch, safely scheduling the premium endpoint. While virtual credit resets across epochs, the physical budget is debited globally. This deterministic tracking guarantees the prefix budget constraint $\sum_{s=1}^t C_{A_s} \leq tb \leq B$ for all t, ensuring strict online feasibility without ex-post rejection steps. The complete procedure is summarized in Algorithm 2, with formal feasibility proofs, fractional rounding loss bounds, and complexity details provided in Appendix B.

5 Theoretical Analysis

Confidence under delayed feedback. For a given confidence level $\delta \in (0, 1)$, we define the radius $\beta_n(\delta) := \sqrt{\log(2KT/\delta)/(2n)}$. By expanding the lower and upper estimates by this radius and projecting them onto $[0, 1]$, we construct the robust confidence bounds $\mathrm{LCB}_i^q(t)$ and $\mathrm{UCB}_i^q(t)$. For arms with no observations, the bounds default to $[0, 1]$.

Theorem 2 (Discount-Censored Confidence)**.** *Under Assumption 1, and any known nonincreasing $g : [0, \infty] \to [0, 1]$, with probability at least $1 - \delta$, simultaneously for all arms and decision times, $q_i \in [\mathrm{LCB}_i^q(t), \mathrm{UCB}_i^q(t)]$. Moreover, whenever $n_i(t) > 0$,*

$$\mathrm{UCB}_i^q(t) - q_i \leq 2\beta_{n_i(t)}(\delta) + \frac{1}{n_i(t)} \sum_{j \in \mathcal{P}_i(t)} u_{i,j}(t). \tag{17}$$

The proof is given in Appendix B.3.

Effective-delay regret. Define $\Psi_g^{\mathrm{cert}}(H) := \sup_s \sum_{k=1}^H u_{s,s+k}$, the largest cumulative envelope over H decision ages. This quantity controls how much a pending request contributes to uncertainty. The next theorem bounds regret under the cost-penalized objective in Eq. (10).

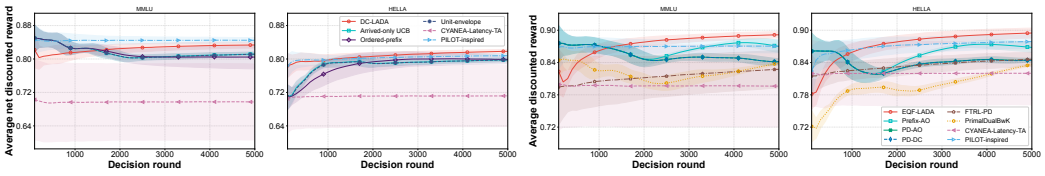

Figure 2: **Learning with delayed feedback.** Left: average net discounted reward for the DC task. Right: average discounted reward for the EQF task under the prefix budget $b = 0.12$.

Theorem 3 (Effective-Delay Regret)**.** *Under Assumption 1 and the exact timeline in Section 2, Discount-Censored* LADA *satisfies*

$$R_u(T) \le \min\left\{ GT, 2K + 4\sqrt{KTL_T} + K\Psi_g^{\mathrm{cert}}(T)(1 + \log T) + \frac{G}{T} \right\}, \tag{18}$$

where $G = 1 + \max_i c_i$ *and* $L_T = \log(4KT^3)$.

The proof is given in Appendix B.4. For fixed K, the delay term is sublinear when $\Psi_g^{\mathrm{cert}}(T) \log T = o(T)$. Summable envelopes give logarithmic delay overhead even with unbounded raw delays.

Hard-budget regret. For EQF−LADA, local credit ensures budget feasibility, while quantization limits the number of epochs. Each binding epoch loses less than one unit relative to its fractional allocation. Combining these losses with confidence widths and quantization error gives the following bound, where M_B is the maximum number of affordable immediate actions.

Theorem 4 (Budget-Uniform Hard-Feasible Regret)**.** *Under Assumption 1, known deterministic costs, and any known nonincreasing kernel* g *under the certified clock, Algorithm 2 with* $\varepsilon_T = \min\{1, \sqrt{(K+1)/T}\}$ *is pathwise budget-feasible for every* $B \ge 0$ *and satisfies*

$$R_B(T) = \widetilde{O}\left(\sqrt{KT} + \sqrt{KM_B} + K\Psi_g^{\mathrm{cert}}(T) \right), \tag{19}$$

where $M_B = \min\{T, \lfloor B/c_{\min} \rfloor\}$ *and* $\widetilde{O}$ *hides logarithmic factors in* K *and* T.

The explicit bounds and proof are given in Appendix B.5.

Lower Bounds. We isolate the regret caused by delayed information when $B = 0$. Under the ceiling clock, let $u_h := \sup_{d>h-1} g(d)$ and $\Phi_g(H) := \max_{1\le h\le H} h u_h$. For exponential discounting, let $\Psi_{\exp}^{\mathrm{real}}(H) := \sum_{h=1}^{H} e^{-\gamma(h-1)}$.

Theorem 5 (Nonzero-Delay Revelation Lower Bound)**.** *Fix a known nonincreasing kernel* g *and the real-valued ceiling clock. For every hard-feasible policy, every horizon* T*, and every* $1 \le H \le T$*, there is a two-arm instance with* $B = 0$*, a common deterministic nonzero delay for both arms, and mode-invariant outcomes such that* $R_0^\pi(T) \ge \frac{1}{8}\Phi_g(H)$*. For exponential discounting, the same construction gives* $\sup_\theta R_0^\pi(T;\theta) \ge \Psi_{\exp}^{\mathrm{real}}(H)/32$.

The proof is given in Appendix B.7. Appendix C gives the kernel comparisons and a separate lower bound for arm uncertainty.

6 Experiments

Experimental Setup We replay paired correctness and latency records from six LLM services on MMLU Hendrycks et al. (2020) and HellaSwag (HELLA) Zellers et al. (2019). Questions are split into 20% development and 80% evaluation sets. Each run samples evaluation questions with replacement, using the same sequence for all methods. Immediate mode returns the recorded correctness at a normalized premium $c_i \in [0.1, 0.3]$, while Delayed mode is free and discounts correctness by $e^{-\gamma D}$. Learners observe only selected-action feedback after arrival. These modes simulate service tiers from historical traces. We use $T = 5{,}000$ submissions and 100 paired seeds. The default setting is $\gamma = 0.02$ and, for EQF, $b = B/T = 0.12$; the sweeps vary b from 0.05 to 0.30 and γ from 0.005 to 0.08. DC measures average discounted reward minus premium cost, whereas EQF measures average

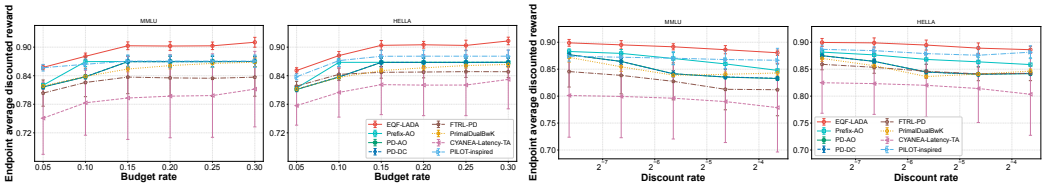

Figure 3: **Budget and discount sensitivity.** Left: six budget rates at $\gamma = 0.02$. Right: five discount rates at $b = 0.12$.

discounted reward subject to cumulative spending $S_t \leq bt$ at every round. Scores include eventual outcomes of all submitted requests. Figures report means and one sample standard deviation across seeds. We evaluate `DC-LADA` and `EQF-LADA`, with hyperparameters fixed across each sweep; the theory-default confidence and regret bounds are not claimed for these settings.

DC baselines compare arrived-only, ordered-prefix, and unit-envelope estimation. EQF baselines compare feasibility-masked UCB (Prefix-AO), primal-dual allocation with arrived-only or discount-censored estimates (PD-AO and PD-DC), FTRL-PD, and PrimalDualBwK Badanidiyuru et al. (2018). Both tasks also include contextual adaptations of CYANEA Li et al. (2025) and PILOT Panda et al. (2025), using the same objectives and spending constraints. Full replay, calibration, and baseline details appear in Appendix F.

Learning from delayed feedback. In Figure 2, `DC-LADA` achieves higher endpoint net reward than all three non-contextual estimators on both datasets. This is consistent with a benefit from using the shrinking value of pending requests to guide exploration. Its comparison with PILOT-inspired varies by dataset: `DC-LADA` performs better on HellaSwag but trails on MMLU, suggesting that contextual prediction can also be valuable. `EQF-LADA` achieves the highest endpoint mean among the evaluated baselines on both datasets. PD-DC and PD-AO perform similarly, while `EQF-LADA` exceeds both, suggesting that the allocation rule matters alongside value estimation. These results support combining discount-censored estimates with budget-aware allocation, although the comparisons do not isolate individual component effects.

Value of additional premium capacity. Figure 3 (left) shows that `EQF-LADA` maintains the highest mean reward across all 12 dataset–budget settings, with a near tie under the tightest MMLU budget. Reward largely levels off between $b = 0.15$ and 0.25, then rises at $b = 0.30$. This pattern suggests that additional premium capacity has uneven marginal value: its benefit depends on which service allocations it enables. All methods record zero prefix-budget violations, so `EQF-LADA`'s advantage reflects higher reward under the same spending constraint.

Sensitivity to the latency penalty. `EQF-LADA` also retains the highest mean reward across all 10 dataset–discount settings in Figure 3 (right). As the latency penalty increases, its reward declines less than that of Prefix-AO on both datasets. This is consistent with accounting for lost Delayed value when allocating scarce Immediate access. Together, the sweeps support `EQF-LADA`'s budgeted allocation design across the tested conditions. These are descriptive comparisons on fixed trace populations; detailed scores and relative gains are provided in Appendix F.5.

7 Conclusion

We investigate online selection among latency-tiered large language model (LLM) services under a limited premium budget. By bounding the remaining value of pending requests, we facilitate learning from incomplete feedback, even when rewards and delays are dependent. Building on these estimates, `EQF-LADA` combines optimistic allocation with deterministic budget control to ensure feasibility for every realized sequence. Trace-replay results demonstrate the practical value of this design across various budget and discount settings. Our guarantees assume fresh stationary outcomes, quality-invariant modes, and a single known deterministic premium resource, while leaving contextual routing, changing service conditions, and stochastic or multiple resources for future research. The lower bound characterizes delayed revelation without establishing full minimax optimality. Trace-replay experiments on MMLU and HellaSwag demonstrate that `EQF-LADA` consistently outperforms evaluated baselines.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization and drafting, language revision, LaTeX editing, clarification of algorithm descriptions and proof explanations, and interpretation of reported experimental results. AI also assisted with executing experimental code, and the authors manually checked this AI-assisted experimental work. The authors take responsibility for the final text, technical claims, and research artifacts.

ETHICS STATEMENT

This work studies budgeted LLM routing through theoretical analysis and historical benchmark replay. Allocating premium access may affect service quality and access across users. Our objective captures discounted utility and premium spending, but does not enforce fairness or task-specific safety. Practical deployment would require evaluating these effects and adding appropriate safeguards for the intended application.

REPRODUCIBILITY STATEMENT

Section 2 specifies the model and assumptions, and Section 4 provides algorithm definitions and pseudocode. Supporting derivations and proofs are given in the appendices referenced by the theoretical results. Appendix F documents the trace processing, data split, feedback protocol, hyperparameters, baseline adaptations, random seeds, and evaluation metrics, together with detailed numerical results. The code will be submitted as supplementary material.

REFERENCES

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014.

Baran Atalar. Neural bandit based optimal llm selection for pipeline of tasks. *ACM SIGMETRICS Performance Evaluation Review*, 53(3):15–17, 2026.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Bandits with replenishable knapsacks: the best of both worlds. In *International Conference on Learning Representations*, volume 2024, pp. 19354–19372, 2024.

Leonardo Cella and Nicolo Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. In *International Conference on Artificial Intelligence and Statistics*, pp. 1168–1177. PMLR, 2020.

Xiangxiang Dai, Jin Li, Xutong Liu, Anqi Yu, and John Lui. Cost-effective online multi-llm selection with versatile reward models. *arXiv preprint arXiv:2405.16587*, 2024.

Eray Can Elumar, Cem Tekin, and Osman Yagan. Bandits with anytime knapsacks. *arXiv preprint arXiv:2501.18560*, 2025.

Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for llm selections. In *International Conference on Learning Representations*, volume 2025, pp. 26186–26203, 2025.

Arthur Flajolet and Patrick Jaillet. Logarithmic regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800*, 2015.

Yichao Fu, Siqi Zhu, Runlong Su, Aurick Qiao, Ion Stoica, and Hao Zhang. Efficient llm scheduling by learning to rank. *Advances in Neural Information Processing Systems*, 37:59006–59029, 2024.

Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. Opposing effects of response time in human–chatbot interaction. *Business & Information Systems Engineering*, 64(6): 773–791, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Shashwat Jaiswal, Kunal Jain, Yogesh Simmhan, Anjaly Parayil, Ankur Mallick, Rujia Wang, Renee St Amant, Chetan Bansal, Victor Rühle, Anoop Kulkarni, et al. Serving models fast and slow: optimizing heterogeneous llm inferencing workloads at scale. *Small*, 13(15k):17k, 2025.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International conference on machine learning*, pp. 1453–1461. PMLR, 2013.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pp. 5969–5978. PMLR, 2021.

Shaoang Li and Jian Li. Near-optimal online deployment and routing for streaming llms. In *International Conference on Learning Representations*, volume 2026, pp. 90170–90202, 2026.

Yixuan Li, Lewis Frampton, Federico Mora, and Elizabeth Polgreen. Online prompt selection for program synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11282–11289, 2025.

Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, et al. Mitigating response delays in free-form conversations with llm-powered intelligent virtual agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–15, 2025.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. *Advances in Neural Information Processing Systems*, 37:141071–141102, 2024.

Adam J Mersereau, Paat Rusmevichientong, and John N Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802, 2009.

Hyungjun Oh, Kihong Kim, Jaemin Kim, Sungkyun Kim, Junyeol Lee, Du-seong Chang, and Jiwon Seo. Exegpt: Constraint-aware resource scheduling for llm inference. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pp. 369–384, 2024.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pp. 2827–2835. PMLR, 2021.

Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. Adaptive llm routing under budget constraints. In *EMNLP (Findings)*, pp. 23934–23949, 2025.

Yifu Tang, Yingfei Wang, and Zeyu Zheng. Stochastic multi-armed bandits with strongly reward-dependent delays. In *International Conference on Artificial Intelligence and Statistics*, pp. 3043–3051. PMLR, 2024.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. *Advances in Neural Information Processing Systems*, 32, 2019.

Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas Jennings. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 1134–1140, 2012.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu, and Haifeng Chen. Mixllm: Dynamic routing in mixed large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10912–10922, 2025.

Bingyang Wu, Yinmin Zhong, Zili Zhang, Shengyu Liu, Fangyue Liu, Yuanhang Sun, Gang Huang, Xuanzhe Liu, and Xin Jin. {FastServe}:{Iteration-Level} preemptive scheduling for large language model inference. In *23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26)*, pp. 57–74, 2026.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International conference on machine learning*, pp. 1254–1262. PMLR, 2016.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Kelly W Zhang, Thomas Baldwin-McDonald, Kamil Ciosek, Lucas Maystre, and Daniel Russo. Impatient bandits: Optimizing for the long-term without delay. *Journal of Machine Learning Research*, 27(156):1–58, 2026.

Mengxiao Zhang, Yingfei Wang, and Haipeng Luo. Contextual linear bandits with delay as payoff. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 76246–76272. PMLR, 2025.

A Analysis for the budgeted oracle

The following lemmas establish Theorem 1.

A.1 From adaptive decisions to fixed allocations

Under Assumption 1, past observations do not change the expected value of a new request assigned to a given action. When these values are known, the oracle therefore gains nothing from adapting its choices to feedback. Its objective depends only on how many times each action is selected, subject to the total request count and premium budget.

For $a \in \mathcal{S}$, let $N_a(T) = \sum_{t=1}^{T} \mathbf{1}\{A_t = a\}$. The corresponding deterministic integer count oracle is defined in Eq. (6).

Lemma 1 (Zero adaptivity gap)**.** *Under Assumption 1, known expected values, deterministic costs, and the expected eventual-value objective,*

$$\mathrm{OPT}_B^{\mathrm{ad}}(T) = \mathrm{OPT}_B^{\mathrm{int}}(T).$$

Proof. Fix any nonanticipating hard-feasible policy. Freshness implies $\mathbb{E}[X_t(A_t) \mid \mathcal{H}_t, A_t] = V_{A_t}$, where $\mathcal{H}_t$ is the information available before decision t. Therefore

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t(A_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} V_{A_t}\right] = \mathbb{E}\left[\sum_{a\in\mathcal{S}} N_a(T) V_a\right]. \tag{20}$$

For every sample path, $N(T)$ is a nonnegative integer vector with $\sum_a N_a(T) = T$ and $\sum_a N_a(T) C_a \le B$. Its deterministic value is therefore at most $\mathrm{OPT}_B^{\mathrm{int}}(T)$, proving the upper bound after taking expectations. Conversely, fix an optimizer of Eq. (6) and execute any nonadaptive sequence with those counts. It is pathwise feasible and has expected value $\sum_a n_a V_a$, which proves the reverse inequality. □

The fixed-allocation problem can be simplified further because all delayed actions have zero premium cost. Let $q_* = \max_{i\in[K]} q_i$ and choose any $i_* \in \arg\max_i q_i$. With known values, assigning a delayed request to arm i_* is at least as good as assigning it to any other arm and uses the same budget.

Lemma 2 (Delayed-point domination)**.** *Every delayed point $(C_{i,\texttt{Del}}, V_{i,\texttt{Del}}) = (0, q_i)$ with $q_i < q_*$ is dominated by $(0, q_*)$. Replacing any occurrence of such an action by $(i_*, \texttt{Del})$ preserves cost and weakly increases expected value. Consequently, both the integer oracle and its LP relaxation need only the $K + 1$ points in Eq. (7).*

Proof. All Delayed actions have the same deterministic cost zero. The replacement keeps the number of actions and total cost unchanged and changes the objective by $q_* - q_i \geq 0$ per replacement. Repeating it eliminates every strictly suboptimal delayed point in either the integer or fractional problem. □

This reduction concerns the known-value oracle. A learner must still identify the best delayed arm, so the reduction does not justify discarding other delayed arms before their values are learned.

A.2 The fractional frontier and its integer approximation

After the reduction, the oracle chooses among one zero-cost baseline and K immediate actions. Allowing fractional allocations gives a linear program (LP) whose optimal value describes the trade-off between premium spending and expected reward. For a per-round budget $b \geq 0$, this frontier is defined in Eq. (8). To describe adjacent frontier vertices, we use a reduced set of points: at each cost, retain only the point with the highest value; remove vertically dominated points and points strictly below the upper concave hull; and remove collinear interior points. On a flat optimal face, select the cheapest maximizer. These conventions preserve the LP's value and give a consistent representation of the frontier.

Because the LP has a single budget constraint, an optimal allocation needs at most two efficient points. When the budget lies between adjacent vertices, their mixture attains the frontier, and the edge slope gives the marginal value of additional premium budget.

Lemma 3 (Concave budget frontier)**.** *The function F is nondecreasing, concave, and piecewise linear. An optimum can always be chosen with support at most two. If b lies in the interior of an efficient edge with adjacent endpoints L, R and $c_L < b < c_R$, its mixing weights are*

$$x_L = \frac{c_R - b}{c_R - c_L}, \qquad x_R = \frac{b - c_L}{c_R - c_L}. \tag{21}$$

Moreover,

$$F(b) = \min_{p \geq 0} \left\{ pb + \max_{0 \leq j \leq K} (v_j - pc_j) \right\}. \tag{22}$$

Every dual optimizer p^* is a supporting slope: for all $b' \geq 0$, $F(b') \leq F(b) + p^*(b' - b)$.

Proof. Feasible sets are nested in b, so F is nondecreasing. If $x^{(1)}$ and $x^{(2)}$ are feasible at b_1 and b_2, then $\theta x^{(1)} + (1 - \theta)x^{(2)}$ is feasible at $\theta b_1 + (1 - \theta)b_2$ and attains the same convex combination of their values. This proves concavity. Because the point set is finite, the feasible cost–value pairs form a polygon; its upper boundary is piecewise linear.

At an extreme optimum, the simplex equality and at most one binding budget inequality leave at most two positive coordinates. When the budget is slack, one coordinate suffices. On an efficient edge with $c_L < b < c_R$, both the simplex and budget constraints bind; solving $x_L + x_R = 1$ and $x_L c_L + x_R c_R = b$ gives Eq. (21).

Finally, attach multiplier $p \geq 0$ to the budget constraint. Maximizing the Lagrangian over the simplex gives $pb + \max_j(v_j - pc_j)$. Finite-dimensional LP strong duality proves Eq. (22). If p^* is optimal at b, then for any b' weak duality at the same price yields

$$F(b') \leq p^* b' + \max_j (v_j - p^* c_j) = F(b) + p^*(b' - b),$$

which is the claimed supporting-line inequality. Thus p^* is the shadow price of one additional unit of per-round budget. □

Scaling the per-round allocation by the horizon T gives the fractional count relaxation

$$\mathrm{LP}_B(T) := \max_{z \geq 0} \left\{ \sum_{j=0}^{K} z_j v_j : \sum_j z_j = T,\ \sum_j z_j c_j \leq B \right\} = TF(B/T). \tag{23}$$

The fractional optimum is not necessarily an executable allocation because request counts must be integers. For a two-endpoint optimum, rounding down the number of requests assigned to the higher-cost endpoint and assigning the remainder to the lower-cost endpoint preserves both the horizon and the budget. The value loss is less than the difference between the endpoint values; a single-vertex optimum is already integral. This gives the following bound, where G denotes the width of the action-value interval.

Lemma 4 (LP–integer gap)**.** *If all action values lie in an interval of width G, then*

$$0 \leq \mathrm{LP}_B(T) - \mathrm{OPT}_B^{\mathrm{ad}}(T) \leq G.$$

In particular, under $R_i, g \in [0,1]$,

$$0 \leq \mathrm{LP}_B(T) - \mathrm{OPT}_B^{\mathrm{ad}}(T) \leq 1. \tag{24}$$

Proof. The first inequality follows because the LP relaxes integer counts and from Lemma 1. Choose an optimal frontier solution with support at most two. A one-point solution repeated T times is already integer and feasible. Otherwise let its adjacent endpoints be L, R, with $c_L < c_R$ and, after dominated points are removed, $v_L \leq v_R$. If the LP uses $z_R = Tx_R$ copies of R, set

$$n_R = \lfloor z_R \rfloor, \qquad n_L = T - n_R.$$

Replacing the fractional remainder of the higher-cost endpoint by L cannot increase cost, so (n_L, n_R) is integer feasible. Its loss relative to the LP is

$$(z_R - \lfloor z_R \rfloor)(v_R - v_L) \leq G.$$

The integer optimum is at least this rounded solution. Lemma 1 completes the proof; $G = 1$ under the paper's normalization. □

Here, G denotes a generic value-interval width, distinct from the net-value range used in the cost-penalized analysis.

A.3 Proof of Theorem 1

Proof. Items 1–4 are Lemmas 1, 2, 3, and 4, respectively, together with the scaling identity in Eq. (23). □

B Regret Analysis

This appendix gives the auxiliary results used in Section 5. We first develop the confidence and effective-delay arguments, then establish the geometry, budget feasibility, and regret bound for `EQF-LADA`.

B.1 Supporting results for delayed-value confidence

For $\delta \in (0,1)$ and $1 \leq n \leq T$, define

$$\beta_n(\delta) := \sqrt{\frac{\log(2KT/\delta)}{2n}}. \tag{25}$$

For $n_i(t) > 0$, the clipped confidence endpoints are

$$\mathrm{LCB}_i^q(t) := \max\{0, \widehat{q}_i^{-}(t) - \beta_{n_i(t)}(\delta)\}, \tag{26}$$

$$\mathrm{UCB}_i^q(t) := \min\{1, \widehat{q}_i^{+}(t) + \beta_{n_i(t)}(\delta)\}. \tag{27}$$

When $n_i(t) = 0$, use the interval $[0, 1]$.

This section constructs a confidence interval for the delayed value $q_i = \mathbb{E}[R_i g(D_i)]$ from every Delayed submission, including requests that have not yet completed. The construction uses submission order rather than arrival order and therefore does not treat the arrival-selected subset as i.i.d. All-submission zero/one imputation is used by OPSE Lancewicki et al. (2021); here the upper imputation is the certified age-dependent bound on $Rg(D)$.

B.1.1 Latent submission streams and observable envelopes

For each arm i, pre-generate an i.i.d. latent stream in Delayed-submission order,

$$(R_{i,j}, D_{i,j}) \overset{\text{iid}}{\sim} \mathsf{P}_i, \qquad Y_{i,j} := R_{i,j} g(D_{i,j}) \in [0, 1], \qquad j \geq 1. \tag{28}$$

The index j is the submission index, not the arrival index. Adaptive action selection only determines how long a prefix of this pre-generated stream has been submitted.

The observable envelopes are defined in Section 4. Also write

$$\overline{Y}_{i,n} := \frac{1}{n} \sum_{j=1}^{n} Y_{i,j}. \tag{29}$$

Lemma 5 (Pathwise latent-prefix sandwich)**.** *For every realization, arm i, and decision time t with $n_i(t) > 0$,*

$$\widehat{q}_i^{-}(t) \leq \overline{Y}_{i,n_i(t)} \leq \widehat{q}_i^{+}(t). \tag{30}$$

No probabilistic independence is required for this inequality.

Proof. The middle term is the sum of the arrived and pending latent values divided by $n_i(t)$. An arrived value is included unchanged in both envelopes. For every pending j, Eq. (59) gives $0 \leq Y_{i,j} \leq u_{i,j}(t)$. Replacing each pending value first by zero and then by its certified envelope proves the two inequalities pathwise. □

Use the radius and clipped endpoints in Eqs.Eq. (25)–equation 27.

Lemma 6 (Simultaneous deterministic-prefix concentration)**.** *With probability at least $1 - \delta$, simultaneously for every $i \in [K]$ and every deterministic $n \in [T]$,*

$$|\overline{Y}_{i,n} - q_i| \leq \beta_n(\delta). \tag{31}$$

Proof. For fixed i, n, the first n latent values in Eq. (28) are i.i.d. in $[0, 1]$ with mean q_i, even though $R_{i,j}$ and $D_{i,j}$ within a pair may be dependent. Hoeffding's inequality gives

$$\Pr\{|\overline{Y}_{i,n} - q_i| > \beta_n(\delta)\} \leq 2\exp(-2n\beta_n(\delta)^2) = \frac{\delta}{KT}.$$

A union bound over the KT deterministic arm–prefix pairs proves the claim. □

B.2 Supporting results for effective-delay regret

We analyze Algorithm 1 under the cost-penalized objective in Eq. (10).

B.2.1 Certified effective-delay complexity

Define the outstanding certified discounted-value mass of arm i before decision t by

$$M_i(t) := \sum_{j \in \mathcal{P}_i(t)} u_{i,j}(t). \tag{32}$$

For an arbitrary clock, the certified effective-delay complexity is

$$\Psi_g^{\text{cert}}(H) := \sup_s \sum_{k=1}^{H} u_{s,s+k}. \tag{33}$$

For a homogeneous clock, $u_{s,s+k} = u_k$, so $\Psi_g^{\mathrm{cert}} = \Psi_g$ with

$$\Psi_g(H) := \sum_{k=1}^{H} u_k. \tag{34}$$

For the real-valued ceiling clock and exponential kernel,

$$u_k = e^{-\gamma(k-1)}, \qquad \Psi_{\exp}^{\mathrm{real}}(H) = \begin{cases} \dfrac{1-e^{-\gamma H}}{1-e^{-\gamma}}, & \gamma > 0, \\ H, & \gamma = 0. \end{cases} \tag{35}$$

The leading unit mass in this expression is a consequence of the real/ceiling clock and must not be replaced by the integer-delay expression $1/(e^{\gamma}-1)$.

Let $N_i^{\mathrm{D}}(T)$ be the number of Delayed submissions of arm i by round T, and let $H_m = \sum_{j=1}^{m} 1/j$ with $H_0 = 0$.

Lemma 7 (Outstanding-mass charging)**.** *For every realized, possibly adaptive action and arrival sequence,*

$$\sum_{t:A_t=(i,\mathtt{Del})} \frac{M_i(t)}{\max\{1, n_i(t)\}} \le \Psi_g^{\mathrm{cert}}(T) H_{N_i^{\mathrm{D}}(T)} \le \Psi_g^{\mathrm{cert}}(T)(1+\log T). \tag{36}$$

The statement permits arbitrary out-of-order arrivals and requests that never complete.

Proof. Let $\tau_{i,r}$ be the round of arm i's r-th Delayed submission and put $N = N_i^{\mathrm{D}}(T)$. If $N \le 1$, there is no earlier Delayed submission at any selected Delayed decision, so the left-hand side is zero and the claim holds (including $H_0 = 0$). Assume henceforth that $N \ge 2$. Immediately before $\tau_{i,r}$, exactly $r-1$ Delayed samples of this arm have been submitted, so $n_i(\tau_{i,r}) = r-1$. The first Delayed selection has zero previous Delayed samples and hence $M_i = 0$; it is handled by the algorithm's zero-count initialization rather than by a nonzero-count charge. The charging sum therefore begins with the second Delayed submission. Expanding those remaining terms and swapping the order of summation gives

$$\begin{aligned} &\sum_{r=2}^{N} \frac{1}{r-1} \sum_{\substack{1\le j<r: \\ j\in\mathcal{P}_i(\tau_{i,r})}} u_{\tau_{i,j},\tau_{i,r}} \\ &\quad = \sum_{j=1}^{N-1} \sum_{\substack{j<r\le N: \\ j\in\mathcal{P}_i(\tau_{i,r})}} \frac{u_{\tau_{i,j},\tau_{i,r}}}{r-1} \\ &\quad \le \sum_{j=1}^{N-1} \frac{1}{j} \sum_{\substack{j<r\le N: \\ j\in\mathcal{P}_i(\tau_{i,r})}} u_{\tau_{i,j},\tau_{i,r}}. \end{aligned}$$

For a fixed j, the calendar ages $\tau_{i,r} - \tau_{i,j}$ in the inner sum are distinct positive integers no larger than T. Nonnegativity and Eq. (33) imply that this possibly sparse sum is at most $\Psi_g^{\mathrm{cert}}(T)$, regardless of which other arms are selected between the displayed times. Hence the last display is at most $\Psi_g^{\mathrm{cert}}(T) H_{N-1} \le \Psi_g^{\mathrm{cert}}(T) H_N$. The standard inequality $H_N \le 1 + \log N \le 1 + \log T$ completes the proof. If a sample never arrives, it remains among the charged terms, but the same envelope sum bounds all of its possible future contributions. □

Let $G := 1 + c_{\max}$ bound the range of cost-penalized action values. With fresh request outcomes, optimism converts the selected net-value gap into the corresponding confidence width; the charging lemma accounts for pending values. This theorem concerns the cost-penalized objective only, with no resource-feasibility assertion. Its exact G/T term is retained when costs are not regarded as fixed constants. Theorem 3 gives the resulting cost-penalized regret bound. Full proof: Appendix B.4.

Why effective delay suffices. Optimism charges each selected action to its confidence width. The ordinary radii sum as square roots of stream counts. In the pending term, reversing the sum charges each request at most its lifetime envelope sum $\Psi_g^{\mathrm{cert}}(T)$, followed by a harmonic factor. This argument applies even to requests that never complete.

For a summable envelope, the delay overhead is $O(K\Psi_g^{\mathrm{cert}}(\infty)\log T)$ without a raw-delay tail assumption. For fixed K, more generally, $\Psi_g^{\mathrm{cert}}(T)\log T = o(T)$ suffices for a sublinear delay term. The harmonic factor in this guarantee cannot be dropped. Appendix B.8 records bounded-delay, exponential, and additional kernel specializations.

B.3 Proof of Theorem 2

Proof. Work on the event in Lemma 6. The adaptive count $n_i(t)$ merely selects one deterministic prefix from an event already valid for every prefix length; we never condition on the random count. Combining Eq. (31) at $n = n_i(t)$ with Lemma 5 proves coverage before clipping, and clipping to $[0, 1]$ preserves it. For the upper width, write

$$\begin{aligned}\mathrm{UCB}_i^q(t) - q_i &\le \widehat{q}_i^{+}(t) + \beta_{n_i(t)}(\delta) - q_i \\ &= \big(\widehat{q}_i^{+}(t) - \overline{Y}_{i,n_i(t)}\big) + \big(\overline{Y}_{i,n_i(t)} - q_i\big) + \beta_{n_i(t)}(\delta) \\ &\le \frac{1}{n_i(t)} \sum_{j \in \mathcal{P}_i(t)} \big(u_{i,j}(t) - Y_{i,j}\big) + 2\beta_{n_i(t)}(\delta),\end{aligned}$$

which is at most Eq. (17). Every step is pathwise except the deterministic-prefix concentration event, so adaptive selection, reward–delay dependence, and out-of-order arrivals do not affect the argument. No bound on delays is needed: even when infinite delay has positive probability, each unfinished request remains in the pending set and obeys its certified envelope. Under the real-valued ceiling clock, this envelope is exactly Eq. (3). □

B.4 Proof of Theorem 3

Proof. Expose, for each arm, one i.i.d. latent stream of Immediate rewards and one i.i.d. latent stream of Delayed values. Instantiate Theorem 2 for the Delayed streams with $\delta_{\mathrm{D}} = 1/(2T^2)$, and use the same radius for the Immediate streams with failure allocation $\delta_{\mathrm{I}} = 1/(2T^2)$. Indeed, for either family there are KT deterministic prefixes, each two-sided Hoeffding failure is at most $2e^{-L_T}$, and

$$2KTe^{-L_T} = 2KTe^{-\log(4KT^3)} = \frac{1}{2T^2}.$$

A union bound over the Immediate and Delayed families therefore fails with probability at most T^{-2}. Equivalently, this explains the exact choice $L_T = \log(4KT^3)$. On the resulting event, simultaneously

$$|\widehat{\mu}_{i,m}^{\mathrm{I}} - \mu_i| \le \sqrt{L_T/(2m)}, \qquad |\overline{Y}_{i,n} - q_i| \le \sqrt{L_T/(2n)} \tag{37}$$

for all i and $1 \le m, n \le T$. This is a deterministic-prefix event; the adaptive counts only select prefixes from it.

On this event all action indices are optimistic. If the selected action is Immediate and $m_i(t) > 0$, its overestimation is at most $2\beta^T_{m_i(t)}$; if its count is zero it is at most one. For a selected Delayed action, Theorem 2 and its proof give, for $n_i(t) > 0$,

$$U_i^{\mathtt{Del}}(t) - q_i \le 2\beta^T_{n_i(t)} + \frac{M_i(t)}{n_i(t)}, \tag{38}$$

and the zero-count width is at most one. Index maximization and optimism of an optimal action imply

$$\Delta_{A_t} \le U_{A_t}(t) - \widetilde{V}_{A_t}. \tag{39}$$

Charge the first selection of each of the $2K$ arm–mode streams by one. At the r-th later selection of a stream, its mode-compatible count is $r - 1$: Immediate observations have returned by the

next decision, while Delayed counts increase at submission rather than completion. Thus the total ordinary-radius contribution is at most

$$2\sqrt{2L_T}\sum_{a\in\mathcal{S}}\sqrt{N_a(T)} \le 2\sqrt{2L_T}\sqrt{2KT} = 4\sqrt{KTL_T},$$

where we used $\sum_{r=1}^{m} r^{-1/2} \le 2\sqrt{m}$ and Cauchy–Schwarz. Lemma 7, summed over arms, bounds the pending term by $K\Psi_g^{\mathrm{cert}}(T)(1+\log T)$. This proves the second branch of Eq. (18) on the good event. On its complement, the per-round net-value gap is at most G, adding at most $GT \cdot T^{-2} = G/T$ in expectation. Finally $R_u(T) \le GT$ holds deterministically, giving the minimum with the trivial cap. The argument uses concentration of bounded discounted values and certified pending envelopes, without any upper bound or tail assumption on raw delays. It therefore allows reward–delay dependence, out-of-order arrivals, and infinite delays. □

B.5 Hard-budget feasibility and regret proof

We analyze `EQF-LADA` in three steps. First, the frontier geometry and local credit rule give pathwise budget feasibility. Second, monotone quantization limits the number of epochs, each of which incurs less than one unit of implementation loss. Finally, we combine this loss with confidence widths and quantization error to bound regret.

Frontier construction and epoch notation. Sort points by cost and, at each duplicated cost, retain the largest quantized immediate value with fixed arm ties. Scan distinct costs in increasing order and discard any point whose value is no larger than a cheaper point's value. Build a stack of the remaining points: whenever the last three points have left slope less than or equal to right slope, remove the middle point and repeat. The resulting hull has strictly decreasing positive slopes. At budget b, choose the last vertex if affordable; otherwise choose the vertex at b, if present, or the adjacent pair straddling b. This also selects the cheapest maximizer on a flat face.

Let epoch e start at decision τ_e and have length H_e, so its local decision r occurs at $t = \tau_e + r - 1$. Write $\widetilde{F}_e(b)$ for its constant frontier value and $\widetilde{v}_{L_e}, \widetilde{v}_{R_e}$ for its endpoint values. Throughout this analysis, $U_{(i,\mathtt{Imm})} = U_i^{\mathrm{I}}$ and $U_{(i,\mathtt{Del})} = U_i^{\mathrm{D}}$ are raw-value UCBs, rather than cost-penalized indices. If $0 < B < c_{\min}$, the prefix budget bound rules out every immediate action.

Frontier geometry and budget feasibility. The action types used by Algorithm 2 follow from the one-dimensional frontier in Section 3.

Lemma 8 (Slack, vertex, or edge classification)**.** *After duplicate costs and dominated or collinear nonvertices are removed, a canonical optimizer of the inequality-constrained optimistic frontier at b has exactly one of three forms:*

S. *a cheapest maximizing vertex v with $c_v < b$;*

V. *a maximizing vertex v with $c_v = b$; or*

E. *the cost-b interpolation of adjacent efficient vertices L, R with $c_L < b < c_R$.*

A flat optimal face is assigned to S or V through its cheapest maximizer.

Proof. If a feasible single point attains the optimum, choose the cheapest one; it has type S or V. Otherwise an extreme LP optimum has support two and its budget constraint must bind, giving two distinct-cost points on an upper-hull edge that straddle b. After nonvertices are removed they are adjacent. A flat face would make its cheaper endpoint optimal by itself, contradicting the two-point case. □

The local credit rule turns each of these optimizers into a sequence of actions. The next result bounds spending within an epoch; summing these bounds establishes feasibility over the full horizon.

Lemma 9 (Epoch-prefix resource certificate)**.** *For every epoch e and every prefix of r decisions within it,*

$$\sum_{s=1}^{r} C_{A_s}^{(e)} \le rb. \tag{40}$$

In a type-E epoch,

$$0 \le D_{e,r} < c_{R_e} - c_{L_e} \le c_{\max}. \tag{41}$$

Proof. The claim is immediate for S and V. In type E, the threshold rule preserves $D_{e,r} \ge 0$. An R step subtracts $c_R - b > 0$, while an L step is taken only when $D_{e,r-1} < c_R - b$ and adds $b - c_L$, so the strict upper bound $c_R - c_L$ is also invariant. Summing the credit update gives $\sum_{s=1}^{r} C_{A_s}^{(e)} = rb - D_{e,r} \le rb$. □

Lemma 10 (Global hard feasibility from local certificates)**.** *After every global prefix of t decisions,* `EQF-LADA` *satisfies*

$$\sum_{s=1}^{t} C_{A_s} \le tb \le B. \tag{42}$$

Hence every selected action is physically affordable.

Proof. For $B = 0$, the boundary policy uses only zero-cost Delayed actions. For $B \ge Tc_{\max}$, every selected cost is at most $c_{\max} \le b$, so the same prefix inequality holds directly. In the intermediate regime, decompose a global prefix into completed epochs and a prefix of the current epoch, and apply Lemma 9 to each piece. Their lengths sum to t, so their costs sum to at most $tb \le Tb = B$. Resetting a virtual credit changes no physical budget. □

Quantization and implementation loss. On the simultaneous confidence event, taking running minima and rounding upward preserve optimism:

$$V_a \le \overline{U}_a(t) \le U_a(t), \qquad V_a \le \widetilde{U}_a(t) \le U_a(t) + \varepsilon. \tag{43}$$

Each of the K immediate coordinates and the aggregate zero-cost coordinate is nonincreasing and has at most

$$L_\varepsilon := \left\lceil \frac{1}{\varepsilon} \right\rceil \tag{44}$$

strict level drops. Since an epoch ends only when the complete quantized vector changes, the number of epochs satisfies

$$J_\varepsilon \le 1 + (K+1)L_\varepsilon. \tag{45}$$

A change in the realizing arm with no change in this vector does not affect the bound. A change hidden by duplicate merging or hull pruning still counts as an epoch boundary.

Within a type-S or type-V epoch, the selected vertex attains the fractional optimum. For a type-E epoch, the credit invariant bounds the cumulative loss from scheduling the two endpoints. The edge slope and cost span cancel, giving a bound independent of the minimum positive cost separation.

Lemma 11 (Unit loss per binding epoch)**.** *Let $\mathcal{E}_{\mathrm{E}}$ be the set of type-E epochs. Then*

$$\sum_{t=1}^{T} \left[\widetilde{F}_t(b) - \widetilde{U}_{A_t}(t) \right] \le |\mathcal{E}_{\mathrm{E}}| \le J_\varepsilon \le 1 + (K+1)L_\varepsilon. \tag{46}$$

The bound has no dependence on a minimum positive cost separation.

Proof. In type S or V, the selected vertex attains $\widetilde{F}_t(b)$, so the implementation loss is zero. Fix a type-E epoch and put

$$\widetilde{p}_e := \frac{\widetilde{v}_{R_e} - \widetilde{v}_{L_e}}{c_{R_e} - c_{L_e}} \ge 0.$$

For either endpoint selected at its r-th decision,

$$\widetilde{F}_e(b) - \widetilde{U}_{A_t}(t) = \widetilde{p}_e(b - C_{A_t}) = \widetilde{p}_e(D_{e,r} - D_{e,r-1}).$$

Summation telescopes to $\widetilde{p}_e D_{e,H_e}$. By Lemma 9,

$$0 \le \widetilde{p}_e D_{e,H_e} < \widetilde{p}_e(c_{R_e} - c_{L_e}) = \widetilde{v}_{R_e} - \widetilde{v}_{L_e} \le 1.$$

Thus each binding epoch contributes less than one and the cost span cancels. Eq. (45) completes the proof. □

Confidence widths and regret. Write $m_i^{\mathrm{I}}(t) = m_i(t)$ for the Immediate count already used in Section 4, and let $M_i^{\mathrm{I}}(T)$ be its horizon total. Define

$$M_B := \min\left\{T, \left\lfloor \frac{B}{c_{\min}} \right\rfloor\right\}, \tag{47}$$

Lemma 12 (Premium confidence-width sum)**.** *For every pathwise hard-feasible policy and* $\beta_m^T = \sqrt{L_T/(2m)}$*,*

$$\sum_{\substack{t:A_t=(i_t,\mathtt{Imm});\\ m_{i_t}^{\mathrm{I}}(t)\ge 1}} \beta^T_{m_{i_t}^{\mathrm{I}}(t)} \le \sqrt{2KM_BL_T}. \tag{48}$$

On the simultaneous confidence event, including zero-count initialization, Immediate UCB overestimation contributes at most

$$\min\{K, M_B\} + 2\sqrt{2KM_BL_T}. \tag{49}$$

Proof. Every Immediate action costs at least $c_{\min}$, hence $\sum_i M_i^{\mathrm{I}}(T) \le M_B$. Using $\sum_{m=1}^n m^{-1/2} \le 2\sqrt{n}$ and Cauchy–Schwarz,

$$\sum_{\substack{t:A_t=(i_t,\mathtt{Imm});\\ m_{i_t}^{\mathrm{I}}(t)\ge 1}} \beta^T_{m_{i_t}^{\mathrm{I}}(t)} \le \sqrt{2L_T}\sum_i \sqrt{M_i^{\mathrm{I}}(T)} \le \sqrt{2KM_BL_T}.$$

At most $\min\{K, M_B\}$ selected Immediate streams have zero count, and a nonzero-count UCB overestimates by at most $2\beta_m^T$. □

Thus premium uncertainty scales with affordable premium opportunities, not with an inverse budget rate. Theorems 2 and 3 similarly give the reusable Delayed width sum

$$K + 2\sqrt{2KTL_T} + K\Psi_g^{\mathrm{cert}}(T)(1+\log T). \tag{50}$$

The reuse in Eq. (50) is a width calculation, not an invocation of the warm-up's action-selection rule. Deterministic-prefix concentration and outstanding-mass charging hold for any adaptive selection sequence, including the current realizers chosen by `EQF-LADA`.

Here "budget-uniform" means that the bound has no inverse budget-rate factor; M_B still depends on B. The symbol $\rho := \min\{B/T, c_{\max}\}$ below refers only to the legacy rate convention.

The theorem combines these widths with epoch implementation and compares to $\mathrm{OPT}_B^{\mathrm{ad}} = \mathrm{OPT}_B^{\mathrm{int}} \le TF(B/T)$. It assumes the fresh law and known deterministic resource of the formulation; it does not cover unknown stochastic costs or a queue whose future law is changed by routing.

The optimized bound uses the quantization choice

$$\varepsilon_T := \min\left\{1, \sqrt{\frac{K+1}{T}}\right\}, \tag{51}$$

Explicit hard-budget regret bounds. Let Assumption 1 hold, let costs be known and deterministic, and let g be any known nonincreasing kernel under the certified clock. Set $L_T = \log(4KT^3)$ and $M_B = \min\{T, \lfloor B/c_{\min}\rfloor\}$. For any $\varepsilon \in (0,1]$, Algorithm 2 is pathwise hard feasible for every $B \ge 0$ and satisfies

$$\begin{aligned} R_B(T) \le \min\Big\{T,\, 1 + (K+1)\left\lceil\frac{1}{\varepsilon}\right\rceil + T\varepsilon + \min\{K, M_B\} + K \\ + 2\sqrt{2KM_BL_T} + 2\sqrt{2KTL_T} + K\Psi_g^{\mathrm{cert}}(T)(1+\log T) + \frac{1}{T}\Big\}. \end{aligned} \tag{52}$$

With $\varepsilon_T := \min\{1, \sqrt{(K+1)/T}\}$ and $T \geq K+1$, this becomes

$$R_B(T) \leq \min\Big\{T,\, 2\sqrt{2KTL_T} + 2\sqrt{2KM_BL_T} + K\Psi_g^{\mathrm{cert}}(T)(1+\log T) \\ + 2\sqrt{(K+1)T} + 2K + 2 + \min\{K, M_B\} + \frac{1}{T}\Big\}. \tag{53}$$

For $T < K+1$, the trivial bound T applies. The result has no inverse budget-rate, depletion-tail, or inverse cost-separation factor.

Proof of Theorem 4.

Proof. The boundary policies are feasible. At $B = 0$, selected Delayed widths are bounded by Eq. (50); at $B \geq Tc_{\max}$, Immediate-only UCB is feasible and $\mu_i \geq q_i$. Both cases are dominated by the displayed bound.

Now suppose $0 < B < Tc_{\max}$. Lemma 10 proves pathwise feasibility. On the simultaneous deterministic-prefix confidence event from Theorem 3, coordinatewise optimism implies $\widetilde{F}_t(b) \geq F(b)$. Hence

$$\begin{aligned} F(b) - V_{A_t} &\leq \widetilde{F}_t(b) - V_{A_t} \\ &\leq [\widetilde{F}_t(b) - \widetilde{U}_{A_t}(t)] + [U_{A_t}(t) - V_{A_t}] + \varepsilon. \end{aligned}$$

Lemma 11 bounds the first term over time by $1 + (K+1)L_\varepsilon$. Lemma 12 bounds the Immediate widths by $\min\{K, M_B\} + 2\sqrt{2KM_BL_T}$, and Eq. (50) bounds the Delayed widths. Quantization contributes $T\varepsilon$.

Theorem 1 gives $\mathrm{OPT}_B^{\mathrm{ad}}(T) \leq TF(B/T)$. Taking expectations therefore controls budgeted regret on the good event. Its failure probability is at most T^{-2}, contributing at most $1/T$. This proves Eq. (52). Finally,

$$1 + (K+1)\left\lceil \frac{1}{\varepsilon} \right\rceil + T\varepsilon \leq K + 2 + \frac{K+1}{\varepsilon} + T\varepsilon,$$

and the stated ε_T gives Eq. (53). □

Reading the expectation step. Fresh outcomes give $\mathbb{E}[\sum_t X_t(A_t)] = \mathbb{E}[\sum_t V_{A_t}]$. Thus the good/bad-event split in the proof applies to $Z = \mathrm{OPT}_B^{\mathrm{ad}}(T) - \sum_t V_{A_t}$, whose expectation is regret, rather than to a conditional expectation of realized rewards. On the good event, bound Z by the frontier decomposition; on its complement, $Z \leq T$. Probability at most T^{-2} therefore contributes at most $1/T$. Also, the implementation gap need not be nonnegative in every round: an expensive endpoint can exceed the fractional frontier value. Its sum within a binding epoch is the nonnegative telescoping quantity in Lemma 11.

Qualitatively, under the displayed logarithmic factors,

$$R_B(T) = \widetilde{O}\Big(\sqrt{KT} + \sqrt{KM_B} + K\Psi_g^{\mathrm{cert}}(T)\Big).$$

Costs enter through M_B, not through a minimum positive cost separation. The local cancellation $\widetilde{p}_e(c_{R_e} - c_{L_e}) = \widetilde{v}_{R_e} - \widetilde{v}_{L_e} \leq 1$ also covers arbitrarily small edge spans. When $0 < B < c_{\min}$, the global prefix certificate still makes every selected action affordable. Slack optima need not spend the surplus budget; exact-cost vertices use type V; flat faces use their cheapest maximizer; and duplicate costs are merged before classification.

B.6 Computational cost and implementation details

With P_t pending requests before round t, a direct implementation takes $O(K + P_t + K\log K)$ time per decision to update the confidence estimates, refresh pending envelopes, and construct the frontier. Processing arrivals adds $O(T\log T)$ total time with an event heap; each finite arrival is processed once, and infinite delays generate no arrival event. Discarding completed records after updating their sums requires $O(K + \max_t P_t)$ online state. Persistent pending requests can make $\sum_t P_t = O(T^2)$.

For exponential discounting, the pending sum can instead be updated by multiplying it by $e^{-\gamma}$ each round and removing each arrival's current contribution. This aggregate implementation requires checks of equivalence and numerical stability against the direct version. Section 6 specifies the calibrated confidence radii and quantization used in the experiments. Budget feasibility holds independently of confidence coverage, while the regret bound requires the radii specified in the analysis.

B.7 Proof of Theorem 5

Proof. Choose $h \in [H]$ attaining $\Phi_g(H)$. If $u_h > 0$, by the definition of a supremum there is a finite d_h with

$$d_h > h - 1, \qquad g(d_h) \geq \frac{1}{2} u_h. \tag{54}$$

In particular $d_h > 0$ and $A(d_h) = \lceil d_h \rceil \geq h$. If $u_h = 0$, the desired lower bound is zero and any finite $d_h > h - 1$ suffices.

Give both arms this same deterministic delay. In environment θ_+, let the rewards be independent Bernoulli random variables with means $(3/4, 1/4)$ for arms $(1, 2)$; in θ_-, swap these means. Reward draws are fresh across requests and the delay law is identical in the two environments. Since $B = 0$ and premium costs are positive, every feasible policy must use Delayed mode.

During decisions $t = 1, \ldots, h$, every request submitted earlier has age at most $h - 1 < A(d_h)$ and is still pending. Thus no reward has been revealed, and the entire observed transcript—including all pending indicators—has the same law under θ_+ and θ_-. Consequently the action distribution at each of these decisions is the same in the two environments. If W_+ is the number of arm-2 choices under θ_+ and W_- is the number of arm-1 choices under θ_- during these h decisions, then

$$\mathbb{E}_{\theta_+}[W_+] + \mathbb{E}_{\theta_-}[W_-] = h. \tag{55}$$

At least one environment therefore has at least $h/2$ expected wrong choices. The delayed-value gap in either environment is $(3/4 - 1/4)g(d_h) = g(d_h)/2$. Its regret over the first h decisions, and hence over horizon T, is at least

$$\frac{h}{2} \cdot \frac{g(d_h)}{2} \geq \frac{1}{8} h u_h = \frac{1}{8} \Phi_g(H).$$

□

B.8 Clock conventions and additional kernels

Feedback timing. The map from continuous delay to feedback rounds is

$$A(d) = \lceil d \rceil, \; A(\infty) = \infty. \tag{56}$$

Immediate feedback is normalized to the end of the submission round; this does not assume instantaneous computation. A completed delayed request reveals (R_s, D_s) and hence its raw reward, discounted reward, and latency loss.

Pending-value envelopes. For a request submitted in round $s < t$, the delays consistent with its pending status before decision t form the set

$$\mathcal{D}_{s,t} := \{d \in [0, \infty] : s + A(d) > t - 1\} = \{d \in [0, \infty] : A(d) \geq t - s\}. \tag{57}$$

Define its lower endpoint and the corresponding discount envelope by

$$L_{s,t} := \inf \mathcal{D}_{s,t}, \qquad u_{s,t} := \sup_{d \in \mathcal{D}_{s,t}} g(d). \tag{58}$$

Since $R_s \in [0, 1]$, this gives the pathwise bound

$$0 \leq R_s g(D_s) \leq u_{s,t}. \tag{59}$$

Under the real-valued ceiling clock, a request of age $k = t - s \geq 1$ has $\mathcal{D}_{s,t} = (k - 1, \infty]$ and $L_{s,t} = k - 1$, yielding Eq. (3). The supremum accounts for the open lower endpoint and permits discontinuous kernels. In particular, pending status implies $D_s > k - 1$, rather than $D_s > k$.

For exponential discounting, the envelope and its sum over the first H decision ages are

$$u_{s,t} = e^{-\gamma(k-1)}, \qquad \sum_{k=1}^{H} u_{s,s+k} = \frac{1-e^{-\gamma H}}{1-e^{-\gamma}} \quad (\gamma > 0). \tag{60}$$

When $\gamma = 0$, the envelope is one and the sum is H.

Corollary 1 (Optional integer-delay clock)**.** *If instead $D_s \in \mathbb{N}_0 \cup \{\infty\}$ and the same availability rule is used, pending at age k implies $D_s \geq k$. Hence $\mathcal{D}_{s,s+k} = \{k, k+1, \ldots\} \cup \{\infty\}$, $L_{s,s+k} = k$, and $u_{s,s+k} = g(k)$. For exponential discounting,*

$$\sum_{k=1}^{H} u_{s,s+k} = \frac{e^{-\gamma}(1-e^{-\gamma H})}{1-e^{-\gamma}} \quad (\gamma > 0).$$

For $\gamma = 0$, it again equals H.

For the homogeneous real/ceiling clock, the envelope sums give the following examples.

- Exponential $g(d) = e^{-\gamma d}$: Eq. (35) is uniformly bounded in H for every fixed $\gamma > 0$.
- Hyperbolic $g(d) = (1+\beta d)^{-1}$: $u_k = (1+\beta(k-1))^{-1}$ and $\Psi_g(H) = \Theta(1+\beta^{-1}\log(1+\beta H))$ for fixed $\beta > 0$.
- Power law $g(d) = (1+d)^{-\alpha}$: $u_k = k^{-\alpha}$, so $\Psi_g(H)$ is $\Theta(1)$ for $\alpha > 1$, $\Theta(\log H)$ for $\alpha = 1$, and $\Theta(H^{1-\alpha})$ for $0 < \alpha < 1$.

These rates are governed by summability of the certified value envelope, not by boundedness of the raw delay. They concern learning q_i only. Pending raw rewards do not inherit the shrinkage. At the boundary $g \equiv 1$, $u_k = 1$ and $\Psi_g(H) = H$; the delay branch of Eq. (18) becomes vacuous after the trivial cap, as it should for tail-free raw-reward learning with unbounded delay.

B.9 Consequences of effective-delay regret

Corollary 2 (Bounded and unbounded raw delays)**.** *If $A(D_i) \leq \overline{D}$ almost surely for every arm, with $\overline{D} \in \mathbb{N}_0$, the delay term in Eq. (18) can be replaced by*

$$K\left(\sup_s \sum_{k=1}^{\min\{T,\overline{D}\}} u_{s,s+k}\right)(1+\log T),$$

because no request can remain pending at a larger age. Without any raw-delay bound, if $\Psi_g^{\mathrm{cert}}(\infty) < \infty$, the certified unresolved mass per submitted sample is horizon-uniform and the regret overhead is $O(K\Psi_g^{\mathrm{cert}}(\infty)\log T)$. More generally, for fixed K the proved delay term is sublinear whenever $\Psi_g^{\mathrm{cert}}(T)\log T = o(T)$. If $K = K(T)$ grows with the horizon, the corresponding sufficient condition is $K(T)\Psi_g^{\mathrm{cert}}(T)(1+\log T) = o(T)$; sublinear total regret additionally requires the statistical and initialization terms to be $o(T)$.

Proof. The bounded-delay statement follows by truncating the per-sample inner sum in the proof of Lemma 7. The other statements follow directly from Eq. (18). □

Corollary 3 (Exact exponential specialization)**.** *For $g(d) = e^{-\gamma d}$ under the real-valued ceiling clock and $\gamma > 0$,*

$$R_u(T) \leq \min\left\{GT, 2K + 4\sqrt{KTL_T} + K\frac{1-e^{-\gamma T}}{1-e^{-\gamma}}(1+\log T) + \frac{G}{T}\right\}. \tag{61}$$

At $\gamma = 0$, the fraction is replaced by T and the trivial GT branch controls the result.

Proof. Substitute Eq. (35) into Theorem 3. □

The harmonic logarithm in these consequences is genuine in the present proof. Even for fixed K, $\Psi_g(T) = o(T)$ alone does not imply that the displayed upper bound is sublinear; the sufficient condition supplied by Theorem 3 is $\Psi_g(T) = o(T/\log T)$.

Additional kernel examples appear in Appendix B.8.

C Lower Bound Analysis

The construction for Theorem 5 compares two fresh stationary environments with the same deterministic nonzero delay and opposite best arms. Before any response arrives, their pending transcripts have the same law. A policy must therefore make suboptimal choices in at least one environment, while discounting reduces the loss of each choice. The resulting lower bound isolates delayed information under a zero premium budget.

Lower-bound quantities. For a homogeneous clock, define

$$u_h := \sup\{g(d) : A(d) \geq h\}. \tag{62}$$

$$\Phi_g(H) := \max_{1\leq h\leq H} h u_h. \tag{63}$$

Under the ceiling clock, $u_h = \sup_{d>h-1} g(d)$. The exponential specialization of Theorem 5 is

$$\sup_\theta R_0^\pi(T;\theta) \geq \frac{1}{32}\Psi_{\exp}^{\text{real}}(H). \tag{64}$$

C.1 Exponential Φ/Ψ equivalence

Lemma 13 (Exponential envelope equivalence)**.** *For every $H \geq 1$ and $\gamma \geq 0$, under the real-valued ceiling clock,*

$$\Phi_{\exp}(H) \leq \Psi_{\exp}^{\text{real}}(H) \leq 4\Phi_{\exp}(H). \tag{65}$$

Proof. Put $r = e^{-\gamma} \in [0,1]$. Then $u_h = r^{h-1}$, $\Psi_{\exp}^{\text{real}}(H) = \sum_{h=1}^H r^{h-1}$, and $\Phi_{\exp}(H) = \max_{h\leq H} hr^{h-1}$. The first inequality follows from $hr^{h-1} \leq \sum_{k=1}^h r^{k-1}$.

For the reverse inequality, the cases $H = 1$ and $r = 1$ are equalities. Assume $H \geq 2$ and write $\lambda = 1 - r$. If $\lambda H \leq 1$, then

$$r^{H-1} \geq (1 - 1/H)^{H-1} \geq e^{-1},$$

and hence $\Psi_{\exp}^{\text{real}}(H) \leq H \leq e\Phi_{\exp}(H)$. If $\lambda H > 1$ and $\lambda \geq 1/2$, then the infinite geometric sum gives $\Psi_{\exp}^{\text{real}}(H) \leq 1/\lambda \leq 2 \leq 2\Phi_{\exp}(H)$. Finally, suppose $\lambda H > 1$ and $0 < \lambda < 1/2$. Let $m = \lceil 1/\lambda \rceil \leq H$. Since $m - 1 < 1/\lambda$ and $(1-\lambda)^{1/\lambda} \geq 1/4$ for $0 < \lambda \leq 1/2$,

$$\Phi_{\exp}(H) \geq mr^{m-1} \geq \frac{1}{\lambda}(1-\lambda)^{1/\lambda} \geq \frac{1}{4\lambda}.$$

Again the geometric sum is at most $1/\lambda$, proving the factor four. These cases include arbitrarily large γ; when $r = 0$, both quantities equal one. □

C.2 General kernels and classical arm uncertainty

Lemma 14 (General-kernel comparison)**.** *For every nonnegative nonincreasing homogeneous envelope and $H \geq 1$,*

$$\Phi_g(H) \leq \Psi_g(H) \leq H_H \Phi_g(H) \leq (1 + \log H)\Phi_g(H). \tag{66}$$

The logarithmic separation is possible.

Proof. Monotonicity gives $hu_h \leq \sum_{k=1}^h u_k \leq \Psi_g(H)$, proving the first inequality after maximizing over h. Conversely, $u_h \leq \Phi_g(H)/h$ for every $h \leq H$, so summation proves the second and third inequalities. For the real/ceiling clock and hyperbolic kernel $g(d) = (1+d)^{-1}$, one has $u_h = 1/h$ exactly. Hence $\Phi_g(H) = 1$ while $\Psi_g(H) = H_H = \Theta(\log H)$. □

Theorem 5 formalizes the two opposing effects of delay: it postpones information, while the same kernel reduces the economic loss of a wrong pre-revelation decision. The general lower-bound scale is Φ_g, not Ψ_g. For the exponential envelope, Lemma 13 establishes equivalence within a universal constant. Accordingly, the theorem does not claim a $K\Psi_g$ lower bound, general-kernel Ψ_g matching, an additive $\sqrt{KT} + \Psi_g$ bound on one instance, or minimax optimality.

C.2.1 Classical arm-identification lower bound

Proposition 1 (Classical arm-identification lower bound)**.** *Suppose $g(0) = 1$. For $2 \le K \le T$, any policy, and any budget $B \ge 0$, there is a zero-delay instance of the budgeted model with $R_B^\pi(T) \ge \sqrt{KT}/32$.*

D Fixed-Frontier Hard-Budget Control

This appendix isolates the deterministic resource-control primitive that motivates the epoch-local certificate in Algorithm 2.

Let $L = (c_L, v_L)$ and $R = (c_R, v_R)$ be adjacent points on an efficient supporting edge of the true frontier, with

$$c_L < c_R, \qquad c_L \le b := B/T \le c_R, \qquad \Delta_c := c_R - c_L, \qquad \Delta_v := v_R - v_L \ge 0. \tag{67}$$

Define the supporting slope and intercept

$$p^* := \frac{\Delta_v}{\Delta_c}, \qquad \nu^* := \max_{a \in \mathcal{S}} (V_a - p^* C_a). \tag{68}$$

Because the edge is globally supporting,

$$v_L - p^* c_L = v_R - p^* c_R = \nu^*. \tag{69}$$

D.1 Dual accounting without a depletion time

Lemma 15 (Exact fixed-frontier dual accounting)**.** *Under Eq. (67),*

$$\mathrm{LP}_B(T) = p^* B + T\nu^*. \tag{70}$$

For every hard-feasible action sequence $A_1, \dots, A_T$,

$$\mathrm{LP}_B(T) - \sum_{t=1}^{T} V_{A_t} = \underbrace{\sum_{t=1}^{T} [\nu^* - (V_{A_t} - p^* C_{A_t})]}_{\textit{reduced-cost loss}} + \underbrace{p^* \left(B - \sum_{t=1}^{T} C_{A_t} \right)}_{\textit{leftover-budget loss}}. \tag{71}$$

Both terms on the right are nonnegative.

Proof. The dual representation in Eq. (22) is minimized by the supporting slope p^* throughout this edge. Since $\max_a (V_a - p^* C_a) = \nu^*$ and $b = B/T$,

$$\mathrm{LP}_B(T) = TF(B/T) = T\big(p^* B/T + \nu^*\big) = p^* B + T\nu^*.$$

Adding and subtracting $p^* \sum_t C_{A_t}$ gives

$$p^* B + T\nu^* - \sum_t V_{A_t} = \sum_t [\nu^* - V_{A_t} + p^* C_{A_t}] + p^* \left(B - \sum_t C_{A_t} \right),$$

which is Eq. (71). The definition of ν^* makes each reduced-cost summand nonnegative. Hard feasibility makes the budget remainder nonnegative, and $p^* \ge 0$ on an increasing efficient edge. □

An active endpoint has zero reduced-cost loss by Eq. (69). A non-frontier action is charged by its reduced-cost gap, while budget under-utilization is charged only by p^* times the terminal remainder. The identity contains neither a post-depletion $T - \tau$ term nor an inverse budget-rate factor.

D.2 INTEGER ENDPOINT ALLOCATION

Define the fractional high-endpoint count

$$z_R^* := \left[\frac{B - Tc_L}{\Delta_c}\right]_{[0,T]}. \tag{72}$$

Under Eq. (67), clipping is inactive. Write

$$N_R^* := \lfloor z_R^* \rfloor, \qquad N_L^* := T - N_R^*, \qquad \alpha := z_R^* - N_R^* \in [0, 1). \tag{73}$$

Lemma 16 (Fixed-edge integer implementation)**.** *Any schedule containing N_R^* copies of R and N_L^* copies of L is hard feasible. Its unused budget and LP value loss are exactly*

$$B - (N_L^* c_L + N_R^* c_R) = \alpha\Delta_c < \Delta_c, \tag{74}$$

$$\mathrm{LP}_B(T) - (N_L^* v_L + N_R^* v_R) = \alpha\Delta_v = p^*\alpha\Delta_c \le \Delta_v. \tag{75}$$

In particular, the loss is at most one under the paper's normalized values.

Proof. The fractional edge allocation has z_R^* copies of R and $T - z_R^*$ copies of L, and it spends exactly B. Replacing the fractional part α of one R by L decreases cost by $\alpha\Delta_c$ and value by $\alpha\Delta_v$. Hence the integer allocation is feasible and the two equalities follow. Because the total cost of the entire nonnegative-cost sequence is at most B, every prefix is also affordable. Finally, $0 \le v_L \le v_R \le 1$ implies $\Delta_v \le 1$. □

This is the LP–integer rounding from Lemma 4, now viewed as a resource-control target that must be implemented online without leaving the fixed edge.

D.3 RESIDUAL RESERVE AND DETERMINISTIC BALANCED ROUNDING

Before decision t, let B_t be remaining budget and $H_t := T - t + 1$ the number of remaining rounds. For a residual state on the fixed edge, define

$$E_t := B_t - H_t c_L, \qquad Q_t := \frac{E_t}{\Delta_c}. \tag{76}$$

The reserve invariant

$$H_t c_L \le B_t \le H_t c_R \tag{77}$$

is equivalent to $0 \le Q_t \le H_t$. Here Q_t is the fractional number of additional R actions that the remaining budget can support beyond serving all remaining rounds with L.

Lemma 17 (Exact residual quota dynamics)**.** *If L is selected, then*

$$E_{t+1} = E_t, \qquad Q_{t+1} = Q_t.$$

If R is selected, then

$$E_{t+1} = E_t - \Delta_c, \qquad Q_{t+1} = Q_t - 1.$$

In both cases $H_{t+1} = H_t - 1$.

Proof. For $a \in \{L, R\}$,

$$E_{t+1} = B_t - c_a - (H_t - 1)c_L = E_t - (c_a - c_L).$$

Substituting $c_a = c_L$ or $c_R = c_L + \Delta_c$ and dividing by Δ_c proves the claim. □

The online rule uses the integer target $n := N_R^*$ and, after k completed decisions, the cumulative target

$$\overline{N}_R(k) := \left\lceil \frac{kn}{T} \right\rceil, \qquad k = 0, \ldots, T. \tag{78}$$

At decision $t = k + 1$, select R exactly when $\overline{N}_R(k+1) > \overline{N}_R(k)$; otherwise select L. This is a deterministic, front-loaded ceiling-balanced schedule. Its increments are in $\{0, 1\}$ and $\overline{N}_R(T) = n$.

Proposition 2 (Fixed-Frontier Hard-Budget Tracking)**.** *Suppose a known fixed efficient edge satisfies Eq. (67). The deterministic rule in Eq. (78) has the following properties.*

1. *It is pathwise hard feasible and, before every decision, $H_t c_L \le B_t \le H_t c_R$; in particular, it always reserves enough budget to execute L in every remaining round.*
2. *For every $0 \le k \le T$, if $N_R(k)$ is the number of R actions in the first k rounds, then*
$$\left| N_R(k) - \frac{k z_R^*}{T} \right| < 1. \tag{79}$$
Consequently, the absolute cumulative cost discrepancy from the fractional edge allocation is less than Δ_c, and the absolute cumulative value discrepancy is at most Δ_v (strict when $\Delta_v > 0$).
3. *It executes exactly $N_R^* = \lfloor z_R^* \rfloor$ high endpoints and ends with*
$$0 \le B_{T+1} = \alpha \Delta_c < \Delta_c. \tag{80}$$
Its value is within $\alpha \Delta_v \le \Delta_v \le 1$ of the fractional frontier.

None of these conclusions contains inverse dependence on $b = B/T$ or on Δ_c.

Proof. By construction, $N_R(k) = \lceil kn/T \rceil$. Since $z_R^* = n + \alpha$ with $\alpha \in [0, 1)$,
$$N_R(k) - \frac{k z_R^*}{T} = \left(\left\lceil \frac{kn}{T} \right\rceil - \frac{kn}{T} \right) - \frac{k\alpha}{T} \in (-1, 1),$$
which proves Eq. (79). Multiplication by Δ_c and Δ_v gives the cost and value discrepancies.

It remains to verify the reserve invariant. After $k < T$ actions, put $H = T - k$ and $m = T - n$. Lemma 17 gives
$$Q_{k+1} = z_R^* - N_R(k) = H - \left(m - \left\lfloor \frac{km}{T} \right\rfloor \right) + \alpha, \tag{81}$$
where we used $\lceil kn/T \rceil = k - \lfloor km/T \rfloor$. The quota is nonnegative because $N_R(k) \le n \le z_R^*$. If $m = 0$, then $n = T$, $\alpha = 0$, and $Q_{k+1} = H$. If $m \ge 1$, then $k < T$ implies $\lfloor km/T \rfloor \le m - 1$, so the parenthesized integer in Eq. (81) is at least one and $Q_{k+1} \le H$ because $\alpha < 1$. Thus $0 \le Q_{k+1} \le H$, which is precisely Eq. (77). In particular, whenever the rule selects R, its post-action excess remains nonnegative, so the current budget pays for R while preserving the all-L reserve.

At $k = T$, the schedule has selected n high endpoints, so $Q_{T+1} = z_R^* - n = \alpha$. There is no remaining-round upper invariant at $H_{T+1} = 0$; instead the actual terminal budget is $B_{T+1} = \alpha \Delta_c < \Delta_c$. Lemma 16 supplies the final value claim. □

D.4 Residual Bellman interpretation

Define
$$J(H, B) := HF(B/H), \qquad H \ge 1. \tag{82}$$

Lemma 18 (Fixed-edge Bellman identity)**.** *Suppose $H \ge 2$, $B/H \in [c_L, c_R]$, and $(B - c_a)/(H - 1) \in [c_L, c_R]$ for an endpoint $a \in \{L, R\}$. Then*
$$J(H, B) = v_a + J(H - 1, B - c_a). \tag{83}$$
Writing $Q = (B - Hc_L)/\Delta_c$, endpoint L has a valid residual state when $Q \le H - 1$, while endpoint R has a valid residual state when $Q \ge 1$.

Proof. On the fixed edge, $F(b) = \nu^* + p^* b$, so $J(H, B) = H\nu^* + p^* B$. If the successor remains on the same edge, then by Eq. (69),
$$\begin{aligned} v_a + J(H - 1, B - c_a) &= \nu^* + p^* c_a + (H - 1)\nu^* + p^*(B - c_a) \\ &= H\nu^* + p^* B = J(H, B). \end{aligned}$$
For L, the successor quota remains Q but its horizon becomes $H - 1$, so validity is equivalent to $Q \le H - 1$. For R, the successor quota is $Q - 1$, so validity is equivalent to $Q \ge 1$; the other side follows from $Q \le H$. □

The identity can fail at an edge boundary if the chosen endpoint sends the residual rate outside $[c_L, c_R]$: at $B = Hc_L$, choosing R violates the low-endpoint reserve, while at $B = Hc_R$, choosing L makes the successor rate exceed c_R. Proposition 2 avoids both invalid moves and keeps every positive-horizon residual state on the edge. The Bellman identity then telescopes through the first $T-1$ transitions. At the terminal transition, the only remainder is $p^* B_{T+1} = p^* \alpha \Delta_c = \alpha \Delta_v$, exactly the rounding loss in Eq. (75). The dual accounting identity is the cleaner main proof because it remains valid for non-frontier actions; the Bellman view explains why a reserve-preserving endpoint schedule has no depletion tail.

D.5 Mode-compatible confidence-width sums

Let $m_i^{\mathrm{I}}(t)$ be the number of Immediate selections of arm i before decision t, and let $M_i^{\mathrm{I}}(T) := \sum_{t=1}^{T} \mathbf{1}\{A_t = (i, \texttt{Imm})\}$ denote its total number over the horizon. Put

$$M_B := \min\left\{T, \left\lfloor \frac{B}{c_{\min}} \right\rfloor\right\}, \qquad c_{\min} := \min_i c_i. \tag{84}$$

Lemma 19 (Premium confidence-width sum)**.** *For every pathwise hard-feasible policy and $\beta_m^T = \sqrt{L_T/(2m)}$,*

$$\sum_{\substack{t: A_t = (i_t, \texttt{Imm});\\ m_{i_t}^{\mathrm{I}}(t) \geq 1}} \beta^T_{m_{i_t}^{\mathrm{I}}(t)} \leq \sqrt{2KM_B L_T}. \tag{85}$$

On the simultaneous confidence event, Immediate UCB overestimation contributes at most

$$\min\{K, M_B\} + 2\sqrt{2KM_B L_T}, \tag{86}$$

including zero-count initialization.

Proof. Every Immediate action costs at least $c_{\min}$, so hard feasibility gives

$$\sum_i M_i^{\mathrm{I}}(T) \leq \min\{T, \lfloor B/c_{\min} \rfloor\} = M_B.$$

For arm i, its nonzero-count radii along successive Immediate selections are indexed by $m = 1, \ldots, M_i^{\mathrm{I}}(T) - 1$. Therefore

$$\begin{aligned}
\sum_{\substack{t: A_t = (i_t, \texttt{Imm});\\ m_{i_t}^{\mathrm{I}}(t) \geq 1}} \beta^T_{m_{i_t}^{\mathrm{I}}(t)} &\leq \sqrt{\frac{L_T}{2}} \sum_i \sum_{m=1}^{M_i^{\mathrm{I}}(T)} \frac{1}{\sqrt{m}} \\
&\leq \sqrt{2L_T} \sum_i \sqrt{M_i^{\mathrm{I}}(T)} \\
&\leq \sqrt{2KM_B L_T},
\end{aligned}$$

using $\sum_{m=1}^{n} m^{-1/2} \leq 2\sqrt{n}$ and Cauchy–Schwarz. At most one zero-count Immediate selection occurs per selected arm, hence at most $\min\{K, M_B\}$ such widths. On the confidence event, each nonzero-count UCB overestimation is at most $2\beta_m^T$, proving Eq. (86). □

Thus premium estimation scales with the number M_B of affordable Immediate opportunities, rather than with $1/(B/T)$. This lemma alone does not prove a budget-uniform learning theorem.

For later use, Theorems 2 and 3 already imply, on the same confidence event, that the sum of Delayed UCB overestimations along any selected Delayed-action sequence is at most

$$K + 2\sqrt{2KTL_T} + K\Psi_g^{\mathrm{cert}}(T)(1 + \log T). \tag{87}$$

The first two terms are Delayed initialization and ordinary statistical radii; the last is the certified pending-envelope charge. This component is pathwise with respect to the selected action sequence and can be reused in the EQF-LADA decomposition in Section 5. The quantized-epoch analysis there controls frontier changes.

D.6 Cost conditioning and boundary audit

Known-edge tracking uses c_L, c_R only through the baseline cost and the quota conversion by Δ_c. Although $p^* = \Delta_v/\Delta_c$ and z_R^* contain divisions by Δ_c, every resource or value loss contains the cancelling products

$$B_{T+1} < \Delta_c, \qquad p^* B_{T+1} \le p^* \Delta_c = \Delta_v \le 1. \tag{88}$$

There is therefore no $1/\Delta_c$ conditioning penalty in tracking a known fixed edge. A small edge span may matter later when noisy values must *identify* which edge is active or when a large price multiplies an off-edge cost error, but it is not a resource-control loss for the fixed-edge schedule proved here. The global quantities $c_{\min}$ and $c_{\max}$ do not enter Proposition 2; $c_{\min}$ enters only the affordable-premium count M_B, and $c_{\max}$ is needed only to identify when the overall budget is inactive.

The boundary cases are as follows.

- If $c_L = 0$, the all-L reserve is free and the same quota and schedule apply; this is the important delayed-baseline edge.
- If $b = c_L$, then $z_R^* = 0$ and the schedule plays L throughout with zero leftover relative to the edge allocation. If $b = c_R$, then $z_R^* = T$ and it plays R throughout.
- If $B < c_{\min}$ or $B = 0$, any applicable active edge must have $c_L = 0$; the schedule may have $N_R^* = 0$ and remains feasible without an inverse budget-rate term.
- If $B \ge T c_{\max}$, the hard budget is inactive after replacing it by the effective cap $B' := T c_{\max}$. Budgets strictly above this cap lie outside the two-endpoint assumption in Eq. (67); their physical surplus need not be below Δ_c, but its shadow price is zero and it creates no value loss.
- Arbitrarily small $\Delta_c > 0$ and arbitrarily large p^* cause no fixed-edge loss blow-up because only $p^* \Delta_c = \Delta_v$ appears.
- If $\Delta_v = 0$, then $p^* = 0$ and rounding has zero value loss even when budget remains.
- Points with duplicated costs were merged before the frontier was formed in Theorem 1; a zero-span "edge" is not an admissible input to this section.

Hard feasibility is pathwise, terminal leftover is smaller than Δ_c, no $O(T)$ replacement tail appears, and no term depends inversely on B/T. The main algorithm replaces one globally fixed edge by quantized geometric epochs and applies a fresh virtual certificate within each epoch.

E Related Work

We compare the objectives, feedback, resource constraints, and benchmarks of related models. Rates below refer to each paper's own assumptions and comparator; they are not directly ordered against our bound. A model without a resource constraint may still use changing action probabilities, but does not address budget-constrained mixture execution.

E.1 Delayed, censored, and progressive feedback

Classical delayed bandits study late revelation of an otherwise unchanged payoff Joulani et al. (2013); Thune et al. (2019); Masoudian et al. (2024). Our setting couples revelation with value: the same request-specific delay determines when feedback arrives and discounts the payoff.

Unrestricted delayed rewards and OPSE. Lancewicki et al. Lancewicki et al. (2021), Section 4, target raw mean reward against the best arm. Completed reward is observed after a possibly reward-dependent, infinite delay. Missing values enter an all-pull zero/one sandwich; their unobserved fraction widens confidence. Theorem 5 has gap-dependent $\log T$ terms and arm-delay quantiles evaluated at gap-dependent levels. There is no resource constraint or optimistic budget-mixture execution. Our additional certificate bounds the pending *discounted* value by $u_{s,t}$ instead of one.

Payoff-determined revelation. Tang et al. Tang et al. (2024), Sections 2–3, target raw survival reward against the best arm, including rewards revealed after the horizon. Delay is its nonnegative ceiling, so non-arrival yields a lower-censoring statement about reward; their estimator truncates at a selected level. Bounds distinguish finite-mean and stronger bounded/sub-Gaussian assumptions. Zhang et al. Zhang et al. (2025), Sections 2–4, instead use linear expected payoffs and proportional payoff delay, a known maximum delay, and phased elimination with volumetric spanners; regret includes a delay term of order $D\Delta_{\max} \log T$ in the fixed-set guarantee. Both have informative pending feedback but no resource constraint or optimistic budget-mixture implementation.

Conversions and progressive observations. Vernade et al. Vernade et al. (2017), Section 2, observe labeled positive conversion events; an absent event can mean no conversion or an unfinished one. Conversion indicators and delays are conditionally independent, with a known delay CDF and an optional censoring window. The comparator is the best conversion arm and regret counts observed conversions by the horizon; confidence and rates depend on the CDF/window. Impatient Bandits Zhang et al. (2026), Sections 3–5, targets long-term reward with progressively revealed surrogate outcomes. Its Bayesian best-item comparison and Gaussian-model guarantee depend on the Value of Progressive Feedback, not a delay-independent surrogate assertion. Neither imposes a resource constraint; their informative partial observations differ from a deterministic envelope obtained solely from non-arrival.

These models make incomplete feedback informative in different ways. Our certificate follows from the known discount kernel and clock under an arbitrary within-request reward–delay law. It requires neither a known delay distribution nor surrogate observations. The distinction is the target $Rg(D)$ and its age-dependent pending bound, rather than all-submission estimation or support for unbounded delays alone. The algorithms estimate immediate rewards from immediate observations and delayed values from all delayed submissions; shared mode outcomes do not make the completed subset unbiased.

E.2 Resource constraints and mixture execution

General BwK models accommodate richer resource structures, including stochastic consumption and contextual information Agrawal & Devanur (2014; 2016). Our one-dimensional frontier uses known deterministic costs and a zero-cost alternative. The compound-action formulation captures the resource problem, while estimating pending discounted values and executing changing allocations require additional design.

General and deterministic knapsack allocation. Badanidiyuru et al. Badanidiyuru et al. (2018) observe rewards and resource consumption after selection, allowing joint stochastic resources. They compare with a known-distribution dynamic policy under exhaustion-based global budgets, with a rate of the form $\widetilde{O}(\sqrt{m\,\mathrm{OPT}} + \mathrm{OPT}\sqrt{m/B})$ for m arms and minimum budget B. This is not the rate-b prefix constraint. Tran-Thanh et al. Tran-Thanh et al. (2012) use known positive deterministic costs, immediate rewards, a budget-limited stopping horizon and a knapsack oracle, with stated gap/cost-dependent logarithmic-budget bounds. Flajolet and Jaillet Flajolet & Jaillet (2015), Section 5, explicitly handle changing optimistic bases with separate deterministic load balances; Theorem 3 gives logarithmic-budget regret with basis-gap and consumption constants. Their comparator is the known-distribution resource-constrained policy, upper-bounded by an LP. These resource models have no pending reward–delay mechanism. Their load balancing is an antecedent of mixture execution, while our certificate concerns every local prefix of a quantized geometric epoch and its concatenation.

Anytime knapsacks. SUAK Elumar et al. (2025), Sections 2–3, observes reward and stochastic cost after each pull and includes a zero-reward, zero-cost null action. The constraint is realized cumulative cost at most tb at every t; a worst-case one-step-cost check permits skipping. It compares with the mean-cost LP and updates optimistic two-arm bases and their sampling probabilities. Theorem 3.2 is instance-dependent logarithmic regret with inverse b and minimum distance of mean costs from b; this cost gap is *not* our separation between two endpoint costs. There is no delayed reward or informative pending signal. Our nonzero-value zero-premium option and known costs support a different, deterministic execution argument.

Other resource and safety constraints. Conservative bandits constrain cumulative reward relative to a default arm, with stochastic guarantees expressed in high probability or expectation Wu et al. (2016). Per-round constrained stochastic bandits constrain the expected cost of a randomized policy Pacchiano et al. (2021); this differs from realized prefix expenditure. Replenishable knapsacks allow negative resource consumption and recovery actions, with both stochastic and adversarial guarantees Bernasconi et al. (2024). These are distinct resource/safety models, not absent antecedents of online feasibility. None of these comparisons substitutes its comparator for ours.

Changing mixtures and prefix budget constraints are therefore established ideas. Our specialization combines certified discounted-value estimates, a useful zero-premium action, quantized epochs, and a local credit rule whose implementation loss is below one per binding epoch. Theorem 4 gives a gap-independent bound without inverse budget-rate or inverse endpoint cost-separation factors under this model.

E.3 LLM ROUTING, SERVING, AND PAYOFF CLOCKS

Model and prompt routing address quality and resource trade-offs Feng et al. (2025); Dai et al. (2024); Wang et al. (2025); Li et al. (2025); Atalar (2026). The closest comparisons differ in how services change over time, how costs are constrained, and what feedback is observed.

Streaming LLM routing. StageRoute Li & Li (2026), Sections 2–4, observes invocation quality and stochastic token cost, optimizes stage-wise active sets and recomputes optimistic routing mixtures. Its budget is expected average cost at most $b + o(1)$, together with probability-based throughput constraints. Its offline comparator optimizes available model sets and routing; the upper bound contains both a square-root learning term and a model-discovery term depending on stages and deployment capacity, yielding a $T^{2/3}$ order under its tuning. Revelation delay and pending-value certificates are not its observation model. Our pathwise budget guarantee therefore addresses a different constraint, while its deployment dynamics are outside our fixed-service abstraction.

Serving and other payoff clocks. FastServe Wu et al. (2026), Sections 3–5, controls token-level preemption, priority queues, and cached execution state to improve service latency. It is a systems scheduler, with no bandit regret comparator or premium-knapsack theorem; its stateful queues are not evidence for our stationarity assumption. Mersereau et al. Mersereau et al. (2009) use immediate observations and global-time discounted reward with structured means, whereas Cella and Cesa-Bianchi Cella & Cesa-Bianchi (2020) use immediately observed payoffs depending on time since an arm's last selection and compare to ranking policies. Neither uses pending completion to discount a submitted request, or implements changing premium-budget mixtures.

Serving systems motivate the service-tier interpretation but do not establish fresh stationary outcomes. Our guarantees concern the learning and allocation layer under that assumption, rather than its implementation in a stateful queue. Likewise, global-time discounting and time-since-selection payoffs use different clocks from the service delay of each submitted request.

F EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

F.1 DATA AND REPLAY PROTOCOL

We use paired binary correctness and latency records for MMLU Hendrycks et al. (2020) and HellaSwag Zellers et al. (2019), comprising 13,946 and 4,988 valid questions, respectively. The six service arms are Doubao-Seed-1.6-nothinking, Gemini 2.5 Flash, Gemini 3 Flash, GPT-3.5-turbo, GPT-5.2, and Qwen2-7B. A deterministic hash of dataset and question ID assigns 20% of questions to development and 80% to evaluation. Each round samples an evaluation question with replacement, preserving its joint outcomes across all arms. Methods share the same sampled question sequence for each seed.

Each arm has Immediate and Delayed modes. Immediate mode returns the recorded correctness with feedback available at the next decision and incurs a known premium $c_i \in [0.1, 0.3]$, obtained by linearly rescaling development-set mean costs. Delayed mode incurs no premium and yields $R\exp(-\gamma D)$, with feedback from a submission at round s arriving at $s + \lceil D \rceil + 1$. We retain

recorded delays in seconds without clipping, with one simulated second per decision round. These service modes and premiums are constructed from historical traces; the experiment is not a live comparison of provider priority tiers. Learners receive only feedback from their selected actions after arrival.

F.2 Configurations and Calibration

All reported runs have horizon $T = 5{,}000$ and use 100 paired evaluation seeds (4000–4099). The learning curves use $\gamma = 0.02$ and, for EQF, $b = B/T = 0.12$. The budget sweep uses $b \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$; the discount sweep uses $\gamma \in \{0.005, 0.01, 0.02, 0.04, 0.08\}$. All budgeted methods enforce $S_t \leq bt$ at every decision, where S_t is cumulative premium expenditure. Hyperparameters are frozen from the corresponding development configurations and are not retuned for individual evaluation budget or discount points.

F.3 Baseline Implementations

For DC, we compare **Arrived-only UCB**, **Ordered-prefix**, and **Unit-envelope**, which use completed feedback only, the completed submission prefix, or unit bounds for unresolved rewards, respectively. They provide alternative estimators within the net-utility selection task. For EQF, **Prefix-AO** applies arrived-only UCB with a feasibility mask; **PD-AO** and **PD-DC** combine primal-dual allocation with arrived-only and discount-censored estimates. **FTRL-PD** uses entropy-regularized selection with primal-dual pricing and delayed feedback updates. **PrimalDualBwK** adapts the primal-dual Bandits with Knapsacks method Badanidiyuru et al. (2018) to known resource costs, arrived-only reward estimates, and the common prefix constraint.

Both tasks also include two contextual routing adaptations. **CYANEA-Latency-TA**, based on CYANEA Li et al. (2025), ranks service–mode actions using the 15 nearest text representations and their arrived successful feedback. **PILOT-inspired**, based on PILOT Panda et al. (2025), uses per-action LinUCB with development-fitted TF-IDF/SVD features (32 components plus an intercept). It starts with $A_a = I$, $\mathbf{b}_a = 0$, and no offline reward or delay supervision; updates use only selected-action feedback after arrival. The routing adaptations use the same task objective and, for EQF, the same prefix feasibility mask. They are not full reproductions of the original routing systems, and their source-method guarantees are not asserted for these adaptations.

F.4 Metrics and Uncertainty

Let $Y_s = R_s$ for an Immediate submission and $Y_s = R_s e^{-\gamma D_s}$ for a Delayed submission, and let C_s be its premium (zero for Delayed). We plot

$$\overline{U}_t = \frac{1}{t}\sum_{s=1}^{t}(Y_s - C_s) \quad \text{for DC}, \qquad \overline{V}_t = \frac{1}{t}\sum_{s=1}^{t} Y_s \quad \text{for EQF}.$$

Thus costs are subtracted in the DC objective but enter only as constraints in EQF. Scores credit eventual outcomes to their submission rounds, including outcomes arriving after T; this evaluation convention does not expose pending outcomes to the learner. Higher values are better. We first compute each seed's average and then report its mean and sample standard deviation (ddof $= 1$). Bands and error bars are standard deviations, not confidence intervals. The reported variability is conditional on the fixed trace populations. Both populations were reused during experiment development, so new replay seeds do not constitute independent replication on new service populations.

F.5 Detailed Scores and Relative Comparisons

Table 1 reports endpoint scores for the learning curves. Tables 2 and 3 report EQF-LADA and the strongest evaluated baseline at each sweep point. All scores are means and sample standard deviations across the same 100 paired evaluation seeds. We compute relative improvement from unrounded means as $100(m_{\mathrm{LADA}} - m_{\mathrm{base}})/m_{\mathrm{base}}$, where m is endpoint average net reward for DC and endpoint average discounted reward for EQF. The strongest baseline is selected separately for each dataset and setting. Reported sweep averages weight each dataset–setting pair equally; they are averages of relative gains, not ratios of pooled rewards or averages of per-seed ratios. Rounded table entries may therefore give slightly different percentages.

Table 1: Endpoint learning performance at $T = 5{,}000$ and $\gamma = 0.02$. The EQF task uses $b = 0.12$. Values are means $\pm$ one sample standard deviation.

Task	Method	MMLU	HellaSwag
DC	DC-LADA	0.8335 ± 0.0071	0.8184 ± 0.0045
DC	Arrived-only UCB	0.8118 ± 0.0049	0.7979 ± 0.0043
DC	Ordered-prefix	0.8045 ± 0.0282	0.7991 ± 0.0101
DC	Unit-envelope	0.8114 ± 0.0049	0.7978 ± 0.0041
DC	CYANEA-Latency-TA	0.6972 ± 0.0908	0.7118 ± 0.0732
DC	PILOT-inspired	0.8449 ± 0.0063	0.8079 ± 0.0054
EQF	EQF-LADA	0.8916 ± 0.0057	0.8948 ± 0.0092
EQF	Prefix-AO	0.8699 ± 0.0110	0.8681 ± 0.0089
EQF	PD-AO	0.8416 ± 0.0055	0.8448 ± 0.0036
EQF	PD-DC	0.8414 ± 0.0055	0.8447 ± 0.0033
EQF	FTRL-PD	0.8275 ± 0.0343	0.8462 ± 0.0130
EQF	PrimalDualBwK	0.8388 ± 0.0025	0.8364 ± 0.0024
EQF	CYANEA-Latency-TA	0.7961 ± 0.0757	0.8201 ± 0.0586
EQF	PILOT-inspired	0.8704 ± 0.0122	0.8789 ± 0.0124

Table 2: Budget sweep at $\gamma = 0.02$. EQF-LADA and the strongest evaluated baseline in each setting. Scores are means $\pm$ one sample standard deviation; gains use unrounded means.

Dataset	b	Best baseline	Baseline score	EQF-LADA	Gain
MMLU	0.05	PILOT-inspired	0.8568 ± 0.0070	0.8575 ± 0.0051	0.09%
MMLU	0.1	Prefix-AO	0.8696 ± 0.0104	0.8808 ± 0.0068	1.29%
MMLU	0.15	PILOT-inspired	0.8707 ± 0.0129	0.9029 ± 0.0082	3.70%
MMLU	0.2	PILOT-inspired	0.8707 ± 0.0129	0.9023 ± 0.0092	3.63%
MMLU	0.25	PILOT-inspired	0.8707 ± 0.0129	0.9028 ± 0.0079	3.69%
MMLU	0.3	PILOT-inspired	0.8707 ± 0.0129	0.9103 ± 0.0104	4.54%
HellaSwag	0.05	PILOT-inspired	0.8378 ± 0.0061	0.8509 ± 0.0058	1.57%
HellaSwag	0.1	PILOT-inspired	0.8717 ± 0.0125	0.8823 ± 0.0086	1.21%
HellaSwag	0.15	PILOT-inspired	0.8808 ± 0.0128	0.9038 ± 0.0102	2.61%
HellaSwag	0.2	PILOT-inspired	0.8810 ± 0.0127	0.9050 ± 0.0071	2.72%
HellaSwag	0.25	PILOT-inspired	0.8810 ± 0.0127	0.9036 ± 0.0110	2.57%
HellaSwag	0.3	PILOT-inspired	0.8810 ± 0.0127	0.9131 ± 0.0079	3.64%

Discount sensitivity. The percentage decline between the endpoints of the discount sweep is $100[m(0.005) - m(0.08)]/m(0.005)$. EQF-LADA declines by 2.04% on MMLU and 1.50% on HellaSwag, compared with 3.97% and 2.63% for Prefix-AO. The relative advantage over the strongest baseline varies across the sweep, including a margin of 0.53% at $\gamma = 0.08$ on HellaSwag. Thus, retaining the highest mean across the tested settings does not imply a constant advantage as the latency penalty increases.

Interpretation and scope. The strongest evidence is the consistency of EQF-LADA's mean reward advantage across the evaluated budget and discount settings. All budgeted methods have zero recorded prefix-budget violations in the budget sweep. The ranking counts and relative gains are descriptive comparisons, not claims of statistical significance; in particular, the 0.09% budget-sweep margin on MMLU at $b = 0.05$ is a near tie. The comparisons evaluate the combined estimation and allocation designs and do not isolate causal effects of individual components. DC-LADA's contextual comparison is mixed across datasets. The results concern frozen practical configurations and the evaluated baseline set, rather than live priority-tier services or new service populations.

Table 3: Discount sweep at $b = 0.12$. EQF-LADA and the strongest evaluated baseline in each setting. Scores are means $\pm$ one sample standard deviation; gains use unrounded means.

Dataset	γ	Best baseline	Baseline score	EQF-LADA	Gain
MMLU	0.005	Prefix-AO	0.8829 ± 0.0049	0.8989 ± 0.0063	1.82%
MMLU	0.01	Prefix-AO	0.8793 ± 0.0072	0.8952 ± 0.0079	1.80%
MMLU	0.02	PILOT-inspired	0.8704 ± 0.0122	0.8916 ± 0.0057	2.43%
MMLU	0.04	PILOT-inspired	0.8679 ± 0.0165	0.8862 ± 0.0076	2.11%
MMLU	0.08	PILOT-inspired	0.8666 ± 0.0222	0.8805 ± 0.0050	1.61%
HellaSwag	0.005	PILOT-inspired	0.8869 ± 0.0080	0.8996 ± 0.0074	1.44%
HellaSwag	0.01	PILOT-inspired	0.8844 ± 0.0071	0.8989 ± 0.0092	1.64%
HellaSwag	0.02	PILOT-inspired	0.8789 ± 0.0124	0.8948 ± 0.0092	1.81%
HellaSwag	0.04	PILOT-inspired	0.8759 ± 0.0161	0.8891 ± 0.0096	1.51%
HellaSwag	0.08	PILOT-inspired	0.8816 ± 0.0119	0.8862 ± 0.0063	0.53%