TO DEFER OR NOT TO DEFER? ONLINE LEARNING WITH LLMS AND CONGESTED EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern AI systems increasingly combine large language models (LLMs) with human experts, using the LLM to handle routine queries while deferring uncertain cases for expert review. When expert capacity is limited, however, deferral decisions must account not only for the reliability of the LLM but also for congestion in the expert queue. We formulate this as an online sequential decision-making problem in which, for each query, the system observes the confidence signal of the LLM's answer and must decide whether to output the answer or defer the query to a capacity-constrained expert. The reliability of the confidence signal is initially unknown and must be learned from expert feedback, which is both selectively observed (only deferred queries are labeled) and delayed by queuing. We propose `KL-PBD`, a learning algorithm that uses Kullback–Leibler lower confidence bounds on the unknown reliability parameters to guide deferral decisions. For K confidence levels and horizon T, `KL-PBD` achieves $O(K \log T)$ gap-dependent regret when the accuracy at every confidence level is bounded away from the deferral boundary, and $\widetilde{O}(\sqrt{KT})$ worst-case regret. We establish matching lower bounds, showing that these rates are optimal up to logarithmic factors. Finally, we validate our approach on synthetic instances and real-world question-answering datasets, namely GSM8K, NuminaMath, and MedMCQA, using Qwen LLMs.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed in settings where automated predictions can be complemented by human expertise, including customer support and healthcare (Huang et al., 2025; Feng et al., 2024; Yang et al., 2025b; Bhattacharya et al., 2025). A natural design is to let the LLM handle routine queries while referring uncertain or high-risk cases to a human expert. Expert attention, however, is limited: reviewing a query takes time, and queries sent for review may have to wait behind previously deferred requests. Consequently, a deferral decision must account not only for whether the LLM is likely to be correct, but also for the congestion created by sending another query to the expert queue. Determining when the LLM is reliable is itself challenging: the correctness of a particular answer is not known unless a ground-truth label is obtained, which in our setting requires expert review. LLMs nevertheless provide signals that may be informative about their reliability: a prominent example is self-consistency (Wang et al., 2023), which measures agreement across multiple generated answers. Such signals can be correlated with correctness, but high confidence or agreement does not guarantee that the answer is correct (Aggarwal et al., 2023; Wang et al., 2024). Thus, a deferral system must learn how the observed confidence signal relates to the probability that the LLM's answer is correct.

We are interested in doing so *online*, without requiring a separate task-specific calibration dataset or retraining the LLM before deployment. Instead, the system learns from expert feedback generated during operation. This gives rise to a coupled learning-and-control problem with an unusual feedback structure. If the system accepts an LLM response, it observes no ground-truth label. If it defers, the label is revealed only after the query has been served by the expert. Feedback is therefore both *selective* and *delayed*. Moreover, acquiring information is itself costly: every deferred query consumes expert capacity, enters the queue, and may increase congestion for future queries. Deferral is therefore simultaneously a prediction decision, an information-acquisition decision, and a queue-control decision. We formalize this interaction as a finite-horizon online sequential decision

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.

problem. At each round, a query arrives together with the LLM's answer and a confidence level, and the system decides whether to accept the answer or defer the query to a capacity-constrained expert queue. The conditional accuracy associated with each confidence level is initially unknown and must be learned from completed expert reviews. Our goal is to design an online policy that minimizes the expected cumulative cost over the horizon, balancing the risk of incorrect LLM answers, the congestion created by expert review, and the value of deferrals as a source of information for future decisions.

Our contributions are as follows.

(a) We propose `KL-PBD`, a queue-aware algorithm that learns the unknown conditional accuracies from delayed expert feedback. The algorithm acts optimally for a pessimistic estimate of the LLM's reliability, favoring deferral, and hence information acquisition, when the available evidence is still limited, while explicitly accounting for the congestion cost of doing so.

(b) We establish near-optimal regret guarantees for `KL-PBD`. For K confidence levels, the algorithm achieves $O(K \log T/\gamma)$ regret when the accuracy at every confidence level is at least γ away from the deferral boundary, and $\widetilde{O}(\sqrt{KT})$ regret in the worst case.

(c) We derive complementary lower bounds showing that logarithmic regret and inverse-gap dependence are unavoidable in separated instances, while the worst-case $\sqrt{KT}$ dependence is optimal up to logarithmic factors.

(d) We evaluate `KL-PBD` on synthetic and real-world question-answering tasks based on GSM8K, NuminaMath, and MedMCQA. We demonstrate the importance of pessimism under selective feedback and show that explicitly accounting for the expert queue is paramount when expert service is congested (see Fig. 1 (right) for an illustrative example).

2 Related Work

Learning to defer. Learning to defer studies systems that either make an automated prediction or delegate the decision to a human expert (Madras et al., 2018; Mozannar & Sontag, 2020). These foundational formulations jointly learn a predictor and a deferral rule, with subsequent work developing calibrated deferral objectives (Verma & Nalisnick, 2022). Post-hoc approaches instead learn when to defer the predictions of a fixed model (Narasimhan et al., 2022). These approaches are formulated as offline supervised learning problems, where the deferral policy is learned from a fixed labeled dataset and evaluated through test-time prediction risk. In our problem, the LLM response rule is fixed, but the system learns a deferral policy online from delayed, selectively observed human feedback generated by its own deferral decisions. Moreover, expert intervention is capacity constrained, making the cost of deferral state dependent rather than fixed. Further supervised and multi-expert formulations are discussed in Appendix D.3.

Online learning and selective feedback. Reid et al. (2024) study online deferral under a total consultation budget. Their bandit feedback model reveals the reward of the selected action, including automated predictions. Online selective classification is closer to our setting: labels are observed only when the learner abstains from answering, and regret guarantees trade off prediction errors against selective label acquisition (Gangrade et al., 2021; Baharav et al., 2026). In our setting, labels are similarly observed only on deferrals, but a deferred query first enters the expert queue, and its label becomes available only once it has been served. Broader active-learning and oracle-query settings are discussed in Appendix D.6.

Deferral under limited review capacity. Alves et al. (2024) consider cost-sensitive deferral to multiple experts under individual workload constraints, optimizing assignments using offline estimates of expert error. More directly related, Lykouris & Weng (2025) learn a binary classifier online while controlling admission to human review and scheduling; as in our model, feedback is selective and becomes available only after a review is completed. In this work, by contrast, the LLM response rule is fixed and the order of service is not controlled by the learner; what is learned is the LLM's accuracy at each self-confidence level, which determines the deferral policy. Their benchmark allocates review capacity using expected arrival and service rates and assumes admitted jobs

are reviewed immediately. We instead compare with the optimal finite-horizon policy subject to the same queue dynamics as the learner, with expert-review and holding costs included in the objective.

Additional related work on uncertainty signals and probabilistic calibration, LLM routing, structured admission-control policies, and human–AI decision support is discussed in Appendix D.

3 Problem formulation and optimal deferral

3.1 LLM–expert system model

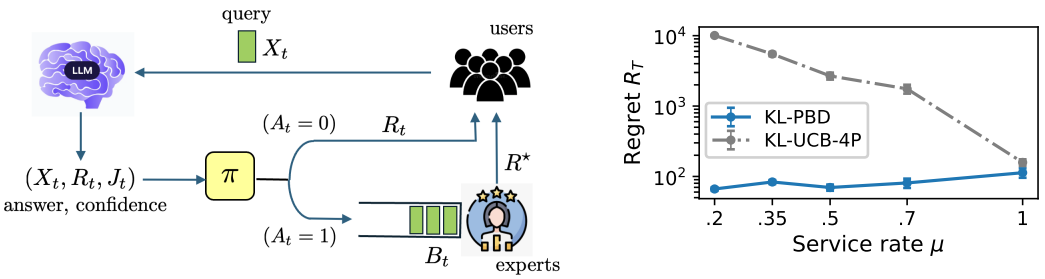


Figure 1: Left: expert-aided LLM system with deferral policy π and delayed feedback from completed expert reviews. Right: regret after $T = 5 \times 10^4$ queries on MedMCQA with Qwen3-4B against the expert service rate μ, comparing the queue-aware `KL-PBD` with the queue-unaware `KL-UCB-4P`. For $\mu = 1$, deferred queries are served before the next round and the queue does not build up, so queue awareness provides no benefit. For $\mu < 1$, deferred queries accumulate and create congestion, making queue-aware deferral essential for low regret. Details in Appendix H.3.

We consider a sequential deferral problem over a known horizon T. At each round $t = 1, \dots, T$, a user query X_t may arrive. Upon receiving a query, the LLM produces a candidate answer R_t together with an index $J_t \in [K] := \{1, \dots, K\}$, which we call its *self-confidence level*. Our model is agnostic to how (R_t, J_t) is generated. A canonical example is self-consistency (Wang et al., 2023): the LLM generates $K - 1$ answers, R_t is chosen from the largest answer-equivalence class, and J_t is one plus the size of that class, so that $J_t = 1$ when no answer can be extracted. Alternative constructions of (R_t, J_t) are discussed in Appendix D.1. Let $R_t^\star$ denote the ground-truth answer and define $Y_t := \mathbf{1}\{R_t \equiv R_t^\star\}$, where $R_1 \equiv R_2$ means that the two answers are equivalent. We set $J_t = 0$, $X_t = R_t = \emptyset$, $Y_t = 1$ when no query arrives and assume that $(J_t, Y_t)_{t \geq 1}$ is an i.i.d. sequence. For $j \in [K] \cup \{0\}$, let $\omega_j := \mathbb{P}(J_t = j)$, and for $j \in [K]$, define[1] $h_j^\star := \mathbb{P}(Y_t = 1 \mid J_t = j)$. Thus, $h_j^\star$ is the LLM's conditional accuracy at self-confidence level j, and we call $h^\star = (h_1^\star, \dots, h_K^\star)$ the *conditional accuracy vector*.

When a query arrives, the decision maker, who observes the self-confidence level J_t but neither X_t nor R_t, either accepts the LLM answer or defers the query to a queue served by a pool of experts, who always provide the correct answer. We encode this decision by $A_t \in \{0, 1\}$, where $A_t = 1$ denotes deferral. We set $A_t = 0$ if $J_t = 0$. Let B_t be the number of queries in the queue at the beginning of round t, and let C_t be the number of queries that can be served by the experts between rounds t and $t + 1$. We assume that $(C_t)_{t \geq 1}$ is i.i.d. with distribution ρ_C and independent of $(J_t, Y_t)_{t \geq 1}$ and the algorithm's internal randomization, and that $\mu := \mathbb{P}(C_t \geq 1) > 0$. The order in which pending queries are reviewed may depend on the observations and the service capacity, but not on unrevealed ground-truth labels. The queue is initially empty, $B_1 = 0$, and has known capacity $\bar{B} \geq 1$, so deferral is infeasible when $B_t = \bar{B}$. Its dynamics are $B_{t+1} = \max\{0, B_t + A_t - C_t\}$.

The cost incurred at round t is $c_t := c_{\text{hold}} B_t + c_{\text{exp}} A_t + c_e(1 - A_t)(1 - Y_t)$, where $c_{\text{hold}} \geq 0$ is the per-round holding cost of each query in the queue, $c_{\text{exp}} \geq 0$ is the cost of deferring a query to the experts, and $c_e > 0$ is the cost of returning an incorrect LLM answer. At the end of the horizon, we account for the cost of serving the queries remaining in the queue. These cost parameters are known to the decision maker. Let $D(b)$ denote the expected holding cost of clearing a queue of size b. It

[1] If $\omega_j = 0$, $h_j^\star$ may be assigned arbitrarily.

satisfies $D(0) = 0$ and, for $b \geq 1$, $D(b) = c_{\text{hold}} b + \mathbb{E}_{C \sim \rho_C}\big[D\big((b - C)^+\big)\big]$. The terminal cost is therefore $D(B_{T+1})$. For example, if $C \sim \text{Bernoulli}(\mu)$, then $D(b) = c_{\text{hold}} b(b+1)/(2\mu)$.

The resulting problem is a finite-horizon Markov decision process (MDP) with state (B_t, J_t), action A_t, per-round cost c_t, and terminal cost $D(B_{T+1})$. A policy $\pi = (\pi_t)_{t=1}^T$ is a sequence of possibly randomized decision rules, where π_t maps the current state (B_t, J_t) to a distribution over feasible actions. Our goal is to learn, online, a policy that performs nearly as well as the optimal policy that knows the conditional accuracy vector $h^\star$. We formalize this objective through regret after first characterizing the optimal policy under known $h^\star$.

3.2 Optimal deferral policy

Fix $h \in [0,1]^K$ and set $h_0 := 1$. For $j \in [K] \cup \{0\}$, define $L^h(j) := c_e(1 - h_j)$, which, for $j \in [K]$, is the expected error cost of accepting an LLM response with self-confidence level j. Let V_t^h and Q_t^h denote the finite-horizon value and state-action value functions of the MDP induced by h, with terminal value $V_{T+1}^h((b,j)) = D(b)$. In other words, $Q_t^h((b,j),a)$ is the expected cost-to-go from state (b,j) when action a is taken at round t, and $V_t^h((b,j))$ is its minimum over feasible actions.

To simplify the notation, let $P_a(b,b')$ denote the transition probability from queue state b to b' under action $a \in \{0,1\}$. Since the queue dynamics depend on the current queue state and action, but not on the current self-confidence level, define, for $b < \bar{B}$,

$$M_t^h(b) := c_{\text{exp}} + \sum_{b'=0}^{\bar{B}} \big(P_1(b,b') - P_0(b,b')\big) \sum_{j=0}^{K} \omega_j V_{t+1}^h((b',j)). \tag{1}$$

We refer to $M_t^h(b)$ as the *marginal cost of deferral*: it captures both the immediate expert-review cost and the expected increase in future cost caused by adding a query to the expert queue. We adopt the convention $M_t^h(\bar{B}) := +\infty$.

For any $b < \bar{B}$ and $j \in [K]$, the difference between the costs of deferring and accepting is

$$Q_t^h((b,j),1) - Q_t^h((b,j),0) = M_t^h(b) - L^h(j).$$

Hence, the optimal decision compares the expected error cost of the LLM response with the marginal cost of deferral. The optimal policy has a threshold structure: a query is deferred whenever the LLM's conditional accuracy is below a threshold determined by the current time and expert backlog. As stated in the following lemma, the threshold is endogenous: through $M_t^h(b)$, it accounts for the future congestion created by the current deferral decision.

Lemma 1 (Optimal deferral policy)**.** *Fix $h \in [0,1]^K$. Resolving ties in favor of acceptance, an optimal action at round t is $A_t^h = \mathbf{1}\{J_t \in [K],\ L^h(J_t) > M_t^h(B_t)\}$. Equivalently, defining the accuracy thresholds $\tau_t^h(b) := 1 - M_t^h(b)/c_e$, for all $b \leq \bar{B}$, the optimal policy defers if and only if $h_{J_t} < \tau_t^h(B_t)$. In particular, $\tau_t^h(\bar{B}) = -\infty$, so deferral is infeasible when the queue is full.*

3.3 Learning problem and regret decomposition

We now turn to the setting in which the true conditional accuracy vector $h^\star$ is unknown. The statistical challenge is to learn $h^\star$ while making deferral decisions. Feedback is both *selective* and *delayed*. If the learner accepts the LLM response ($A_t = 0$), its ground-truth label Y_t is not revealed, and hence the corresponding error cost is not observed. If the learner defers ($A_t = 1$), the ground-truth label becomes available only once the query is processed by an expert. Thus, information about $h^\star$ can be acquired only through the capacity-constrained experts, and the resulting feedback is delayed by the same queue that governs the deferral decisions. In particular, exploration affects not only the information available to the learner but also the future state of the system through congestion.

To isolate this learning challenge, our main analysis assumes that the distributions of the confidence levels ω and of the service capacity ρ_C are known. This assumption does not require analogous active exploration: the self-confidence level J_t is observed whenever a query arrives, so ω can be estimated passively from arrival frequencies, while observed service capacities provide samples from ρ_C independently of the deferral decisions. In Appendix B.6, we extend our analysis to the setting where ω and ρ_C are unknown and show that our worst-case regret guarantees are preserved.

A learning algorithm π is analogous to a policy, except that since it learns on the way, the decision taken at time t may depend on *all* observations made until time t. The expected regret of an algorithm π is defined by:

$$R_T(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{T} c_t + D(B_{T+1})\right] - \mathbb{E}^{\pi^\star}\left[\sum_{t=1}^{T} c_t + D(B_{T+1})\right].$$

where $\pi^\star$ is the optimal deferral policy for the conditional accuracy vector $h^\star$, identified in Lemma 1. Note that the oracle, applying $\pi^\star$, and the learner use the same K self-confidence levels; hence, regret measures only the additional cost incurred from learning the unknown conditional accuracy vector $h^\star$, rather than any approximation error induced by the choice of confidence discretization.

For brevity, a superscript $\star$ on a quantity denotes evaluation at $h = h^\star$ (e.g. $L^\star = L^{h^\star}$). Define the instantaneous decision gap $\Delta_t := \left|L^\star(J_t) - M_t^\star(B_t)\right| \mathbf{1}\{A_t \neq A_t^\star\}$ when $J_t \in [K]$ and $B_t < \bar{B}$, and set $\Delta_t := 0$ otherwise.

Lemma 2 (Regret decomposition)**.** *For any learning algorithm π, $R_T(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{T} \Delta_t\right]$.*

Lemma 2 shows that regret is the expected cumulative decision gap over the rounds in which the learner takes a strictly suboptimal action. Consequently, the learning problem reduces to estimating $h^\star$ accurately enough to identify on which side of the time- and backlog-dependent oracle threshold $\tau_t^\star(B_t)$ each self-confidence level lies.

4 The KL−PBD algorithm and its regret

4.1 Pessimistic Bellman deferral

We now introduce our algorithm KL−PBD, which learns the unknown conditional accuracy vector $h^\star$ from delayed expert feedback while accounting for the congestion induced by deferral. KL−PBD is summarized in Algorithm 1.

For each self-confidence level $j \in [K]$, let $N_{j,t}$ denote the number of level-j queries whose expert reviews have been *completed* before the decision at round t. Importantly, $N_{j,t}$ need not equal the number of level-j queries deferred before t: a deferred query contributes to $N_{j,t}$ only once it is served and its ground-truth label becomes available. The selective and delayed nature of the feedback thus enters the algorithm through $N_{j,t}$. Let $Y_{j,n}$ denote the label revealed by the nth completed level-j review, and, for $n \geq 1$, define $\widehat{h}_{j,n} := \frac{1}{n}\sum_{\ell=1}^{n} Y_{j,\ell}$. For $T \geq 3$, let $\phi(T) := \log T + \log\log T$ denote the confidence budget. Given $N_{j,t} \geq 1$ observed labels, we form the KL lower confidence bound

$$\underline{h}_{j,t} := \inf\left\{q \in [0, \widehat{h}_{j,N_{j,t}}] : N_{j,t}\ \mathrm{kl}\Big(\widehat{h}_{j,N_{j,t}}, q\Big) \leq \phi(T)\right\}$$

where kl denotes the Bernoulli relative entropy, and set $\underline{h}_{j,t} = 0$ when $N_{j,t} = 0$. At round t, KL−PBD treats $\underline{h}_t$ as the current conditional accuracy vector and applies the corresponding optimal policy (Algorithm 1, lines 5–6). Equivalently, the learner defers whenever the pessimistic expected error cost of the LLM exceeds the Bellman marginal cost of expert review.

Algorithm 1 KL−PBD

1: Set $\underline{h}_{j,1} = 0$ for every $j \in [K]$.
2: **for** $t = 1, \ldots, T$ **do**
3: Incorporate all expert labels received before round t and update $N_{j,t}$, $\widehat{h}_{j,N_{j,t}}$, and $\underline{h}_{j,t}$.
4: Observe (B_t, J_t).
5: Compute $\widehat{\tau}_t(B_t) = 1 - M_t^{\underline{h}_t}(B_t)/c_e$.
6: Choose $A_t = \mathbf{1}\big\{J_t \in [K],\ \underline{h}_{J_t,t} < \widehat{\tau}_t(B_t)\big\}$.
7: **end for**

The use of a lower confidence bound serves both learning and control. Because correctness is observed only for deferred queries, accepting a query provides no information about $h^\star$. A plug-in estimate can therefore become prematurely optimistic about a self-confidence level, causing the learner to accept subsequent queries from that level and thereby suppress the feedback needed to correct the estimate (see Appendix H.5).

In contrast, KL−PBD treats uncertain levels pessimistically: before sufficient expert feedback has accumulated, a smaller $\underline{h}_{j,t}$ makes deferral more attractive and hence promotes the collection of

additional labels. This exploration is nevertheless congestion-aware, since the decision is made through the threshold $\widehat{\tau}_t(B_t)$ and therefore accounts for the current queue and the future cost of expert review.

4.2 Regret guarantees

We first give a gap-dependent regret guarantee. Recall that the oracle defers a level-j query at state (t, b) when $h_j^\star < \tau_t^\star(b) = 1 - M_t^\star(b)/c_e$. The distance $\left|h_j^\star - \tau_t^\star(b)\right|$ therefore measures how far the conditional accuracy of level j is from the oracle threshold at time t and queue length b. Equivalently, for $b < \bar{B}$, $|L^\star(j) - M_t^\star(b)| = c_e \left|h_j^\star - \tau_t^\star(b)\right|$ is the difference between the oracle Q-values of accepting and deferring. Thus, the farther $h_j^\star$ is from the threshold, the less accurately it needs to be estimated to identify the oracle action.

Importantly, states at which $h_j^\star = \tau_t^\star(b)$ should not be treated as having a problematic "zero gap." At such a state, $L^\star(j) = M_t^\star(b)$, so accepting and deferring have the same oracle Q-value. Choosing either action therefore incurs no regret. Accordingly, the relevant notion of separation is the smallest *nonzero* distance from an oracle threshold. To formalize this, define the set of nonzero oracle gaps over the horizon by

$$\mathcal{G}_T := \left\{ \left|h_j^\star - \tau_t^\star(b)\right| : j \in [K],\ t \le T,\ b < \bar{B},\ h_j^\star \neq \tau_t^\star(b) \right\}.$$

If $\mathcal{G}_T = \emptyset$, every feasible oracle comparison is a tie. Hence the decision gap in Lemma 2 is zero at every round, and every algorithm has zero regret. We therefore restrict attention to the nontrivial case $\mathcal{G}_T \neq \emptyset$ and define $\gamma_T := \min\{1, \min \mathcal{G}_T\}$. This quantity captures the smallest separation that the learner must resolve in order to distinguish a strict oracle preference between accepting and deferring. When this separation is bounded away from zero, sufficiently accurate estimates of $h^\star$ identify the oracle action.

Theorem 1 (Gap-dependent regret of KL-PBD)**.** *Suppose that $\mathcal{G}_T \neq \emptyset$. For every $T \ge 3$,*

$$R_T(\pi^{\texttt{KL-PBD}}) \le K\kappa\phi(T) + 2c_e \sum_{j=1}^{K} \min\left\{ \omega_j T, \bar{B} + \frac{4\phi(T)}{\gamma_T} \right\} + 2\bar{B}c_e \left[1 + 4 \sum_{j=1}^{K} \log\left(1 + \frac{\omega_j T}{\phi(T)}\right) \right]$$

where $\kappa := 4e(\max\{c_e, c_{\exp}\} + c_{\text{hold}}\bar{B}/\mu)$.

In particular, if $\gamma_T \ge \gamma > 0$ uniformly in T, then $R_T(\pi^{\texttt{KL-PBD}}) = O(\frac{K \log T}{\gamma})$. Thus, when the conditional accuracies remain separated from the oracle thresholds, KL-PBD incurs only logarithmic regret. A stronger version of Theorem 1 is proven in Appendix B.4.

We next give a gap-free guarantee, which holds without any separation from the oracle thresholds. A stronger version of Theorem 2 is proven in Appendix B.3.

Theorem 2 (Worst-case regret of KL-PBD)**.** *For every $T \ge 3$,*

$$R_T(\pi^{\texttt{KL-PBD}}) \le K\kappa\phi(T) + c_e \sum_{j=1}^{K} \min\left\{ \omega_j T, \bar{B} + 2\sqrt{2\omega_j T \phi(T)} \right\} + \bar{B}c_e \left[1 + 2 \sum_{j=1}^{K} \log\left(1 + \frac{\omega_j T}{\phi(T)}\right) \right].$$

Up to logarithmic factors, Theorem 2 gives $R_T(\pi^{\texttt{KL-PBD}}) = \widetilde{O}(K + \sum_{j=1}^{K} \min\{\omega_j T, \sqrt{\omega_j T}\})$. In particular, the worst case over arrival distributions yields $\widetilde{O}(\sqrt{KT})$ for $K \le T$ (Remark 1).

The main difficulty in the regret analysis is that $\underline{h}_t$ also enters the marginal cost of deferral, so pessimism may cause KL-PBD not only to defer queries that the oracle accepts, but also to accept queries that the oracle defers. For a high-level summary of our proof technique, see Appendix B.1.

5 Lower bounds

We now show that the dependence on K, T, and the gaps in Theorems 1 and 2 is optimal in general (up to logarithmic factors for worst-case regret) by deriving matching lower bounds, which are

proven in Appendix C. Write $R_T(\pi; h)$ for the regret when the conditional accuracy vector is h, and define $\tau := 1 - [c_{\exp} + (1-\mu)c_{\text{hold}}/\mu]/c_e$. This is the largest accuracy threshold the oracle ever uses: $\tau_t^h(b) \le \tau$ for all h, t and b, with equality at the empty queue in the last round (Lemma 18).

We first give an instance-dependent lower bound. We call an algorithm π uniformly efficient if, run with horizon T, it satisfies $R_T(\pi; h) = o(T^\alpha)$ for every $h \in [0,1]^K$ and every $\alpha > 0$. Let $\overline{\tau}^\star := \lim_{T-t\to\infty} \tau_t^\star(0)$ be the long-horizon oracle threshold at the empty queue, which exists by Lemma 18. For every t and b, $\tau_t^\star(b) \le \tau_t^\star(0)$ (Proposition 1) and $\overline{\tau}^\star \le \tau_t^\star(0) \le \tau$ (Lemma 18). Thus, $\overline{\tau}^\star$ is the long-horizon limit of the largest threshold the oracle uses.

Theorem 3 (Instance-dependent lower bound)**.** *If $\overline{\tau}^\star \in (0,1)$, every uniformly efficient algorithm π satisfies*

$$\liminf_{T\to\infty} \frac{R_T(\pi; h^\star)}{\log T} \ge c_e \sum_{\substack{j\in[K]:\\ h_j^\star > \overline{\tau}^\star,\ \omega_j > 0}} \frac{h_j^\star - \overline{\tau}^\star}{\mathrm{kl}(h_j^\star, \overline{\tau}^\star)}. \tag{2}$$

As the positive gaps $h_j^\star - \overline{\tau}^\star$ tend to zero, with $\overline{\tau}^\star$ bounded away from 0 and 1, each summand in the lower bound scales as $2\overline{\tau}^\star(1-\overline{\tau}^\star)/(h_j^\star - \overline{\tau}^\star)$. Hence, the logarithmic growth in T and inverse-gap dependence of Theorem 1 are necessary in general. When $\mu = 1$, $\overline{\tau}^\star = \tau$ and `KL-PBD` attains equation 2 with equality (Proposition 3). Additionally, the stronger version of Theorem 1 proved in Appendix B.4 matches the dependence of Theorem 3 on the conditional accuracies for small gaps.

Our proof relies on a change-of-measure argument, as in Lai & Robbins (1985). The main difficulty in our setting is that the marginal cost of deferral depends on the round and the queue state, and that the oracle threshold changes with the instance. For a high-level summary of our proof technique, see Appendix C.1.

We next give a minimax lower bound.

Theorem 4 (Minimax lower bound)**.** *Assume $\tau \in (0,1)$. For every $T \ge 1$,*

$$\inf_{\pi} \sup_{h\in[0,1]^K} R_T(\pi; h) \ge \frac{1}{256} \sum_{j=1}^{K} \min\Big\{ \mu^3 c_e \min\{\tau, 1-\tau\} \min\{\omega_j T, \sqrt{\omega_j T}\},\ \mu^2 c_{\text{hold}}\, \omega_j T \Big\}. \tag{3}$$

When $\mu = 1$, the last term can be dropped and the constant $1/256$ improved to $1/12$.

If $c_{\text{hold}} > 0$ or $\mu = 1$, the lower bound scales as $\sum_{j=1}^K \min\{\omega_j T, \sqrt{\omega_j T}\}$. Thus, its dependence on T and ω agrees with Theorem 2 up to logarithmic factors and the upper bound's additive K term. The lower bound is largest for balanced levels, $\omega_0 = 0$ and $\omega_j = 1/K$ (Remark 3), giving $\min\{T, \sqrt{KT}\}$ and matching the upper bound up to logarithmic factors for $K \le T$.

The main difficulty in the minimax lower-bound construction is that hard instances cannot be obtained by perturbing the conditional accuracies around a fixed threshold, since the threshold itself depends on them, and that nearby instances are only guaranteed to have different optimal actions when the queue is empty. For a high-level summary of our proof technique, see Appendix C.1.

6 Numerical experiments

We now compare `KL-PBD` with baselines and examine the structure of the oracle policy on instances built from LLM outputs on GSM8K, NuminaMath-TIR and MedMCQA. Additional results and details are deferred to Appendices G and H.

Instances. We use two LLMs, Qwen3-4B-Instruct-2507 and Qwen3-0.6B, and build one instance per dataset–LLM pair; results for Qwen2.5-0.5B-Instruct are reported in Appendix H. Each instance uses the self-consistency construction of Section 3: for every question, the LLM generates 20 answers, so there are $K = 21$ levels. For a query at level j, let p_j denote its self-consistency score, i.e., the fraction of generated answers that agree with R_t, so that $p = (p_1, \dots, p_K) = (0, .05, \dots, 1)$, where $p_1 = 0$ corresponds to queries for which no answer could be extracted. We take ω_j and $h_j^\star$ to be the empirical frequency and conditional accuracy of self-consistency level j over the whole

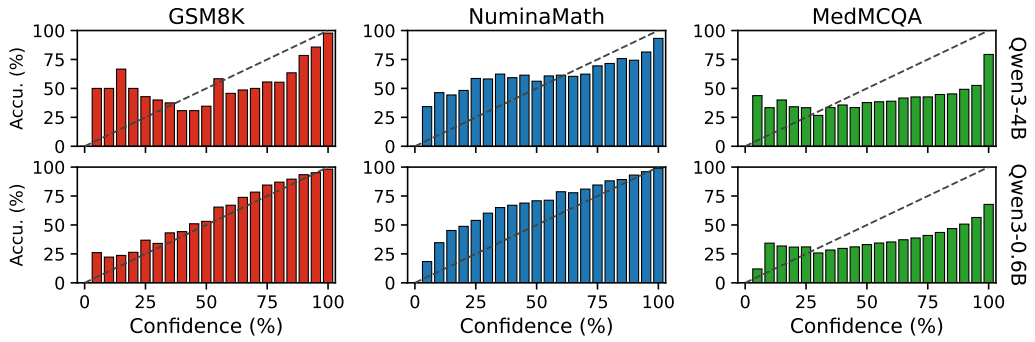


Figure 2: Conditional accuracy of the self-consistency score (dashed line: perfect calibration).

dataset, and sample queries independently with replacement from that dataset, so that a query arrives at every round ($\omega_0 = 0$). Regret is evaluated against the optimal policy for this empirical instance. Figure 2 shows three calibration regimes. On GSM8K and NuminaMath the score is informative: Qwen3-0.6B is especially close to calibrated on GSM8K, while Qwen3-4B is underconfident at low scores and overconfident at high scores. On MedMCQA, the score is nearly uninformative: accuracy barely varies with the score below 100%, and the 100% level has accuracy only .79 (Qwen3-4B) and .68 (Qwen3-0.6B).

Setup and baselines. We set $(c_e, c_{\text{exp}}, c_{\text{hold}}) = (1, 0.25, 0.03)$, $\bar{B} = 8$, $C_t \sim \text{Bernoulli}(0.35)$, and $T = 10^5$. The arrival distribution ω is unknown to the algorithms and is estimated as $\hat{\omega}_{j,t} = t^{-1} \sum_{s=1}^{t} \mathbf{1}\{J_s = j\}$ by all methods that require it, except `FQT`, which is tuned under a uniform ω. We compare `KL-PBD` with: i) `SALMUT-SF`, a threshold-based actor–critic policy adapted from Roy et al. (2021) to handle unknown rewards and selective feedback; ii) Fixed Queue Threshold (`FQT`), a non-learning baseline that uses a queue threshold optimized under the surrogate $h = p$; iii) variants of `KL-PBD`, namely `KL-PBD+`, which uses a more aggressive confidence budget inspired by KL-UCB+ type rules (Garivier & Cappé, 2011; Achab et al., 2018), Greedy Bellman Deferral (`GBD`), which replaces the lower confidence bound $\underline{h}_t$ by the empirical mean $\widehat{h}_{j,N_{j,t}}$, and Naive Bellman Deferral (`NBD`), which freezes the estimate at $h = p$, i.e., treats the score as calibrated. These baselines are described in more detail in Appendix F.

Regret of `KL-PBD` against baselines. In Figure 3, we report regret averaged over 100 independent runs with pointwise 95% confidence bands. Solid lines are algorithms that learn $h^\star$, and dashed lines are the baselines using the $h = p$ surrogate. The results illustrate the impact of score calibration: on GSM8K and NuminaMath, `NBD` is generally the best non-learning baseline and competes with `KL-PBD` over this horizon, whereas on MedMCQA it incurs 20 to 65 times the regret of `KL-PBD+`. Among learning algorithms, `GBD` suffers from the selective-feedback failure described in Appendix H.5, and `SALMUT-SF` adapts too slowly to be competitive. `KL-PBD+` attains the lowest regret in all six instances, up to a factor 4 below `KL-PBD`: its more aggressive level-dependent confidence budget reduces the conservatism of the deferral rule. Appendix H reports regret values at horizons $T \in \{10^4, 5 \cdot 10^4, 10^5, 2 \cdot 10^5\}$; the gap between `KL-PBD+` and the $h = p$ baselines keeps widening with T. Comparisons to queue-unaware baselines, including adaptations of bandit algorithms from Achab et al. (2018), are deferred to Appendix H.3: `KL-PBD` and `KL-PBD+` generally beat them by a wide margin when $\mu < 1$, and remain competitive under immediate service ($\mu = 1$).

Structure of the oracle policy. Figure 4 shows the oracle threshold and the resulting decisions for Qwen3-0.6B. For every μ, the threshold decreases with the queue state, as established in Proposition 1. Higher values of μ lead to higher deferral thresholds, as faster expert service reduces the congestion cost of deferral: the oracle then defers even queries on which the LLM is fairly accurate. The bottom row shows the oracle deferral decisions for the service rate $\mu = 0.35$ used in the regret experiments. Accordingly, the oracle defers fewer levels as the queue becomes more congested. In contrast, it does not always defer fewer levels as self-consistency increases: since the oracle defers exactly the levels whose conditional accuracy lies below the threshold, this occurs be-

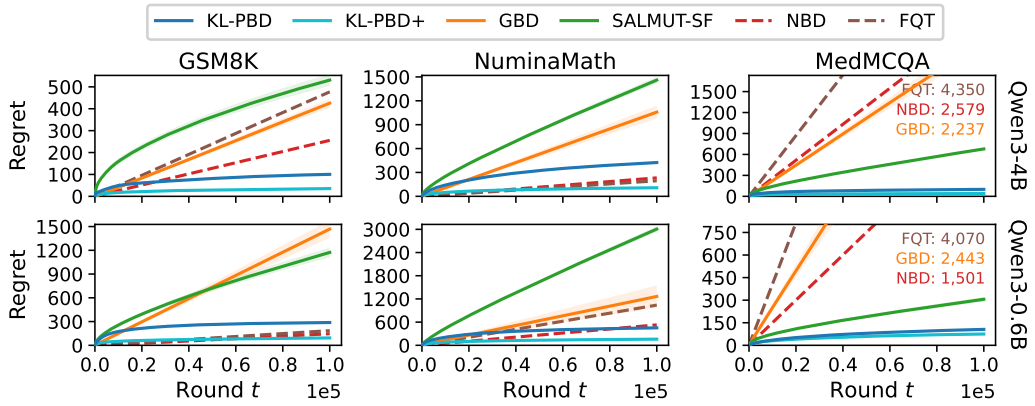


Figure 3: Regret averaged over 100 runs. Curves leaving the frame are listed with their final regret.

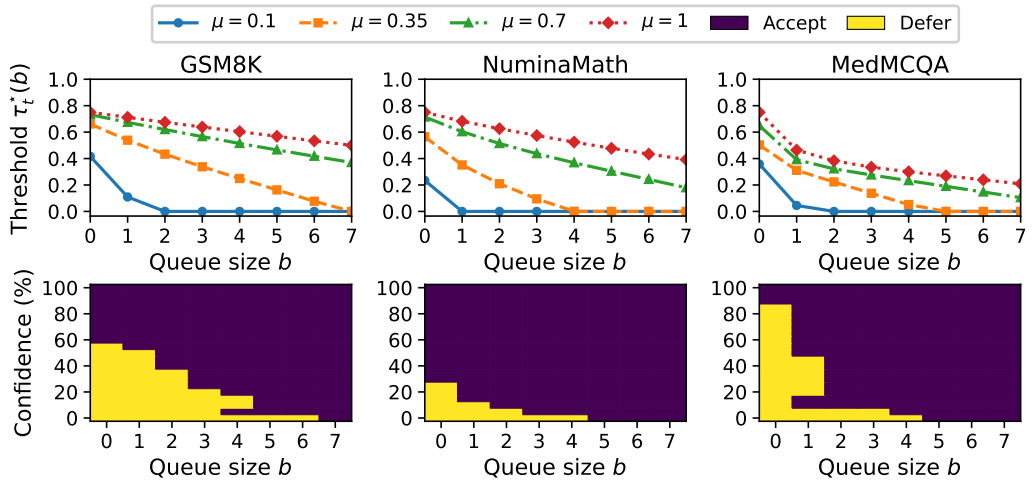


Figure 4: Oracle thresholds for varying service rate μ (top) and oracle decisions for $\mu = 0.35$ (bottom), for Qwen3-0.6B at round $t = T/2$.

cause the conditional accuracy is non-monotone and the threshold falls between the accuracies of two levels whose order is reversed relative to their self-consistency. Similar plots for Qwen3-4B and comparisons with the thresholds and decisions learned by KL−PBD are deferred to Appendix H.4.

7 Conclusion

We studied online deferral to a congested expert, where deferral costs depend on the queue state and expert feedback is both selective and delayed. This creates a feedback loop: learning when to trust the LLM requires expert labels, but acquiring those labels consumes the same limited resource that the learner is trying to manage. KL−PBD addresses this through pessimistic estimates of the unknown conditional accuracies, preventing premature acceptance from cutting off the feedback needed for learning. Its regret guarantees are near-optimal in both the worst-case and gap-dependent regimes.

Several extensions are natural. First, expert effort could depend on the query: low-confidence queries may also be harder to review, making service times confidence-level dependent. Second, our benchmark fixes a discretization of the confidence score; for continuous scores especially, choosing the number of levels would trade approximation error against statistical complexity. Third, the number of LLM generations could itself be treated as a design variable, trading the quality of the confidence signal against inference cost.

AI USE STATEMENT

In this work, we used generative AI tools to assist with: i) writing and editing, and in particular, to improve the presentation of the proofs; ii) coding, to help implement our algorithms and write data-processing and parsing scripts; iii) literature search, to help identify potentially relevant work.

The coding assistance was used strictly to implement our ideas and methods. We conducted our own literature review, and AI was used as a complementary tool for discovering potentially relevant references that we might have missed, not for synthesizing the literature on our behalf. We checked and revised all AI-assisted text and code, and independently verified all AI-suggested references and literature-related claims against the original sources. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide a complete description of our problem setting and assumptions in Section 3, and of our algorithm in Section 4. Full proofs of our results are given in Appendices A–C, with accompanying proof sketches to help the reader navigate them. Our experimental setup is described in Section 6 and Appendices E–H, including the datasets used and how they were parsed, the baselines we compare our algorithm to, and our evaluation protocol. The code used to reproduce our experiments is included in the supplementary material.

REFERENCES

Mastane Achab, Stephan Clémençon, and Aurélien Garivier. Profitable bandits. In Jun Zhu and Ichiro Takeuchi (eds.), *Proceedings of The 10th Asian Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, pp. 694–709. PMLR, 14–16 Nov 2018. URL https://proceedings.mlr.press/v95/achab18a.html.

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761/.

Jean Vieira Alves, Diogo Leitão, Sérgio Jesus, Marco O. P. Sampaio, Javier Liébana, Pedro Saleiro, Mario A. T. Figueiredo, and Pedro Bizarro. Cost-Sensitive Learning to Defer to Multiple Experts with Workload Constraints. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/alves2024tmlr-costsensitive/.

Tavor Z Baharav, Spyros Dragazis, and Aldo Pacchiano. The good, the bad, and the sampled: a no-regret approach to safe online classification. *International Conference on Artificial Intelligence and Statistics*, 2026.

Nina Corvelo Benz, Eleni Straitouri, and Manuel Gomez-Rodriguez. Learning to decide with AI assistance under human-alignment. *arXiv preprint arXiv:2605.12646*, 2026.

Aditya Bhattacharya, Simone Stumpf, Robin De Croon, and Katrien Verbert. Explanatory debiasing: Involving domain experts in the data generation process to mitigate representation bias in AI systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713497. URL https://doi.org/10.1145/3706598.3713497.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2972–3005. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/charusaie22a.html.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R. Featured Certification.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Corinna Cortes, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Optimized deferral for imbalanced settings. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*, 2026. URL https://openreview.net/pdf?id=CfqUMty4Ns.

Nina Corvelo Benz and Manuel Rodriguez. Human-aligned calibration for AI-assisted decision making. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 14609–14636. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2f1d1196426ba84f47d115cac3dcb9d8-Paper-Conference.pdf.

Nina L. Corvelo Benz and Manuel Gomez Rodriguez. Human-alignment influences the utility of AI-assisted decision making. *Scientific Reports*, 15(1):29154, 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-12205-1. URL https://doi.org/10.1038/s41598-025-12205-1.

Sanjoy Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781, April 2011. ISSN 0304-3975. doi: 10.1016/j.tcs.2010.12.054. URL https://doi.org/10.1016/j.tcs.2010.12.054.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. Large language model-based human-agent collaboration for complex task solving. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1336–1357, 2024.

Aditya Gangrade, Anil Kag, Ashok Cutkosky, and Venkatesh Saligrama. Online selective classification with limited feedback. *Advances in Neural Information Processing Systems*, 34:14529–14541, 2021.

Aurélien Garivier. Informational confidence bounds for self-normalized averages and applications. In *2013 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2013. doi: 10.1109/ITW.2013.6691311.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In Sham M. Kakade and Ulrike von Luxburg (eds.), *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 359–376, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL https://proceedings.mlr.press/v19/garivier11a.html.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4147–4180, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/11f5520daf9132775e8604e89f53925a-Paper-Conference.pdf.

Steve Hanneke. Theory of disagreement-based active learning. *Found. Trends Mach. Learn.*, 7(2–3):131–309, June 2014. ISSN 1935-8237. doi: 10.1561/2200000037. URL https://doi.org/10.1561/2200000037.

Shuji Hao, Peiying Hu, Peilin Zhao, Steven C. H. Hoi, and Chunyan Miao. Online active learning with expert advice. *ACM Trans. Knowl. Discov. Data*, 12(5), June 2018. ISSN 1556-4681. doi: 10.1145/3201604. URL https://doi.org/10.1145/3201604.

Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2478–2484. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/344. URL https://doi.org/10.24963/ijcai.2022/344. Main Track.

Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. Learning to defer with limited expert predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6002–6011, 2023.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Xiangji Huang. How to enable effective cooperation between humans and NLP models: A survey of principles, formalizations, and beyond. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 466–488, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.22. URL https://aclanthology.org/2025.acl-long.22/.

Hugging Face. Math-Verify. https://github.com/huggingface/Math-Verify, 2025.

Chunyang Jiang, Yonggang Zhang, Yiyang Cai, Chi-Min Chan, Yulong Liu, Mingming Chen, Wei Xue, and Yike Guo. Semantic voting: A self-evaluation-free approach for efficient LLM self-improvement on unverifiable open-ended tasks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 36536–36556, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3db3e1f192877e47bf48c93cae238e51-Paper-Conference.pdf.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, Aditya Krishna Menon, and Sanjiv Kumar. Universal model routing for efficient LLM inference. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 10169–10218, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/113ae3a9762ca2168f860a8501d6ae25-Paper-Conference.pdf.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 154–165, 2021.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

T. Lattimore and Cs. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 08 2020. URL https://tor-lattimore.com/downloads/book/book.pdf.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath TIR. https://huggingface.co/datasets/AI-MO/NuminaMath-TIR, 2024.

Thodoris Lykouris and Wentao Weng. Learning to defer in congested systems: The AI-human interplay, 2025. URL https://arxiv.org/abs/2402.12237.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in neural information processing systems*, 36:3578–3606, 2023.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable H-consistent and Bayes-consistent loss functions for learning to defer. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 73638–73671. Curran Associates, Inc., 2024a. doi: 10.52202/079017-2342. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/86a11cc08e221cfef301dabc735bbe92-Paper-Conference.pdf.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. *arXiv preprint arXiv:2403.19494*, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable h-consistency and strong guarantees for learning to defer. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43035–43066. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/mao25c.html.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Why ask one when you can ask k? learning-to-defer to the top-k experts. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 66968–67011, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6dc129f0383f9fa270347bb09f0030d4-Paper-Conference.pdf.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 29292–29304. Curran Associates, Inc., 2022. doi: 10.52202/068431-2124. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bc8f76d9caadd48f77025b1c889d2e2d-Paper-Conference.pdf.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8901–8929. Curran Associates, Inc., 2024. doi: 10.52202/079017-0283. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10c456d2160517581a234dfde15a7505-Paper-Conference.pdf.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d693d554e0ede0d75f7d2873b015f228-Paper.pdf.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, Cambridge, MA, 1999.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Guillem Ramírez, Alexandra Birch, and Ivan Titov. Optimising calls to large language models with uncertainty-based two-tier selection. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=T9cOYH0wGF.

Mirabel Reid, Tom Sühr, Claire Vernade, and Samira Samadi. Online decision deferral under budget constraints, 2024. URL https://arxiv.org/abs/2409.20489.

William Réveillard, Vasileios Saketos, Alexandre Proutiere, and Richard Combes. Minimizing human intervention in online classification. *AISTATS 2026*, 2026.

Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester, 1988. ISBN 9780471917878.

Arghyadip Roy, Vivek Borkar, Abhay Karandikar, and Prasanna Chaporkar. Online reinforcement learning of optimal threshold policies for Markov decision processes. *IEEE Transactions on Automatic Control*, 67(7):3722–3729, 2021.

Matteo Russo, Andrea Celli, Riccardo Colini-Baldeschi, Federico Fusco, Daniel Haimovich, Dima Karamshuk, Stefano Leonardi, and Niek Tax. Online learning with sublinear best-action queries. *Advances in Neural Information Processing Systems*, 37:40407–40433, 2024.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3475–3483. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/tailor24a.html.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pp. 22184–22202. PMLR, 2022.

Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11415–11434. PMLR, 2023.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. Self-consistency boosts calibration for math reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6023–6029, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.349. URL https://aclanthology.org/2024.findings-emnlp.349/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ICLR*, 2023.

Lucas Weber, Ana Busic, and Jiamin Zhu. Reinforcement learning and regret bounds for admission control. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52403–52427. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/weber24a.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Xinwei Yang, Zhaofeng Liu, Chen Huang, Jiashuai Zhang, Tong Zhang, Yifan Zhang, and Wenqiang Lei. ELABORATION: A comprehensive benchmark on human-LLM competitive programming. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 59–104, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.4. URL https://aclanthology.org/2025.acl-long.4/.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21691–21728, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5de11e930c1bbfda5d4fc9d2b0924032-Paper-Conference.pdf.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699. ACM, 2002. doi: 10.1145/775047.775151.

Tuo Zhang, Asal Mehradfar, Dimitrios Dimitriadis, and Salman Avestimehr. Leveraging uncertainty estimation for efficient LLM routing. In *ICML 2025 Workshop on Collaborative and Federated Agentic Workflows*, 2025. URL https://openreview.net/forum?id=DJpEIwKJt7.

Appendix Content

A Proofs for Section 3 **17**

A.1 Optimal deferral and regret decomposition 17

A.2 Service-capacity structure 17

B Proofs for Section 4.2 **20**

B.1 Proof sketch 20

B.2 Notation 20

B.3 Worst-case regret of KL-PBD 20

B.4 Gap-dependent regret 23

B.5 Technical lemmas . . . 26
B.6 Extension to unknown arrival and service distributions . . . 33

C Proofs for Section 5 **38**
C.1 Proof sketches . . . 38
C.2 Common ingredients . . . 38
C.3 Instance-dependent regret lower bound . . . 40
C.4 KL−PBD is asymptotically optimal under immediate-service . . . 42
C.5 Worst-case regret lower bound . . . 44
C.6 Technical lemmas . . . 47

D Additional related work **52**
D.1 Alternative self-confidence signals . . . 52
D.2 Probabilistic calibration . . . 53
D.3 Additional learning-to-defer formulations . . . 53
D.4 LLM routing . . . 53
D.5 Structured admission control . . . 54
D.6 Active learning and oracle queries . . . 54
D.7 Human–AI decision support . . . 54

E Datasets **55**
E.1 Synthetic datasets . . . 55
E.2 Real-world datasets . . . 55

F Baselines and implementation details **57**

G Synthetic experiments **59**
G.1 Effect of K and of the oracle gap on KL−PBD regret . . . 59
G.2 Effect of service rate and cost parameters . . . 60

H Additional real-world experiments and details **64**
H.1 Regret of all algorithms . . . 64
H.2 Regret as a function of the horizon . . . 64
H.3 Service rate and queue-unaware algorithms . . . 64
H.4 Learned threshold and decisions . . . 67
H.5 The importance of pessimism . . . 68

A PROOFS FOR SECTION 3

Throughout the appendices, write $\mathcal{B} := \{0, \dots, \bar{B}\}$ and $S_t := (B_t, J_t)$. Let $\mathcal{A}(b, j) := \{0, 1\}$ when $j \in [K]$ and $b < \bar{B}$, and $\mathcal{A}(b, j) := \{0\}$ otherwise. We recall that action 1 corresponds to deferral, and action 0 to acceptance. Let $\mathcal{F}_t$ be the sigma-field generated by the self-confidence levels, actions, service capacities, completed-query identities and their revealed ground-truth labels through the end of round t, together with past internal randomization. The algorithm chooses A_t using only $\mathcal{F}_{t-1}$, J_t, and independent randomization. The queries completed in each round are selected using only the observed history and service capacity, before any of their ground-truth labels are revealed. Thus B_t is $\mathcal{F}_{t-1}$-measurable, whereas J_t is independent of $\mathcal{F}_{t-1}$ with distribution ω. For a real-valued function f on a finite domain, we write $\mathrm{sp}(f) := \max_x f(x) - \min_x f(x)$.

A.1 OPTIMAL DEFERRAL AND REGRET DECOMPOSITION

Proof of Lemma 1. For $b < \bar{B}$, $j \in [K]$, and $s = (b, j)$,

$$Q_t^h(s, 1) - Q_t^h(s, 0) = M_t^h(b) - L^h(j). \tag{4}$$

Hence deferral is optimal when $L^h(j) > M_t^h(b)$, acceptance is optimal when $L^h(j) < M_t^h(b)$, and both actions are optimal at equality. On no-query and full-queue states, acceptance is the only feasible action. Resolving ties in favor of acceptance gives the stated action rule and its equivalent threshold formulation. □

Proof of Lemma 2. All expectations are under the true conditional accuracy vector $h^\star$. By the Bellman definition and the observation model,

$$Q_t^\star(S_t, A_t) = \mathbb{E}^\pi\left[c_t + V_{t+1}^\star(S_{t+1}) \mid \mathcal{F}_{t-1}, J_t, A_t\right].$$

Taking expectations and summing over t, the value terms telescope:

$$\mathbb{E}^\pi\left[\sum_{t=1}^{T}\big(Q_t^\star(S_t, A_t) - V_t^\star(S_t)\big)\right] = \mathbb{E}^\pi\left[\sum_{t=1}^{T} c_t + D(B_{T+1})\right] - \mathbb{E}^\pi[V_1^\star(S_1)].$$

Since $B_1 = 0$ and $J_1 \sim \omega$ under both the algorithm and the oracle policy,

$$\mathbb{E}^\pi[V_1^\star(S_1)] = \sum_{j=0}^{K} \omega_j V_1^\star((0, j)) = \mathbb{E}^{\pi^\star}\left[\sum_{t=1}^{T} c_t + D(B_{T+1})\right].$$

Consequently,

$$R_T(\pi) = \mathbb{E}^\pi\left[\sum_{t=1}^{T}\big(Q_t^\star(S_t, A_t) - V_t^\star(S_t)\big)\right]. \tag{5}$$

When $J_t \in [K]$ and $B_t < \bar{B}$, equation 4 gives

$$Q_t^\star(S_t, A_t) - V_t^\star(S_t) = \big|L^\star(J_t) - M_t^\star(B_t)\big|\mathbf{1}\{A_t \neq A_t^\star\} = \Delta_t$$

since the action space is binary. Otherwise, acceptance is the only feasible action, and both sides are zero. Substitution into equation 5 proves the claim. □

A.2 SERVICE-CAPACITY STRUCTURE

In this section, we prove several structural properties induced by our service-capacity model. The proposition below states that the marginal cost of deferral increases with the expert queue state.

Proposition 1 (Monotone marginal cost)**.** *For every conditional accuracy vector h and every $t \le T$, the Bellman marginal cost is nondecreasing in the queue state:*

$$c_{\mathrm{exp}} \le M_t^h(0) \le M_t^h(1) \le \cdots \le M_t^h(\bar{B} - 1).$$

Consequently, for every round t and self-confidence level $j \in [K]$, there exists a threshold $\tau_{t,j}^{\mathrm{queue}}(h) \in \{0, \dots, \bar{B}\}$ such that, resolving ties in favor of acceptance, the optimal policy defers a query with self-confidence level j if and only if

$$B_t < \tau_{t,j}^{\mathrm{queue}}(h).$$

For a conditional accuracy vector h, consider the queue value

$$U_t^h(b) := \sum_{j=0}^{K} \omega_j V_t^h((b,j)),$$

i.e., the value at queue state b before the level of the incoming query is revealed. For $t \le T$ and $b \in \mathcal{B}$, let $F_t^h(b) := \mathbb{E}[U_{t+1}^h((b - C_t)^+)]$, and, for $b < \bar{B}$, define $d_t^h(b) := F_t^h(b+1) - F_t^h(b)$. Then the Bellman marginal cost can be written as

$$M_t^h(b) = c_{\exp} + d_t^h(b), \qquad b < \bar{B}. \tag{6}$$

We first establish the structural properties of the queue value function that will also be useful later. Define its adjacent increments by

$$x_t^h(q) := U_t^h(q) - U_t^h(q-1), \qquad q = 1, \dots, \bar{B}.$$

Lemma 3 (Queue-value increments under service capacity)**.** *For every conditional accuracy vector h and every $t \le T+1$,*

$$0 \le x_t^h(1) \le x_t^h(2) \le \cdots \le x_t^h(\bar{B}).$$

Consequently, U_t^h is nondecreasing and discrete convex in the queue state.

Moreover, for every $t \le T$ and $b < \bar{B}$,

$$d_t^h(b) = \sum_{c=0}^{b} \rho_C(c)\, x_{t+1}^h(b+1-c). \tag{7}$$

Proof. Equation 7 follows directly from the service-capacity dynamics. Indeed,

$$d_t^h(b) = \mathbb{E}\big[U_{t+1}^h((b+1-C_t)^+) - U_{t+1}^h((b-C_t)^+)\big]\,.$$

If $C_t = c \le b$, the two queue states differ by one and the difference equals $x_{t+1}^h(b+1-c)$. If $c > b$, both queue states are zero. Hence equation 7 follows.

We prove the monotonicity of the increments by backward induction. At $t = T+1$, we have $U_{T+1}^h = D$, independently of h. Let $S_0 := 0$ and $S_k := C_1 + \cdots + C_k$ for $k \ge 1$, where $C_1, C_2, \dots$ are i.i.d. with law ρ_C. Iterating the definition of D gives

$$D(b) = c_{\text{hold}} \sum_{k \ge 0} \mathbb{E}[(b - S_k)^+].$$

The series is finite since $\mu > 0$. Let $g(q) := D(q) - D(q-1)$. Since $(q - S_k)^+ - (q - 1 - S_k)^+ = \mathbf{1}\{S_k < q\}$, we have

$$g(q) = c_{\text{hold}} \sum_{k \ge 0} \mathbb{P}(S_k < q).$$

Thus $g(q) \ge 0$ and is nondecreasing in q, so

$$0 \le g(1) \le g(2) \le \cdots \le g(\bar{B}).$$

Therefore $U_{T+1}^h = D$ is nondecreasing and discrete convex, establishing the base case.

Suppose now that

$$0 \le x_{t+1}^h(1) \le \cdots \le x_{t+1}^h(\bar{B}).$$

Equation 7 immediately gives $d_t^h(b) \ge 0$. Moreover, for $b < \bar{B} - 1$,

$$\begin{aligned} d_t^h(b+1) - d_t^h(b) = &\sum_{c=0}^{b} \rho_C(c)\, \big[x_{t+1}^h(b+2-c) - x_{t+1}^h(b+1-c)\big] \\ &+ \rho_C(b+1) x_{t+1}^h(1) \ge 0. \end{aligned}$$

Thus

$$0 \le d_t^h(0) \le d_t^h(1) \le \cdots \le d_t^h(\bar{B}-1). \tag{8}$$

For $j = 0, \dots, K$, write $r_j^h := L^h(j) - c_{\text{exp}}$. In particular, $r_0^h = -c_{\text{exp}}$ because $L^h(0) = 0$.

For $b < \bar{B}$, the Bellman equation for the queue value function gives

$$U_t^h(b) = c_{\text{hold}} b + F_t^h(b) + c_{\text{exp}} + \sum_{j=0}^{K} \omega_j \min\{r_j^h, d_t^h(b)\}.$$

Indeed, for $j \in [K]$ this represents the minimum between acceptance and deferral costs. For $j = 0$, action 0 is forced; since $r_0^h = -c_{\text{exp}}$ and $d_t^h(b) \geq 0$, the corresponding term satisfies $c_{\text{exp}} + \min\{r_0^h, d_t^h(b)\} = 0$, and therefore represents the no-query state correctly.

Consequently, for $1 \leq q < \bar{B}$, subtracting the Bellman equations for the queue value function at q and $q-1$ gives

$$\begin{aligned} x_t^h(q) &= c_{\text{hold}} + d_t^h(q-1) + \sum_{j=0}^{K} \omega_j \left[\min\{r_j^h, d_t^h(q)\} - \min\{r_j^h, d_t^h(q-1)\}\right] \\ &= c_{\text{hold}} + \sum_{j=0}^{K} \omega_j \operatorname{med}\{d_t^h(q-1), d_t^h(q), r_j^h\}, \end{aligned} \tag{9}$$

where the second equality uses $\sum_{j=0}^{K} \omega_j = 1$ and $d_t^h(q-1) \leq d_t^h(q)$, and $\operatorname{med}\{x, y, z\}$ denotes the median of its three arguments.

At $q = \bar{B}$, action 1 is infeasible. Hence

$$U_t^h(\bar{B}) = c_{\text{hold}} \bar{B} + F_t^h(\bar{B}) + \sum_{j=0}^{K} \omega_j L^h(j),$$

which together with the Bellman equation at $\bar{B} - 1$ yields

$$x_t^h(\bar{B}) = c_{\text{hold}} + \sum_{j=0}^{K} \omega_j \max\left\{d_t^h(\bar{B}-1), r_j^h\right\}. \tag{10}$$

By equation 8, for every fixed r, $\operatorname{med}\{d_t^h(q-1), d_t^h(q), r\}$ is nonnegative and nondecreasing in q. Therefore

$$0 \leq x_t^h(1) \leq \cdots \leq x_t^h(\bar{B}-1).$$

Finally,

$$\max\left\{d_t^h(\bar{B}-1), r\right\} \geq \operatorname{med}\left\{d_t^h(\bar{B}-2), d_t^h(\bar{B}-1), r\right\},$$

whenever $\bar{B} \geq 2$. Combining equation 9 and equation 10 therefore gives $x_t^h(\bar{B}) \geq x_t^h(\bar{B}-1)$. For $\bar{B} = 1$, the conclusion follows directly from equation 10. This completes the backward induction. □

Proof of Proposition 1. By Lemma 3, $x_{t+1}^h(q)$ is nonnegative and nondecreasing in q. Hence equation 7 and equation 6 give

$$\begin{aligned} 0 \leq d_t^h(0) \leq d_t^h(1) \leq \cdots \leq d_t^h(\bar{B}-1), \\ c_{\text{exp}} \leq M_t^h(0) \leq M_t^h(1) \leq \cdots \leq M_t^h(\bar{B}-1). \end{aligned}$$

Fix t and self-confidence level j, and define

$$\tau_{t,j}^{\text{queue}}(h) := \min\left(\left\{b \in \{0, \dots, \bar{B}-1\} : L^h(j) \leq M_t^h(b)\right\} \cup \{\bar{B}\}\right).$$

Since $M_t^h(b)$ is nondecreasing in b, $L^h(j) > M_t^h(b)$ if and only if $b < \tau_{t,j}^{\text{queue}}(h)$. By equation 4, the optimal policy with ties resolved in favor of acceptance therefore defers if and only if $B_t < \tau_{t,j}^{\text{queue}}(h)$. □

B Proofs for Section 4.2

B.1 Proof sketch

To ease navigation within this appendix, we first provide a high-level proof sketch of our gap-dependent and worst-case upper bounds. They are proven in Appendices B.4 and B.3 respectively.

Proof sketch. The analysis exploits that pessimism makes KL−PBD almost one-sided. On the high-probability event $\underline{h}_t \le h^\star$, if the marginal cost of deferral were computed under the true model $h^\star$, the learner would accept only if the oracle does, and regret would come from deferrals only. This ordering is not exact, because $\underline{h}_t$ is also plugged into the marginal cost of deferral. To account for these additional errors, we track the difference between the value functions of the pessimistic and true models: Proposition 2 relates this misspecification to the estimation error. Updating $\underline{h}_t$ leaves an additional term controlled by its total variation, which is logarithmic. The remaining estimation error is charged only on deferred rounds, which are exactly the rounds that produce labels, so the confidence widths shrink as they are charged. Summing these widths incurs only an additive $\bar{B}$ per level from delayed feedback, since at most $\bar{B}$ previously deferred queries can still be unlabeled, and yields Theorem 2. For Theorem 1, we exploit that, by Proposition 2(ii), estimation errors smaller than $\gamma_T/2$ cannot change strict oracle decisions. We therefore compare $\underline{h}_t$ with the intermediate model obtained by replacing $\underline{h}_{j,t}$ with $\max\{\underline{h}_{j,t}, h^\star_j - \gamma_T/2\}$, which agrees with the oracle on every strict comparison. Only the confidence width beyond $\gamma_T/2$ needs to be charged, yielding the $1/\gamma_T$ dependence. □

B.2 Notation

All expectations in this appendix are under $h^\star$ and $\pi^{\text{KL−PBD}}$. For $j \in [K]$ and $t \ge 0$, define $D_{j,t} := \sum_{s=1}^{t} \mathbf{1}\{J_s = j, A_s = 1\}$. Recall that $N_{j,t}$ counts completed self-confidence level j labels available before round t. In this appendix, we prove more general regret upper bounds from which the bounds of Theorems 2 and 1 directly follow. The technical lemmas used in the regret proofs are stated and proved in Section B.5.

We first establish bounds for a general confidence budget $\phi(T) > 1$, then specialize to $\phi(T) = \log T + \log\log T$ to obtain the main-text bounds. For $T \ge 2$, define

$$\eta_T := T \min\left\{1,\ 2KTe^{-\phi(T)},\ 2eK\lceil \phi(T)\log T\rceil e^{-\phi(T)}\right\}, \tag{11}$$

set $\eta_1 := \min\{1, 2Ke^{-\phi(1)}\}$, and write

$$\Lambda_T := \sum_{j=1}^{K} \log\left(1 + \frac{\omega_j T}{\phi(T)}\right).$$

These two quantities are used only to state the appendix bounds.

For a fixed budget $\phi(T) > 1$, let $\widehat{h}_{j,n}$ denote the empirical mean of the first n ground-truth labels of self-confidence level j, extended to an infinite i.i.d. $\mathrm{Bernoulli}(h^\star_j)$ sequence as in Lemma 8, and define the concentration event

$$\mathcal{E} := \bigcap_{j=1}^{K}\bigcap_{n=1}^{T}\left\{n\,\mathrm{kl}(\widehat{h}_{j,n}, h^\star_j) \le \phi(T)\right\}. \tag{12}$$

Let also

$$g_{j,0} := 0, \qquad g_{j,n} := \inf\left\{q \in [0, \widehat{h}_{j,n}] : n\,\mathrm{kl}(\widehat{h}_{j,n}, q) \le \phi(T)\right\}, \quad n \ge 1. \tag{13}$$

Thus $\underline{h}_{j,t} = g_{j,N_{j,t}}$, where $N_{j,t}$ counts completed labels available before round t.

B.3 Worst-case regret of KL−PBD

We first prove a refined bound that retains the Bernoulli KL geometry. The main-text theorem follows by bounding the KL confidence width using Pinsker's inequality.

For a Bernoulli parameter $p \in [0,1]$ and $x \geq 0$, define

$$w_p(x) := \sup\{p - q : 0 \leq q \leq p,\ \exists z \in [0,1],\ \mathrm{kl}(z,p) \leq x,\ \mathrm{kl}(z,q) \leq x\}. \tag{14}$$

This function is nondecreasing in x and takes values in $[0,p]$. For $\Gamma > 0$ and $d \geq 0$, let

$$F_{p,\Gamma}(d) := \int_0^d w_p(\Gamma/u)\, du, \tag{15}$$

with the integrand at zero defined by its right limit.

Theorem 5 (Refined regret bound)**.** *For any fixed budget $\phi(T) > 1$,*

$$\begin{aligned} R_T(\pi^{\textit{KL-PBD}}) \leq \eta_T \Delta + c_e \sum_{j=1}^{K} \min\left\{h_j^\star \omega_j T,\ \bar{B} h_j^\star + F_{h_j^\star, \phi(T)}(\omega_j T)\right\} \\ + \bar{B} c_e (1 + 2\Lambda_T). \end{aligned} \tag{16}$$

where $\Delta := \max\{c_e, c_{\exp}\} + \frac{c_{\mathrm{hold}} \bar{B}}{\mu}$.

Recall the history $\mathcal{F}_t$ from Appendix A. Immediately before J_t is observed, B_t, the completed-label counts, and $\underline{h}_t$ are $\mathcal{F}_{t-1}$-measurable, while J_t is independent of $\mathcal{F}_{t-1}$ with distribution ω. Define

$$\mathcal{E}_t := \left\{N_{j,t}\, \mathrm{kl}(\widehat{h}_{j,N_{j,t}}, h_j^\star) \leq \phi(T) \text{ for every } j \text{ with } N_{j,t} \geq 1\right\}, \quad \sigma := \inf\{t \in [T] : \mathcal{E}_t^c\} \wedge (T+1).$$

Then $\{t < \sigma\} \in \mathcal{F}_{t-1}$. Since $N_{j,t} \leq T$, we have $\mathcal{E}_t^c \subseteq \mathcal{E}^c$ for every t, so Lemma 8 gives

$$\mathbb{P}(\sigma \leq T) \leq \mathbb{P}(\mathcal{E}^c) \leq \frac{\eta_T}{T}. \tag{17}$$

For $t < \sigma$, let $b_{0,t} := 0$ and for $j \in [K]$, define

$$b_{j,t} := c_e(h_j^\star - \underline{h}_{j,t}). \tag{18}$$

For $j \in [K]$, Lemma 9 gives

$$0 \leq b_{j,t} \leq \begin{cases} c_e h_j^\star, & N_{j,t} = 0, \\ c_e w_{h_j^\star}(\phi(T)/N_{j,t}), & N_{j,t} \geq 1. \end{cases} \tag{19}$$

Regret decomposition. Define

$$\begin{aligned} \mathcal{R}_{\mathrm{fail}} &:= \mathbb{E}\left[\sum_{t=\sigma}^{T} \Delta_t \mathbf{1}\{\sigma \leq T\}\right], \\ \mathcal{R}_{\mathrm{defer}} &:= \mathbb{E}\left[\sum_{t<\sigma} b_{J_t,t} \mathbf{1}\{A_t = 1\}\right], \\ \mathcal{R}_{\mathrm{update}} &:= \bar{B} c_e \left[1 + \mathbb{E}\left[\sum_{t=2}^{T} \mathbf{1}\{t < \sigma\} \left\|\underline{h}_t - \underline{h}_{t-1}\right\|_\infty\right]\right]. \end{aligned}$$

Lemma 4 (Regret decomposition)**.** *The regret satisfies*

$$R_T(\pi^{\textit{KL-PBD}}) \leq \underbrace{\mathcal{R}_{\mathrm{fail}}}_{\text{(I)}} + \underbrace{\mathcal{R}_{\mathrm{defer}}}_{\text{(II)}} + \underbrace{\mathcal{R}_{\mathrm{update}}}_{\text{(III)}}. \tag{20}$$

Proof of Lemma 4. By Lemma 2,

$$R_T(\pi^{\mathrm{KL-PBD}}) = \mathbb{E}\left[\sum_{t<\sigma} \Delta_t\right] + \mathcal{R}_{\mathrm{fail}}.$$

Thus it remains to control the regret accumulated before σ. On these rounds, $\underline{h}_t \le h^\star$ coordinate-wise. Apply Lemma 13 with $\widetilde{h}_t = h^\star$. Then $\widetilde{\Delta}_t = \Delta_t$, $e_{j,t} = b_{j,t}$, and the reference vector does not change with t. Consequently,

$$\mathbb{E}\left[\sum_{t<\sigma} \Delta_t\right] \le \mathbb{E}\left[\sum_{t<\sigma} b_{J_t,t}\mathbf{1}\{A_t = 1\}\right] + \bar{B}c_e\left(1 + \mathbb{E}\left[\sum_{t=2}^{\sigma-1} \|\underline{h}_t - \underline{h}_{t-1}\|_\infty\right]\right)$$
$$= \mathcal{R}_{\mathrm{defer}} + \mathcal{R}_{\mathrm{update}}.$$

Adding $\mathcal{R}_{\mathrm{fail}}$ proves equation 20. □

Lemma 5 (Term (I): concentration failure)**.** *We have* $\mathcal{R}_{\mathrm{fail}} \le \eta_T \Delta$.

Proof. Lemma 12 and equation 17 give

$$\mathcal{R}_{\mathrm{fail}} \le T\Delta\mathbb{P}(\sigma \le T) \le \eta_T\Delta.$$

□

Lemma 6 (Term (II): deferred-round contribution)**.** *We have*

$$\mathcal{R}_{\mathrm{defer}} \le c_e \sum_{j=1}^{K} \min\left\{h_j^\star \omega_j T,\ \bar{B}h_j^\star + F_{h_j^\star,\phi(T)}(\omega_j T)\right\}. \tag{21}$$

Proof. Fix a self-confidence level j. Define $f_j(0) := h_j^\star$ and $f_j(u) := w_{h_j^\star}(\phi(T)/u)$ for $u > 0$. Since $w_{h_j^\star}$ is nondecreasing and takes values in $[0, h_j^\star]$, f_j is nonnegative and nonincreasing. By equation 19,

$$b_{j,t} \le c_e f_j(N_{j,t}), \qquad t < \sigma.$$

Moreover, the definition of $F_{p,\Gamma}$ gives $\int_0^{\omega_j T} f_j(u)\,du = F_{h_j^\star,\phi(T)}(\omega_j T)$. Applying Lemma 14, we obtain

$$\mathbb{E}\left[\sum_{t<\sigma} b_{j,t}\mathbf{1}\{J_t = j, A_t = 1\}\right] \le c_e\mathbb{E}\left[\sum_{t<\sigma} \mathbf{1}\{J_t = j, A_t = 1\} f_j(N_{j,t})\right]$$
$$\le c_e \min\left\{h_j^\star\omega_j T,\ \bar{B}h_j^\star + F_{h_j^\star,\phi(T)}(\omega_j T)\right\}.$$

Summing over $j \in [K]$ proves equation 21. □

Lemma 7 (Term (III): Bellman-model updates)**.** *We have*

$$\mathcal{R}_{\mathrm{update}} \le \bar{B}c_e(1 + 2\Lambda_T).$$

Proof. Since $\mathbf{1}\{t < \sigma\} \le 1$, Lemma 10 gives

$$\mathcal{R}_{\mathrm{update}} \le \bar{B}c_e\left[1 + \mathbb{E}\left[\sum_{t=2}^{T} \left\|\underline{h}_t - \underline{h}_{t-1}\right\|_\infty\right]\right]$$
$$\le \bar{B}c_e(1 + 2\Lambda_T).$$

□

Proof of Theorem 5. Combine Lemmas 4, 5, 6, and 7. □

Proof of Theorem 2. For the main-text budget $\phi(T) = \log T + \log\log T$, Lemma 8 gives $\eta_T \le 4eK\phi(T)$. Using $h_j^\star \le 1$ and $F_{h_j^\star,\phi(T)}(\omega_j T) \le 2\sqrt{2\omega_j T\phi(T)}$ in Theorem 5 gives

$$R_T(\pi^{\texttt{KL-PBD}}) \le 4eK\phi(T)\Delta + c_e\sum_{j=1}^{K} \min\left\{\omega_j T,\ \bar{B} + 2\sqrt{2\omega_j T\phi(T)}\right\}$$
$$+ \bar{B}c_e\left[1 + 2\sum_{j=1}^{K} \log\left(1 + \frac{\omega_j T}{\phi(T)}\right)\right].$$

Since $\kappa = 4e\Delta$, this is the main-text bound. □

Remark 1 (Worst case over ω). Since $\sum_{j=1}^{K} \omega_j \le 1$, Jensen's inequality gives

$$\begin{aligned}\Lambda_T &= \sum_{j=1}^{K} \log\left(1 + \frac{\omega_j T}{\phi(T)}\right) \\ &\le K \log\left(1 + \frac{T}{K\phi(T)} \sum_{j=1}^{K} \omega_j\right) \\ &\le K \log\left(1 + \frac{T}{K\phi(T)}\right).\end{aligned}$$

Moreover,

$$\sum_{j=1}^{K} \sqrt{\omega_j T} \le \sqrt{KT \sum_{j=1}^{K} \omega_j} \le \sqrt{KT}, \qquad \sum_{j=1}^{K} \omega_j T \le T.$$

Hence

$$\sum_{j=1}^{K} \min\left\{\omega_j T,\ \bar{B} + 2\sqrt{2\omega_j T \phi(T)}\right\} \le \min\left\{T,\ K\bar{B} + 2\sqrt{2KT\phi(T)}\right\},$$

and therefore

$$\begin{aligned}R_T(\pi^{\texttt{KL-PBD}}) \le\ & 4eK\phi(T)\Delta + c_e \min\{T,\ K\bar{B} + 2\sqrt{2KT\phi(T)}\} \\ & + \bar{B}c_e\left[1 + 2K \log\left(1 + \frac{T}{K\phi(T)}\right)\right].\end{aligned}$$

These inequalities hold with equality when $\omega_0 = 0$ and $\omega_j = 1/K$ for every $j \in [K]$, so this bound is the largest value of the bound of Theorem 2 over ω.

Alternatively, retaining the variance-sensitive term in equation 30 gives, for any fixed budget $\phi(T) > 1$,

$$\begin{aligned}R_T(\pi^{\texttt{KL-PBD}}) \le\ & \eta_T \Delta + \bar{B}c_e(K+1) \\ & + 4c_e\sqrt{2T\phi(T)} \sum_{j=1}^{K} \sqrt{h_j^\star(1-h_j^\star)\omega_j} + c_e(7\phi(T) + 2\bar{B})\Lambda_T.\end{aligned}$$

The term proportional to $\phi(T)\Lambda_T$ is a statistical remainder; it is separate from the Bellman-model update contribution.

B.4 Gap-dependent regret

We use the gap definitions from the main text. For $p \in [0,1]$ and $a > 0$, define

$$I_p(a) := \int_0^\infty \big(w_p(1/u) - a\big)_+ \, du, \tag{22}$$

with the integrand at zero defined by its right limit.

Theorem 6 (Refined gap-dependent regret bound)**.** *If $\mathcal{G}_T = \emptyset$, the regret is zero. Otherwise, let $a_T = \gamma_T/2$. For any budget $\phi(T) > 1$ held fixed throughout the run,*

$$\begin{aligned}R_T(\pi^{\textit{KL-PBD}}) \le\ & \eta_T \Delta + 2c_e \sum_{j=1}^{K} \min\left\{(h_j^\star - a_T)_+ \omega_j T,\ \bar{B}(h_j^\star - a_T)_+ + \phi(T) I_{h_j^\star}(a_T)\right\} \\ & + 2\bar{B}c_e(1 + 4\Lambda_T).\end{aligned} \tag{23}$$

Proof of Theorem 6. If $\mathcal{G}_T = \emptyset$, then $h_j^\star = \tau_t^\star(b)$ for every $j \in [K]$, $t \le T$, and $b < \bar{B}$. Hence every feasible action contributes 0 to regret and $R_T(\pi^{\texttt{KL-PBD}}) = 0$. We therefore assume below that $\mathcal{G}_T \neq \emptyset$.

Set $a := a_T = \frac{\gamma_T}{2}$, $\theta := c_e a$. Recall from equation 18 that, for $t < \sigma$, $b_{j,t} = c_e(h_j^\star - \underline{h}_{j,t}) \ge 0$.

Clipped intermediate model. For $t < \sigma$ and $j \in [K]$, define

$$b^\circ_{j,t} := \min\{b_{j,t}, \theta\}, \qquad h^\circ_{j,t} := h^\star_j - \frac{b^\circ_{j,t}}{c_e},$$

and let

$$x_{j,t} := b_{j,t} - b^\circ_{j,t} = (b_{j,t} - \theta)_+.$$

For notational convenience, set $x_{0,t} := 0$. Thus

$$\underline{h}_t \le h^\circ_t \le h^\star, \qquad \|h^\circ_t - h^\star\|_\infty \le a.$$

For $j \in [K]$, $t \le T$, and $b < \bar{B}$, define the action margin

$$D^h_t(j,b) := M^h_t(b) - c_e(1 - h_j).$$

Under $h^\star$, the margin satisfies $D^\star_t(j,b) = c_e\big(h^\star_j - \tau^\star_t(b)\big)$. Since $h^\circ_t \le h^\star$, Proposition 2(ii) gives, for every $b < \bar{B}$,

$$0 \le M^{h^\circ_t}_t(b) - M^\star_t(b) \le c_e\|h^\circ_t - h^\star\|_\infty \le \theta.$$

Moreover, for every $j \in [K]$,

$$c_e(1 - h^\circ_{j,t}) - c_e(1 - h^\star_j) = b^\circ_{j,t} \in [0, \theta].$$

Consequently, for every $j \in [K]$ and $b < \bar{B}$,

$$\left|D^{h^\circ_t}_t(j,b) - D^\star_t(j,b)\right| \le \theta.$$

Whenever $D^\star_t(j,b) \ne 0$, the definition of γ_T gives $|D^\star_t(j,b)| = c_e|h^\star_j - \tau^\star_t(b)| \ge c_e\gamma_T = 2\theta$.

Hence $D^{h^\circ_t}_t(j,b)$ has the same sign as $D^\star_t(j,b)$ and

$$\left|D^{h^\circ_t}_t(j,b)\right| \ge \frac{1}{2}\left|D^\star_t(j,b)\right|.$$

For $t < \sigma$, define

$$\Delta^\circ_t := Q^{h^\circ_t}_t(S_t, A_t) - V^{h^\circ_t}_t(S_t).$$

If A_t is optimal under $h^\star$, then $\Delta_t = 0 \le 2\Delta^\circ_t$. Otherwise, necessarily $J_t \in [K]$ and $B_t < \bar{B}$, because action 0 is forced on no-query and full-queue states. The sign comparison above then shows that A_t is also suboptimal under h°_t, and

$$\Delta^\circ_t = \left|D^{h^\circ_t}_t(J_t, B_t)\right| \ge \frac{1}{2}\left|D^\star_t(J_t, B_t)\right| = \frac{1}{2}\Delta_t.$$

Therefore

$$\Delta_t \le 2\Delta^\circ_t, \qquad t < \sigma. \tag{24}$$

Regret relative to the clipped model. For every $j \in [K]$ and $t < \sigma$, the conditional accuracy vectors $\underline{h}_t$ and h°_t satisfy

$$c_e(h^\circ_{j,t} - \underline{h}_{j,t}) = c_e(h^\star_j - \underline{h}_{j,t}) - b^\circ_{j,t} = b_{j,t} - b^\circ_{j,t} = x_{j,t}.$$

Apply Lemma 13 with $\widetilde{h}_{j,t} := \max\{\underline{h}_{j,t}, h^\star_j - a\}$. This vector is $\mathcal{F}_{t-1}$-measurable, satisfies $\underline{h}_t \le \widetilde{h}_t$, and equals h°_t for $t < \sigma$. Thus $e_{j,t} = x_{j,t}$ and $\widetilde{\Delta}_t = \Delta^\circ_t$ on these rounds, giving

$$\begin{aligned}\mathbb{E}\left[\sum_{t<\sigma}\Delta^\circ_t\right] &\le \mathbb{E}\left[\sum_{t<\sigma} x_{J_t,t}\mathbf{1}\{A_t = 1\}\right] \\ &\quad + \bar{B}c_e\left(1 + \mathbb{E}\left[\sum_{t=2}^{\sigma-1}\left(\|\underline{h}_t - \underline{h}_{t-1}\|_\infty + \|h^\circ_t - h^\circ_{t-1}\|_\infty\right)\right]\right).\end{aligned}$$

It remains to control the variation of the clipped model. The clipping map is 1-Lipschitz: for every $j \in [K]$ and $2 \le t < \sigma$,

$$\begin{aligned}|h^\circ_{j,t} - h^\circ_{j,t-1}| &= \left|\max\{\underline{h}_{j,t}, h^\star_j - a\} - \max\{\underline{h}_{j,t-1}, h^\star_j - a\}\right| \\ &\le |\underline{h}_{j,t} - \underline{h}_{j,t-1}|.\end{aligned}$$

Hence $\|h^\circ_t - h^\circ_{t-1}\|_\infty \le \|\underline{h}_t - \underline{h}_{t-1}\|_\infty$. By Lemma 10,

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=2}^{\sigma-1} \left(\|\underline{h}_t - \underline{h}_{t-1}\|_\infty + \|h^\circ_t - h^\circ_{t-1}\|_\infty\right)\right] \\ &\qquad \le 2\mathbb{E}\left[\sum_{t=2}^{T} \|\underline{h}_t - \underline{h}_{t-1}\|_\infty\right] \le 4\Lambda_T.\end{aligned}$$

It follows that

$$\mathbb{E}\left[\sum_{t<\sigma} \Delta^\circ_t\right] \le \mathbb{E}\left[\sum_{t<\sigma} x_{J_t,t}\mathbf{1}\{A_t = 1\}\right] + \bar{B}c_e(1 + 4\Lambda_T).$$

Combining this with equation 24,

$$\mathbb{E}\left[\sum_{t<\sigma} \Delta_t\right] \le 2\mathbb{E}\left[\sum_{t<\sigma} x_{J_t,t}\mathbf{1}\{A_t = 1\}\right] + 2\bar{B}c_e(1 + 4\Lambda_T). \tag{25}$$

Residual error on deferred rounds. Fix a self-confidence level j. Define $f_j(0) := (h^\star_j - a)_+$ and $f_j(u) := (w_{h^\star_j}(\phi(T)/u) - a)_+$ for $u > 0$. This function is nonnegative and nonincreasing. Since $\theta = c_e a$, equation 19 gives

$$x_{j,t} = (b_{j,t} - \theta)_+ \le \begin{cases} c_e(h^\star_j - a)_+, & N_{j,t} = 0, \\ c_e\big(w_{h^\star_j}(\phi(T)/N_{j,t}) - a\big)_+, & N_{j,t} \ge 1, \end{cases} \qquad t < \sigma.$$

Thus $x_{j,t} \le c_e f_j(N_{j,t})$. Applying Lemma 14,

$$\begin{aligned}\mathbb{E}\left[\sum_{t<\sigma} x_{j,t}\mathbf{1}\{J_t = j, A_t = 1\}\right] &\le c_e\mathbb{E}\left[\sum_{t<\sigma} \mathbf{1}\{J_t = j, A_t = 1\} f_j(N_{j,t})\right] \\ &\le c_e \min\left\{(h^\star_j - a)_+\omega_j T,\ \bar{B}(h^\star_j - a)_+ + \int_0^{\omega_j T} f_j(u)\,du\right\}.\end{aligned}$$

The integral is bounded using nonnegativity and the change of variables $u = \phi(T)v$:

$$\begin{aligned}\int_0^{\omega_j T} f_j(u)\,du &\le \int_0^\infty \big(w_{h^\star_j}(\phi(T)/u) - a\big)_+\,du \\ &= \phi(T)\int_0^\infty \big(w_{h^\star_j}(1/v) - a\big)_+\,dv \\ &= \phi(T) I_{h^\star_j}(a).\end{aligned}$$

Substituting this bound and summing over $j \in [K]$ gives

$$\mathbb{E}\left[\sum_{t<\sigma} x_{J_t,t}\mathbf{1}\{A_t = 1\}\right] \le c_e \sum_{j=1}^K \min\left\{(h^\star_j - a)_+\omega_j T,\ \bar{B}(h^\star_j - a)_+ + \phi(T) I_{h^\star_j}(a)\right\}.$$

Substituting into equation 25 and adding the failure contribution $\mathcal{R}_{\text{fail}} \le \eta_T \Delta$ from Lemma 5 proves equation 23. □

Proof of Theorem 1. If $\mathcal{G}_T = \emptyset$, the refined theorem gives zero regret. Otherwise, we apply Theorem 6 with $a_T = \gamma_T/2$. We have $(h^\star_j - a_T)_+ \le 1$ and, by Lemma 11, $I_{h^\star_j}(a_T) \le 2/a_T = 4/\gamma_T$.

For $\phi(T) = \log T + \log\log T$, Lemma 8 gives $\eta_T \le 4eK\phi(T)$. Consequently,

$$R_T(\pi^{\texttt{KL-PBD}}) \le 4eK\phi(T)\Delta + 2c_e \sum_{j=1}^{K} \min\left\{\omega_j T,\ \bar{B} + \frac{4\phi(T)}{\gamma_T}\right\} + 2\bar{B}c_e\left[1 + 4\sum_{j=1}^{K}\log\left(1 + \frac{\omega_j T}{\phi(T)}\right)\right],$$

which proves the main-text theorem since $\kappa = 4e\Delta$. □

In particular, the main-text bound implies

$$R_T(\pi^{\texttt{KL-PBD}}) \le 2\bar{B}c_e(K+1) + K\phi(T)\left(\frac{8c_e}{\gamma_T} + 4e\Delta\right) + 8\bar{B}c_e\Lambda_T.$$

When the gap is bounded away from zero, the regret remains logarithmic in T. The failure contribution now has the same logarithmic order, but does not worsen the inverse-gap dependence. The refined theorem and equation 35 retain the variance-sensitive improvement.

B.5 Technical lemmas

B.5.1 Concentration and KL confidence bounds

Lemma 8 (Concentration within each self-confidence level under delayed feedback)**.** *For every $j \in [K]$, the observed-label sequence can be extended to an infinite i.i.d.* $\mathrm{Bernoulli}(h_j^\star)$ *sequence. For any fixed budget $\phi(T) > 1$, the event $\mathcal{E}$ of equation 12 satisfies $\mathbb{P}(\mathcal{E}^c) \le \eta_T/T$. For $T \ge 3$ and $\phi(T) = \log T + \log\log T$, we also have*

$$\eta_T \le 4eK\phi(T). \tag{26}$$

Proof. Conditional on the self-confidence level sequence, the ground-truth labels are independent, with $Y_t \sim \mathrm{Bernoulli}(h_j^\star)$ when $J_t = j$. Since deferral and completion decisions use only self-confidence levels, service information and previously revealed labels, conditional on the history immediately before a level-j label is revealed, that label has distribution $\mathrm{Bernoulli}(h_j^\star)$. Revealing completed labels in submission order within each round, and appending independent $\mathrm{Bernoulli}(h_j^\star)$ labels after the observed sequence for level j, gives the claimed infinite i.i.d. sequence (Lattimore & Szepesvári, 2020, Section 4.6).

For a Bernoulli sample mean, Chernoff's inequality and a union bound give

$$\mathbb{P}(\mathcal{E}^c) \le 2KTe^{-\phi(T)}.$$

For $T \ge 2$, applying the Bernoulli maximal inequality (Garivier, 2013, equation (5)) to the ground-truth label sequence for each self-confidence level and taking a union bound over levels gives

$$\mathbb{P}(\mathcal{E}^c) \le 2eK\lceil\phi(T)\log T\rceil e^{-\phi(T)}, \qquad \phi(T) > 1. \tag{27}$$

If $h_j^\star \in \{0,1\}$, then $\widehat{h}_{j,n} = h_j^\star$ almost surely for every $n \ge 1$, so the KL concentration event holds automatically for that j.

Combining these bounds with $\mathbb{P}(\mathcal{E}^c) \le 1$ gives $\mathbb{P}(\mathcal{E}^c) \le \eta_T/T$ for $T \ge 2$. For $T = 1$, the same conclusion follows from the Chernoff union bound and the definition of η_1. Picking $\phi(T) = \log T + \log\log T$ gives, using $Te^{-\phi(T)} = (\log T)^{-1}$,

$$\eta_T \le 2eK\frac{\lceil\phi(T)\log T\rceil}{\log T} \le 2eK\left(\phi(T) + \frac{1}{\log T}\right) \le 4eK\phi(T).$$

The last inequality uses $\log T > 1$ and $\phi(T) > 1$. □

Lemma 9 (KL diameter and its integral bounds)**.** *For every $p \in [0,1]$ and $x \ge 0$,*

$$w_p(x) \le \min\{p, \sqrt{2x}\}. \tag{28}$$

Whenever $n\,\mathrm{kl}(\widehat{h}_{j,n}, h_j^\star) \le \phi(T)$ *with* $n \ge 1$,

$$0 \le h_j^\star - g_{j,n} \le w_{h_j^\star}(\phi(T)/n). \tag{29}$$

Moreover, $F_{p,\Gamma}$ *is nondecreasing and concave, and*

$$F_{p,\Gamma}(d) \le \min\left\{pd,\, 2\sqrt{2\Gamma d}\right\}. \tag{30}$$

Proof. The full confidence set $\{q \in [0,1] : n\,\mathrm{kl}(\widehat{h}_{j,n}, q) \le \phi(T)\}$ is nonempty and compact. Its minimum is $g_{j,n}$, since it contains $\widehat{h}_{j,n}$. Under the stated KL constraint, it also contains $h_j^\star$. Thus $g_{j,n} \le h_j^\star$, and the definition of $w_{h_j^\star}$ gives equation 29.

The bound $w_p(x) \le p$ is immediate. Pinsker's inequality applied to each of the two KL constraints gives $|z-p|, |z-q| \le \sqrt{x/2}$, hence $p - q \le \sqrt{2x}$, proving equation 28.

As w_p is nondecreasing, the integrand defining $F_{p,\Gamma}$ is nonincreasing in u, so $F_{p,\Gamma}$ is nondecreasing and concave. Integrating the two bounds in equation 28 gives equation 30. □

Lemma 10 (Pathwise total variation)**.** *For any binary sequence and any fixed* $\Gamma > 0$*, let* $g_0 = 0$ *and let* g_n *be its Bernoulli KL lower confidence bound with budget* Γ*. For every integer* $N \ge 1$*,*

$$\sum_{n=1}^{N} |g_n - g_{n-1}| \le -2\log(1 - e^{-\Gamma/N}) \le 2\log(1 + N/\Gamma). \tag{31}$$

For `KL-PBD`*, on every sample path,*

$$\sum_{t=2}^{T} \|\underline{h}_t - \underline{h}_{t-1}\|_\infty \le 2\sum_{j=1}^{K} \log\left(1 + \frac{N_{j,T}}{\phi(T)}\right). \tag{32}$$

Consequently,

$$\mathbb{E}\left[\sum_{t=2}^{T} \|\underline{h}_t - \underline{h}_{t-1}\|_\infty\right] \le 2\Lambda_T. \tag{33}$$

Proof. We interpolate the number of successes s and failures f continuously. For $s > 0$, put $m = s + f$, $z = s/m$, and let $q = q(s,f) \in (0,z)$ solve

$$m\,\mathrm{kl}(z,q) = \Gamma.$$

For $s = 0$, set $q(0,f) = 0$, and set $q(0,0) = 0$. These definitions agree continuously at the boundaries relevant below: $q \le z$ gives continuity as $s \downarrow 0$ with $f > 0$, while $\mathrm{kl}(z,q) \le \mathrm{kl}(1,q) = -\log q$ gives $q \le e^{-\Gamma/m}$ as $m \downarrow 0$. In particular, $q(s,0) = e^{-\Gamma/s}$ for $s > 0$.

Consider first adding a success, keeping f fixed. In the interior, implicit differentiation gives

$$\frac{\partial}{\partial s}\big[m\,\mathrm{kl}(z,q)\big]_{q \text{ fixed}} = \log(z/q), \qquad \frac{\partial}{\partial q}\big[m\,\mathrm{kl}(z,q)\big] = -\frac{m(z-q)}{q(1-q)}.$$

Writing $a = \log(z/q) > 0$ and $x = \Gamma/m$, we obtain

$$\frac{\partial q}{\partial s} = \frac{1-q}{m}\frac{a}{e^a - 1} \le \frac{\Gamma}{m^2(e^{\Gamma/m} - 1)}. \tag{34}$$

Indeed, $x = za + (1-z)\log((1-z)/(1-q)) \le a$, and $a \mapsto a/(e^a - 1)$ is decreasing on $(0,\infty)$. The same inequality holds when $f = 0$, either directly from $q(s,0) = e^{-\Gamma/s}$ or by a limit.

Adding a failure while keeping s fixed gives instead the partial derivative $\log((1-z)/(1-q)) \le 0$ at fixed q. Implicit differentiation therefore gives $\partial q/\partial f \le 0$. Continuity extends this monotonicity to $f = 0$ and $s = 0$. Thus successes can only increase the lower confidence bound, and failures can only decrease it.

We interpolate each observation over its own interval $m \in [n-1, n]$. The total upward variation is at most the integral of the right-hand side of equation 34 over all intervals:

$$\sum_{n=1}^{N} (g_n - g_{n-1})_+ \le \int_0^N \frac{\Gamma}{m^2(e^{\Gamma/m} - 1)}\, dm = -\log(1 - e^{-\Gamma/N}).$$

By continuity of the lower confidence bound at $m = 0$ and $s = 0$, the integrals may start at 0. Since $g_0 = 0$ and $g_N \ge 0$, downward variation is no larger than upward variation. Finally, $e^x \ge 1 + x$ implies $-\log(1 - e^{-x}) \le \log(1 + 1/x)$, proving equation 31.

Fix a self-confidence level j. Since $N_{j,t}$ is nondecreasing and $\underline{h}_{j,t} = g_{j,N_{j,t}}$, the triangle inequality gives

$$\sum_{t=2}^{T} \left|\underline{h}_{j,t} - \underline{h}_{j,t-1}\right| \le \sum_{n=1}^{N_{j,T}} |g_{j,n} - g_{j,n-1}|.$$

Applying equation 31, with an empty sum when $N_{j,T} = 0$, and using

$$\left\|\underline{h}_t - \underline{h}_{t-1}\right\|_\infty \le \sum_{j=1}^{K} \left|\underline{h}_{j,t} - \underline{h}_{j,t-1}\right|$$

proves equation 32.

It remains to take expectations. Since $N_{j,T}$ counts completed self-confidence level j labels available before round T, it is bounded by the number of self-confidence level j arrivals before that round. Thus $\mathbb{E}[N_{j,T}] \le \omega_j T$. By concavity of $x \mapsto \log(1 + x)$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=2}^{T} \left\|\underline{h}_t - \underline{h}_{t-1}\right\|_\infty\right] &\le 2\sum_{j=1}^{K} \mathbb{E}\left[\log\left(1 + \frac{N_{j,T}}{\phi(T)}\right)\right] \\ &\le 2\sum_{j=1}^{K} \log\left(1 + \frac{\omega_j T}{\phi(T)}\right) = 2\Lambda_T.\end{aligned}$$

□

Lemma 11 (Integrated residual width)**.** *If $p \le a$, then $I_p(a) = 0$. For $p > a > 0$,*

$$I_p(a) \le \min\left\{\frac{2}{a},\, 8p(1-p)\left(\frac{1}{a} - \frac{1}{p}\right) + 12\log\frac{p}{a}\right\}. \tag{35}$$

Proof. For $p \le a$, $I_p(a) = 0$. Pinsker's bound gives, for $a > 0$,

$$I_p(a) \le \int_0^\infty (\sqrt{2/u} - a)_+\, du = \frac{2}{a}.$$

For the other bound, assume $p > a$ and put $A = 2\sqrt{2p(1-p)}$. By Tonelli's theorem,

$$I_p(a) = \int_a^p \mathrm{Leb}\{u > 0 : w_p(1/u) > r\}\, dr. \tag{36}$$

If $w_p(1/u) > r$, then equation 28 gives $r < A/\sqrt{u} + 6/u$. By $2xy \le x^2 + y^2$, applied with $x = \sqrt{r}$ and $y = A/\sqrt{ru}$,

$$\frac{A}{\sqrt{u}} \le \frac{r}{2} + \frac{A^2}{2ru}.$$

Consequently,

$$r < \frac{r}{2} + \frac{A^2}{2ru} + \frac{6}{u}, \qquad \text{so} \qquad u < \frac{A^2}{r^2} + \frac{12}{r}.$$

Thus

$$\mathrm{Leb}\{u > 0 : w_p(1/u) > r\} \le \frac{A^2}{r^2} + \frac{12}{r}.$$

This argument also applies when $A = 0$. Substituting this length bound into equation 36 yields

$$I_p(a) \le 8p(1-p)\left(\frac{1}{a} - \frac{1}{p}\right) + 12\log\frac{p}{a},$$

as claimed. In particular, all integrals used in the theorems are finite. □

B.5.2 SERVICE-CAPACITY STABILITY BOUNDS

Recall from Appendix A the queue value $U_t^h(b) := \sum_{j=0}^K \omega_j V_t^h((b,j))$. Since $V_{T+1}^h((b,j)) = D(b)$ for every $j = 0, \dots, K$ and $\sum_{j=0}^K \omega_j = 1$, we have $U_{T+1}^h = D$, independently of h, and the Bellman recursion can be written as

$$U_t^h(b) = c_{\text{hold}} b + \sum_{j=0}^{K} \omega_j \min_{a \in \mathcal{A}(b,j)} \left\{ c_{\text{exp}} a + L^h(j)(1-a) + \mathbb{E}\big[U_{t+1}^h((b+a-C_t)^+)\big] \right\}.$$

We first show that the service-capacity structure yields a stability property useful to control the effect of a pessimistic conditional accuracy vector on the Bellman recursion.

Proposition 2 (Bellman stability)**.** *Let $h, \widetilde{h} \in [0,1]^K$. For every $t \le T$:*

(i) For arbitrary $h, \widetilde{h}$, $\operatorname{sp}\left(U_t^h - U_t^{\widetilde{h}}\right) \le \bar{B} c_e \|h - \widetilde{h}\|_\infty$, and $\left|M_t^h(b) - M_t^{\widetilde{h}}(b)\right| \le c_e \|h - \widetilde{h}\|_\infty$ for every $b < \bar{B}$.

(ii) If $h \le \widetilde{h}$ coordinatewise, then $U_t^h - U_t^{\widetilde{h}}$ is nondecreasing in the queue state and, for every $b < \bar{B}$,

$$0 \le M_t^h(b) - M_t^{\widetilde{h}}(b) \le c_e \|h - \widetilde{h}\|_\infty.$$

Part (i) controls changes in the Bellman quantities under arbitrary conditional accuracy perturbations, while part (ii) shows that a pessimistic conditional accuracy vector increases the Bellman marginal deferral costs and induces an ordered change in the queue values.

Proof of Proposition 2. Let $h, \widetilde{h} \in [0,1]^K$ and set $\varepsilon := c_e \|h - \widetilde{h}\|_\infty$. Recall from Appendix A the notations $x_t^h(q) = U_t^h(q) - U_t^h(q-1)$ and $d_t^h(b) = F_t^h(b+1) - F_t^h(b)$. We prove both statements by a common backward induction. For arbitrary $h, \widetilde{h}$, we show that

$$\left| x_t^h(q) - x_t^{\widetilde{h}}(q) \right| \le \varepsilon, \tag{37}$$

and, if $h \le \widetilde{h}$ coordinatewise, that

$$0 \le x_t^h(q) - x_t^{\widetilde{h}}(q) \le \varepsilon, \tag{38}$$

for every $t \le T+1$ and $q = 1, \dots, \bar{B}$.

At $t = T+1$, both queue-value functions are equal. Hence $x_{T+1}^h(q) = x_{T+1}^{\widetilde{h}}(q)$ for every $q = 1, \dots, \bar{B}$, so both claims hold. Suppose they hold at time $t+1$. By equation 7, for every $b < \bar{B}$,

$$d_t^h(b) - d_t^{\widetilde{h}}(b) = \sum_{c=0}^{b} \rho_C(c) \left[x_{t+1}^h(b+1-c) - x_{t+1}^{\widetilde{h}}(b+1-c) \right].$$

The weights are nonnegative and sum to at most one. Therefore,

$$\left| d_t^h(b) - d_t^{\widetilde{h}}(b) \right| \le \varepsilon,$$

and, when $h \le \widetilde{h}$,

$$0 \le d_t^h(b) - d_t^{\widetilde{h}}(b) \le \varepsilon.$$

Recall the notation $r_j^h = L^h(j) - c_{\text{exp}}$. For $j \in [K]$, we have $r_j^h - r_j^{\widetilde{h}} = c_e(\widetilde{h}_j - h_j)$, while $r_0^h = r_0^{\widetilde{h}} = -c_{\text{exp}}$. Thus $|r_j^h - r_j^{\widetilde{h}}| \le \varepsilon$ for $j \in \{0\} \cup [K]$ and these differences are nonnegative when $h \le \widetilde{h}$.

For $1 \le q < \bar{B}$, the preceding bounds and equation 9 give

$$\begin{aligned} |x_t^h(q) - x_t^{\widetilde{h}}(q)| &\le \sum_{j=0}^{K} \omega_j \left| \operatorname{med}\{d_t^h(q-1), d_t^h(q), r_j^h\} - \operatorname{med}\{d_t^{\widetilde{h}}(q-1), d_t^{\widetilde{h}}(q), r_j^{\widetilde{h}}\} \right| \\ &\le \sum_{j=0}^{K} \omega_j \varepsilon = \varepsilon, \end{aligned}$$

where we used that for any $a, \widetilde{a} \in \mathbb{R}^3$, $|\mathrm{med}\{a_1, a_2, a_3\} - \mathrm{med}\{\widetilde{a}_1, \widetilde{a}_2, \widetilde{a}_3\}| \le \max_{i=1,2,3} |a_i - \widetilde{a}_i|$. At $q = \bar{B}$, the same conclusion follows from equation 10 and $|\max\{a_1, a_2\} - \max\{\widetilde{a}_1, \widetilde{a}_2\}| \le \max_{i=1,2} |a_i - \widetilde{a}_i|$. This proves equation 37 at time t.

If $h \le \widetilde{h}$, each argument of the median or maximum under h is at least the corresponding argument under $\widetilde{h}$. Since increasing any argument cannot decrease either its median or its maximum, equation 9 and equation 10 also give $x_t^h(q) \ge x_t^{\widetilde{h}}(q)$ for $q = 1, \dots, \bar{B}$. Combining this with the absolute bound proves equation 38 at time t, completing the common induction.

Now define $Z_t(b) := U_t^h(b) - U_t^{\widetilde{h}}(b)$. By equation 37, $|Z_t(b) - Z_t(b-1)| \le \varepsilon$ for $b = 1, \dots, \bar{B}$. Hence,

$$\mathrm{sp}(Z_t) \le \bar{B}\varepsilon = \bar{B} c_e \|h - \widetilde{h}\|_\infty.$$

Moreover, by equation 6, $M_t^h(b) - M_t^{\widetilde{h}}(b) = d_t^h(b) - d_t^{\widetilde{h}}(b)$, so

$$|M_t^h(b) - M_t^{\widetilde{h}}(b)| \le c_e \|h - \widetilde{h}\|_\infty, \qquad b < \bar{B}.$$

This proves part (i).

If $h \le \widetilde{h}$, equation 38 shows that Z_t is nondecreasing in the queue state. The corresponding bound on the smoothed increments also gives

$$0 \le M_t^h(b) - M_t^{\widetilde{h}}(b) \le c_e \|h - \widetilde{h}\|_\infty, \qquad b < \bar{B}.$$

This proves part (ii). □

We next record a uniform bound on the gaps Δ_t.

Lemma 12 (Uniform gap bound)**.** *For every conditional accuracy vector h, every $t \le T+1$, and every $q = 1, \dots, \bar{B}$,*

$$0 \le U_t^h(q) - U_t^h(q-1) \le \frac{q\, c_{\mathrm{hold}}}{\mu} + (c_e - c_{\mathrm{exp}})_+. \tag{39}$$

Consequently, for every $t \le T$ and every $b < \bar{B}$,

$$c_{\mathrm{exp}} \le M_t^h(b) \le \max\{c_e, c_{\mathrm{exp}}\} + \frac{c_{\mathrm{hold}} \bar{B}}{\mu}. \tag{40}$$

In particular, $0 \le \Delta_t \le \Delta$ on every sample path.

Proof. The lower bound in equation 39 follows from Lemma 3. For the upper bound, fix t and $q \ge 1$, and couple two systems starting from queue states q and $q-1$, respectively, using the same self-confidence level and service realizations through round T, and the same service realizations thereafter. Let the lower system follow an optimal policy.

To account for the terminal value, we regard both systems as continuing after round T with no further arrivals and incurring holding costs until their queues are empty; by definition, the expected cost of this continuation from queue state b is $D(b)$.

While the queues differ, they differ by exactly one job. If the lower system takes action 0, let the upper system also take action 0. If the lower system defers, let the upper system take action 0 instead. In the latter case, the two pre-service queue states after the deferral decision coincide, and hence the queues couple at the end of that round. Once they have coupled, let the upper system follow the same actions as the lower system.

Before coupling, the difference in holding cost is c_{hold} per round. If the lower system never triggers coupling by deferring, the queues couple once q rounds with $C_t \ge 1$ have occurred. The expected time required for q such service opportunities is q/μ. If coupling instead occurs because the lower system defers, the only additional non-holding cost incurred by the upper system on that round is $c_e(1 - h_{J_t}) - c_{\mathrm{exp}} \le (c_e - c_{\mathrm{exp}})_+$. This can occur at most once. Thus the constructed algorithm for the upper system has expected cost at most

$$U_t^h(q-1) + \frac{q\, c_{\mathrm{hold}}}{\mu} + (c_e - c_{\mathrm{exp}})_+.$$

Optimality of $U_t^h(q)$ proves equation 39.

By equation 7,

$$d_t^h(b) = \sum_{c=0}^{b} \rho_C(c)\, x_{t+1}^h(b+1-c).$$

Using equation 39 and $b+1 \le \bar{B}$,

$$0 \le d_t^h(b) \le \frac{c_{\text{hold}}\bar{B}}{\mu} + (c_e - c_{\text{exp}})_+.$$

Together with equation 6, this yields equation 40.

Finally, for $j \in [K]$, $c_e(1-h_j^\star) \in [0, c_e]$, while equation 40 gives $0 \le M_t^\star(b) \le \Delta$. Hence, for $b < \bar{B}$,

$$\left|c_e(1-h_j^\star) - M_t^\star(b)\right| \le \Delta.$$

On no-query or full-queue rounds $\Delta_t = 0$. Thus the definition of Δ_t gives $0 \le \Delta_t \le \Delta$ on every sample path.

□

B.5.3 Model comparison and pending labels

Lemma 13 (Bellman comparison)**.** *Let $\widetilde{h}_t \in [0,1]^K$ be $\mathcal{F}_{t-1}$-measurable and satisfy $\underline{h}_t \le \widetilde{h}_t$ coordinatewise for $t < \sigma$. For $j \in [K]$, define $e_{j,t} := c_e(\widetilde{h}_{j,t} - \underline{h}_{j,t})$, set $e_{0,t} := 0$, and let*

$$\widetilde{\Delta}_t := Q_t^{\widetilde{h}_t}(S_t, A_t) - V_t^{\widetilde{h}_t}(S_t)$$

be the decision gap of the learner's action under $\widetilde{h}_t$. Then

$$\mathbb{E}\left[\sum_{t<\sigma} \widetilde{\Delta}_t\right] \le \mathbb{E}\left[\sum_{t<\sigma} e_{J_t,t}\mathbf{1}\{A_t = 1\}\right] + \bar{B}c_e\left(1 + \mathbb{E}\left[\sum_{t=2}^{\sigma-1}\left(\|\underline{h}_t - \underline{h}_{t-1}\|_\infty + \|\widetilde{h}_t - \widetilde{h}_{t-1}\|_\infty\right)\right]\right). \tag{41}$$

Proof. Fix $t < \sigma$. For $t \le s \le T+1$, define $Z_s^t(b) := U_s^{\underline{h}_t}(b) - U_s^{\widetilde{h}_t}(b)$. By Proposition 2(ii), Z_s^t is nondecreasing in b. Define the centered potential $\Phi_s^t(b) := Z_s^t(\bar{B}) - Z_s^t(b) \ge 0$, and write $\bar{e}_t := \sum_{j=1}^K \omega_j e_{j,t}$.

At queue state $\bar{B}$, action 0 is the only feasible action. Therefore, for $t \le s \le T$,

$$Z_s^t(\bar{B}) = \bar{e}_t + \mathbb{E}\big[Z_{s+1}^t((\bar{B} - C_s)^+) \mid \mathcal{F}_{t-1}\big] \le \bar{e}_t + Z_{s+1}^t(\bar{B}),$$

and hence

$$Z_s^t(\bar{B}) - \bar{e}_t \le Z_{s+1}^t(\bar{B}). \tag{42}$$

Since the `KL-PBD` action at round t is optimal under $\underline{h}_t$, the definitions of Q_t^h and V_t^h give

$$\widetilde{\Delta}_t = V_t^{\underline{h}_t}(S_t) - V_t^{\widetilde{h}_t}(S_t) - e_{J_t,t}(1 - A_t) - \mathbb{E}\big[Z_{t+1}^t(B_{t+1}) \mid \mathcal{F}_{t-1}, J_t, A_t\big]. \tag{43}$$

Because J_t is independent of $\mathcal{F}_{t-1}$ with distribution ω,

$$\mathbb{E}[e_{J_t,t} \mid \mathcal{F}_{t-1}] = \sum_{j=0}^{K} \omega_j e_{j,t} = \bar{e}_t,$$

and

$$\mathbb{E}\Big[V_t^{\underline{h}_t}(S_t) - V_t^{\widetilde{h}_t}(S_t) \mid \mathcal{F}_{t-1}\Big] = Z_t^t(B_t).$$

Taking conditional expectation in equation 43 therefore yields

$$\mathbb{E}[\widetilde{\Delta}_t \mid \mathcal{F}_{t-1}] = Z_t^t(B_t) - \bar{e}_t + \mathbb{E}[e_{J_t,t}\mathbf{1}\{A_t = 1\} \mid \mathcal{F}_{t-1}] - \mathbb{E}\big[Z_{t+1}^t(B_{t+1}) \mid \mathcal{F}_{t-1}\big].$$

Using $Z_t^t(B_t) = Z_t^t(\bar{B}) - \Phi_t^t(B_t)$ and equation 42, we obtain

$$\mathbb{E}[\widetilde{\Delta}_t \mid \mathcal{F}_{t-1}] \le \mathbb{E}[e_{J_t,t}\mathbf{1}\{A_t = 1\} \mid \mathcal{F}_{t-1}] + \mathbb{E}\big[\Phi_{t+1}^t(B_{t+1}) \mid \mathcal{F}_{t-1}\big] - \Phi_t^t(B_t). \tag{44}$$

Since $\{t < \sigma\} \in \mathcal{F}_{t-1}$, we may multiply equation 44 by $\mathbf{1}\{t < \sigma\}$, take expectations, and sum over t. Let $m := \min\{T, \sigma - 1\}$. On paths with $m \ge 1$,

$$\sum_{t=1}^{m} \big[\Phi_{t+1}^t(B_{t+1}) - \Phi_t^t(B_t)\big] = -\Phi_1^1(B_1) + \Phi_{m+1}^m(B_{m+1}) + \sum_{t=2}^{m} \big[\Phi_t^{t-1}(B_t) - \Phi_t^t(B_t)\big].$$

The first term is nonpositive. If $m = T$, then $Z_{T+1}^T = D - D = 0$, and therefore $\Phi_{T+1}^T \equiv 0$. Otherwise, $m + 1 = \sigma$, and Proposition 2(i) gives

$$0 \le \Phi_\sigma^{\sigma-1}(B_\sigma) \le \bar{B}c_e\|\underline{h}_{\sigma-1} - \widetilde{h}_{\sigma-1}\|_\infty \le \bar{B}c_e.$$

It remains to control the change of potential when the conditional accuracy vectors are updated. Successive vectors need not be coordinatewise ordered, so here we use the two-sided stability in Proposition 2(i). For $2 \le t < \sigma$,

$$\begin{aligned}
|\Phi_t^{t-1}(B_t) - \Phi_t^t(B_t)| &= \big|(Z_t^{t-1} - Z_t^t)(\bar{B}) - (Z_t^{t-1} - Z_t^t)(B_t)\big| \\
&\le \operatorname{sp}(Z_t^{t-1} - Z_t^t) \\
&\le \operatorname{sp}\left(U_t^{\underline{h}_{t-1}} - U_t^{\underline{h}_t}\right) + \operatorname{sp}\left(U_t^{\widetilde{h}_{t-1}} - U_t^{\widetilde{h}_t}\right) \\
&\le \bar{B}c_e\left(\|\underline{h}_t - \underline{h}_{t-1}\|_\infty + \|\widetilde{h}_t - \widetilde{h}_{t-1}\|_\infty\right).
\end{aligned}$$

Here the third line uses subadditivity of the span. Combining these bounds controls the potential sum by

$$\bar{B}c_e\left(1 + \sum_{t=2}^{\sigma-1}\left(\|\underline{h}_t - \underline{h}_{t-1}\|_\infty + \|\widetilde{h}_t - \widetilde{h}_{t-1}\|_\infty\right)\right).$$

On paths with $m = 0$, the potential sum is empty and the same upper bound holds. Taking expectations and substituting into the summed one-step inequality proves equation 41. □

Lemma 14 (Pending-label bound for deferred queries)**.** *For $j \in [K]$ and any nonnegative nonincreasing function $f : [0,\infty) \to [0,\infty)$,*

$$\mathbb{E}\left[\sum_{t<\sigma} \mathbf{1}\{J_t = j, A_t = 1\} f(N_{j,t})\right] \le \min\left\{f(0)\omega_j T,\ \bar{B}f(0) + \int_0^{\omega_j T} f(u)\,du\right\}.$$

Proof. Fix a self-confidence level j, write $D_j := D_{j,\sigma-1}$, and enumerate its deferrals before σ as $t_{j,1}, \ldots, t_{j,D_j}$. Immediately before the ℓ-th such deferral, $\ell - 1$ previous self-confidence level j jobs have been submitted. Since deferral is feasible at $t_{j,\ell}$, at most $\bar{B} - 1$ of them can still be pending. Hence

$$N_{j,t_{j,\ell}} \ge (\ell - \bar{B})_+. \tag{45}$$

For the first at most $\bar{B}$ deferrals, use $f(N_{j,t_{j,\ell}}) \le f(0)$. For $\ell > \bar{B}$, monotonicity of f and equation 45 give $f(N_{j,t_{j,\ell}}) \le f(\ell - \bar{B})$. Therefore,

$$\sum_{\ell=1}^{D_j} f(N_{j,t_{j,\ell}}) \le f(0)\min\{D_j, \bar{B}\} + \sum_{n=1}^{(D_j - \bar{B})_+} f(n).$$

Since f is nonincreasing, $f(n) \le \int_{n-1}^{n} f(u)\,du$. Thus

$$\sum_{\ell=1}^{D_j} f(N_{j,t_{j,\ell}}) \le f(0)\min\{D_j, \bar{B}\} + \int_0^{(D_j-\bar{B})_+} f(u)\,du$$
$$\le \bar{B}f(0) + \int_0^{D_j} f(u)\,du.$$

Every such deferral occurs on a self-confidence level j arrival, so $\mathbb{E}[D_j] \le \omega_j T$. The map $d \mapsto \int_0^d f(u)\,du$ is nondecreasing and concave because f is nonnegative and nonincreasing. Consequently, Jensen's inequality gives

$$\mathbb{E}\left[\sum_{\ell=1}^{D_j} f(N_{j,t_{j,\ell}})\right] \le \bar{B}f(0) + \mathbb{E}\left[\int_0^{D_j} f(u)\,du\right]$$
$$\le \bar{B}f(0) + \int_0^{\mathbb{E}[D_j]} f(u)\,du$$
$$\le \bar{B}f(0) + \int_0^{\omega_j T} f(u)\,du.$$

On the other hand, $f(N_{j,t}) \le f(0)$ gives

$$\mathbb{E}\left[\sum_{\ell=1}^{D_j} f(N_{j,t_{j,\ell}})\right] \le f(0)\mathbb{E}[D_j] \le f(0)\omega_j T.$$

Taking the better of these two bounds proves the claim. □

B.6 Extension to unknown arrival and service distributions

The arrival distribution ω and the truncated service distribution defined below can be learned without additional deferrals when their realizations are observed independently of the deferral decisions. In this section, we show that using `KL-PBD` with plug-in empirical estimates of these distributions preserves the worst-case regret rate of Theorem 2, including its dependence on the arrival distribution ω. Throughout this section, assume $c_{\text{hold}} > 0$. For unknown ρ_C, additionally assume that $\bar{C}_t := \min\{C_t, \bar{B}\}$ is observed at the end of every round, including rounds with an empty queue. Write $\bar{\rho}_c := \mathbb{P}(\bar{C}_t = c)$ for $c = 0, \dots, \bar{B}$, so $\mu = 1 - \bar{\rho}_0 > 0$, and set $\Omega := \mathbb{P}(J_t \ne 0) = 1 - \omega_0$. Only this truncated service distribution affects the queue dynamics and terminal cost. We write $Q_t^{h,\omega,\bar{\rho}}$, $V_t^{h,\omega,\bar{\rho}}$ and $M_t^{h,\omega,\bar{\rho}}$ for the corresponding Bellman quantities.

Estimators of ω and $\bar{\rho}$. For $t \ge 1$ and $t \ge 2$ respectively, we form the estimates

$$\widehat{\omega}_{j,t} = \frac{1}{t}\sum_{s\le t} \mathbf{1}\{J_s = j\}, \quad j = 0, \dots, K, \qquad \widehat{\rho}_{c,t} = \frac{1}{t-1}\sum_{s<t} \mathbf{1}\{\bar{C}_s = c\}, \quad c = 0, \dots, \bar{B}. \tag{46}$$

The arrival estimate uses the current level J_t, which is observed before the decision; the capacity $\bar{C}_t$ is realized after it.

When $\bar{\rho}$ is known, set $\widetilde{\rho}_t := \bar{\rho}$. When service is estimated and $t \ge 2$, set $\widehat{\mu}_t := 1 - \widehat{\rho}_{0,t}$ and, if $\widehat{\mu}_t > 0$, set $\widetilde{\rho}_t := \widehat{\rho}_t$. Whenever $\widetilde{\rho}_t$ is defined, we solve the remaining-horizon Bellman recursion using $\widehat{\omega}_t$ and $\widetilde{\rho}_t$, with terminal cost $D_{\widetilde{\rho}_t}$, and choose

$$A_t = \mathbf{1}\big\{J_t \in [K],\ \underline{h}_{J_t,t} < \widehat{\tau}_t(B_t)\big\}, \qquad \widehat{\tau}_t(b) := 1 - \frac{M_t^{\underline{h}_t,\widehat{\omega}_t,\widetilde{\rho}_t}(b)}{c_e}, \tag{47}$$

with the convention $M_t(\bar{B}) = +\infty$ of Section 3. If service is estimated and either $t = 1$ or $\widehat{\mu}_t = 0$, we set $A_t = 0$ instead.

Define

$$\Psi_T(\bar{\rho}) := \sum_{c=0}^{\bar{B}} \min\{\bar{\rho}_c T, \sqrt{\bar{\rho}_c(1-\bar{\rho}_c)T}\}, \qquad S_T(\omega) := \sum_{j=1}^{K} \min\{\omega_j T, \sqrt{\omega_j T}\}.$$

Denote the known-parameter worst-case bound by

$$\mathcal{B}_T := K\kappa\phi(T) + c_e \sum_{j=1}^{K} \min\{\omega_j T, \bar{B} + 2\sqrt{2\omega_j T\phi(T)}\} + \bar{B}c_e \left[1 + 2\sum_{j=1}^{K} \log\left(1 + \frac{\omega_j T}{\phi(T)}\right)\right] \tag{48}$$

where we recall $\kappa = 4e\Delta = 4e(\max\{c_e, c_{\exp}\} + c_{\text{hold}}\bar{B}/\mu)$. For $\mu_0 > 0$, set

$$L_{\mu_0} := \frac{\bar{B}c_{\text{hold}}}{\mu_0} + (c_e - c_{\exp})_+, \qquad a_\omega(\mu_0) := \frac{c_e L_{\mu_0}}{c_{\text{hold}}}, \qquad a_C(\mu_0) := \frac{L_{\mu_0}^2}{c_{\text{hold}}} + L_{\mu_0} + \frac{c_{\text{hold}}\bar{B}^2}{\mu_0^2}. \tag{49}$$

In the rest of this section, we prove the following theorem.

Theorem 7 (Regret of `KL-PBD` for unknown ω and ρ_C)**.** *For $T \geq 3$, the plug-in variants of `KL-PBD` defined above satisfy*

$$R_T \leq \mathcal{B}_T + 4\, a_\omega(\mu) S_T(\omega), \qquad \textit{if } \bar{\rho} \textit{ is known}, \tag{50}$$

$$R_T \leq \mathcal{B}_T + \Omega\Delta(1 + 8/\mu) + 4\, a_\omega(\mu/2) S_T(\omega) + 2\Omega\, a_C(\mu/2)\Psi_T(\bar{\rho}), \qquad \textit{if } \bar{\rho} \textit{ is estimated}. \tag{51}$$

Both bounds are $\widetilde{O}(K + S_T(\omega))$, uniformly over arrival distributions.

Theorem 7 shows that learning these distributions requires no additional exploration under the stated observation assumptions and does not worsen the worst-case rate of Theorem 2.

Proof of Theorem 7. All Bellman quantities used for comparison below are evaluated at the fixed true distributions $(\omega, \bar{\rho})$. We use the history from Appendix A and set

$$e_t := Q_t^{\underline{h}_t, \omega, \bar{\rho}}(S_t, A_t) - V_t^{\underline{h}_t, \omega, \bar{\rho}}(S_t) \geq 0.$$

By Lemma 12, $e_t \leq \Delta \mathbf{1}\{J_t \neq 0\}$; also $e_t = 0$ when $B_t = \bar{B}$. Whenever A_t is computed from equation 47, equation 4 gives

$$e_t \leq \mathbf{1}\{J_t \neq 0\} \sup_{b < \bar{B}} \left| M_t^{\underline{h}_t, \widehat{\omega}_t, \widetilde{\rho}_t}(b) - M_t^{\underline{h}_t, \omega, \bar{\rho}}(b) \right|. \tag{52}$$

The proof of Theorem 2 then extends as

$$R_T \leq \mathcal{B}_T + \sum_{t=1}^{T} \mathbb{E}[e_t]. \tag{53}$$

To see this, we use the proof of Lemma 13 with $\widetilde{h}_t = h^\star$.

The only use of exact optimality under $\underline{h}_t$ is in equation 43. Here, by definition of e_t,

$$Q_t^{\underline{h}_t, \omega, \bar{\rho}}(S_t, A_t) = V_t^{\underline{h}_t, \omega, \bar{\rho}}(S_t) + e_t,$$

so equation 43 has the additional term e_t, and equation 44 has the additional term $\mathbb{E}[e_t \mid \mathcal{F}_{t-1}]$.

The potential bounds from Proposition 2 and the subsequent telescoping argument are unchanged. Therefore, equation 41, and hence the decomposition in equation 20, hold with the additional term $\mathbb{E}[\sum_{t<\sigma} e_t]$.

The bounds in Lemmas 5, 6 and 7 still apply: the concentration, pending-label counting and lower-confidence-bound variation arguments in Lemmas 8, 14 and 10 hold for adaptive feasible actions. The same simplification as in the proof of Theorem 2, together with $e_t \geq 0$, proves equation 53. No variation bound for $\widehat{\omega}_t$ or $\widehat{\rho}_t$ is needed, because the Bellman comparison in the proof is carried out under the fixed true distributions $(\omega, \bar{\rho})$; the effect of using the empirical distributions to choose A_t is isolated in e_t.

For an empirical distribution $\widehat{v}_n$ based on n i.i.d. samples from v, we have $\mathbb{E}[\widehat{v}_{i,n}] = v_i$ and $\mathrm{Var}(\widehat{v}_{i,n}) = v_i(1-v_i)/n$. Cauchy–Schwarz and $|\widehat{v}_{i,n} - v_i| \le \widehat{v}_{i,n} + v_i$ therefore give

$$\mathbb{E}[|\widehat{v}_{i,n} - v_i|] \le \min\{2v_i, \sqrt{v_i(1-v_i)/n}\}, \qquad \sum_{n=1}^{T} \mathbb{E}[|\widehat{v}_{i,n} - v_i|] \le 2\min\{v_i T, \sqrt{v_i(1-v_i)T}\}, \tag{54}$$

where the second inequality uses $\sum_{n=1}^{T} n^{-1/2} \le 2\sqrt{T}$. Since $\widehat{\omega}_t$ and ω are probability vectors, $|\widehat{\omega}_{0,t} - \omega_0| = \left|\sum_{j=1}^{K}(\widehat{\omega}_{j,t} - \omega_j)\right|$, so

$$\|\widehat{\omega}_t - \omega\|_1 \le 2\sum_{j=1}^{K} |\widehat{\omega}_{j,t} - \omega_j|, \qquad \sum_{t=1}^{T} \mathbb{E}[\|\widehat{\omega}_t - \omega\|_1] \le 4S_T(\omega). \tag{55}$$

Similarly, since $\widehat{\rho}_t$ is based on $t-1$ samples, equation 54 gives

$$\sum_{t=2}^{T} \mathbb{E}[\|\widehat{\rho}_t - \bar{\rho}\|_1] \le 2\Psi_T(\bar{\rho}). \tag{56}$$

For known $\bar{\rho}$, applying Lemma 15 with $\mu_0 = \mu$ in equation 52 and bounding the indicator by one gives $\mathbb{E}[e_t] \le a_\omega(\mu)\mathbb{E}[\|\widehat{\omega}_t - \omega\|_1]$ for every $t \ge 1$. Equation 53 and equation 55 prove equation 50.

For estimated $\bar{\rho}$, let $G_t := \{\widehat{\mu}_t \ge \mu/2\}$ for $t \ge 2$. On G_t, Lemma 15 applies with $\mu_0 = \mu/2$; on G_t^c, we have $e_t \le \Delta \mathbf{1}\{J_t \ne 0\}$.

Both G_t and $\widehat{\rho}_t$ are $\mathcal{F}_{t-1}$-measurable, whereas J_t is independent of $\mathcal{F}_{t-1}$. Bounding $\mathbf{1}\{J_t \ne 0\}$ by one in the $\widehat{\omega}_t$ term and using independence in the other two terms therefore gives, for $t \ge 2$,

$$\mathbb{E}[e_t] \le a_\omega(\mu/2)\mathbb{E}[\|\widehat{\omega}_t - \omega\|_1] + \Omega\, a_C(\mu/2)\mathbb{E}[\|\widehat{\rho}_t - \bar{\rho}\|_1] + \Omega\Delta\mathbb{P}(G_t^c).$$

The binomial lower-tail bound gives $\mathbb{P}(G_t^c) \le e^{-(t-1)\mu/8}$, so that $\sum_{t=2}^{T} \mathbb{P}(G_t^c) \le 8/\mu$. Together with $\mathbb{E}[e_1] \le \Omega\Delta$, equation 55, equation 56 and equation 53 prove equation 51.

Finally, for $T \ge 1$,

$$S_T(\omega) \ge \Omega\sqrt{T}, \qquad \Omega\Psi_T(\bar{\rho}) \le \Omega\sqrt{\bar{B}T} \le \sqrt{\bar{B}}\, S_T(\omega).$$

Indeed, $S_T(\omega) \ge \Omega\sqrt{T}$ follows from $\min\{\omega_j T, \sqrt{\omega_j T}\} \ge \omega_j\sqrt{T}$ for every $j \in [K]$. Moreover, Cauchy–Schwarz gives $\sum_{c=0}^{\bar{B}} \sqrt{\bar{\rho}_c(1-\bar{\rho}_c)} \le \sqrt{\bar{B}}$ and hence $\Psi_T(\bar{\rho}) \le \sqrt{\bar{B}T}$. Thus the additional terms are $O(S_T(\omega))$, with constants independent of ω. If $\Omega = 0$, all actions are forced and regret is zero. □

Lemma 15 (Uniform stability of the marginal cost)**.** *Fix $h \in [0,1]^K$, with $h_0 = 1$. For two models*[2] *(ω, ρ) and (ω', ρ') satisfying $1-\rho_0, 1-\rho_0' \ge \mu_0 > 0$, for every $t \le T$ and $b < \bar{B}$,*

$$\left|M_t^{h,\omega,\rho}(b) - M_t^{h,\omega',\rho'}(b)\right| \le a_\omega(\mu_0)\|\omega - \omega'\|_1 + a_C(\mu_0)\|\rho - \rho'\|_1. \tag{57}$$

Proof. We use the notation x_t, d_t and r_j from the proof of Lemma 3, suppressing model superscripts. The recursions in equation 7, equation 9 and equation 10 apply with $\rho_C(c) = \rho_c$ for $c < \bar{B}$ and terminal cost D_ρ.

We interpolate linearly between the two models:

$$(\omega^\lambda, \rho^\lambda) := (1-\lambda)(\omega, \rho) + \lambda(\omega', \rho'), \qquad \lambda \in [0,1].$$

Since $1 - \rho_0^\lambda \ge \mu_0$, equation 39 gives $0 \le x_t(i) \le L_{\mu_0}$ for every $\lambda \in [0,1]$. We bound $\mathrm{d}M_t(b)/\mathrm{d}\lambda$ uniformly in λ and then integrate over $[0,1]$.

For the model with service law ρ^λ, define $g(i) := D_{\rho^\lambda}(i) - D_{\rho^\lambda}(i-1)$. Subtracting the defining recursions for $D_{\rho^\lambda}(i)$ and $D_{\rho^\lambda}(i-1)$ gives, for $1 \le i \le \bar{B}$,

$$(1-\rho_0^\lambda)g(i) = c_{\mathrm{hold}} + \sum_{c=1}^{i-1} \rho_c^\lambda g(i-c),$$

[2]Each model uses its own terminal draining cost.

where the sum is empty for $i = 1$. Hence each $g(i)$ is absolutely continuous in λ. The recursions in equation 7, equation 9 and equation 10 then imply by backward induction that all $d_t(b)$ and $x_t(i)$ are absolutely continuous in λ, since these recursions involve only finite sums, products, and the Lipschitz maps $\min$, $\max$ and med. Below, derivatives are taken at values of λ where these functions are differentiable, which holds almost everywhere. Let

$$F := c_e\|\omega' - \omega\|_1 + L_{\mu_0}\|\rho' - \rho\|_1, \qquad G := \frac{c_{\mathrm{hold}}\bar{B}^2}{\mu_0^2}\|\rho' - \rho\|_1.$$

We first bound $\mathrm{d}g(i)/\mathrm{d}\lambda$. Since $\sum_{c=1}^{i-1}\rho_c^\lambda \le 1 - \rho_0^\lambda$ and $1 - \rho_0^\lambda \ge \mu_0$, the preceding recursion gives $g(i) \le c_{\mathrm{hold}}/\mu_0 + \max_{k<i} g(k)$, where the maximum over an empty set is zero, so that $0 \le g(i) \le ic_{\mathrm{hold}}/\mu_0$. Differentiating the recursion gives

$$(1-\rho_0^\lambda)\frac{\mathrm{d}g(i)}{\mathrm{d}\lambda} = \sum_{c=1}^{i-1}\rho_c^\lambda\frac{\mathrm{d}g(i-c)}{\mathrm{d}\lambda} + \sum_{c=0}^{i-1}(\rho_c' - \rho_c)g(i-c),$$

so that

$$\left|\frac{\mathrm{d}g(i)}{\mathrm{d}\lambda}\right| \le \max_{k<i}\left|\frac{\mathrm{d}g(k)}{\mathrm{d}\lambda}\right| + \frac{c_{\mathrm{hold}}\bar{B}}{\mu_0^2}\|\rho' - \rho\|_1.$$

A second induction on $i \le \bar{B}$ gives $|\mathrm{d}g(i)/\mathrm{d}\lambda| \le G$.

We next bound $\mathrm{d}x_t(i)/\mathrm{d}\lambda$ recursively in terms of $\mathrm{d}x_{t+1}(k)/\mathrm{d}\lambda$. Define

$$\psi_{j,t}(i) := \begin{cases}\mathrm{med}\{d_t(i-1), d_t(i), r_j\}, & i < \bar{B},\\ \max\{d_t(\bar{B}-1), r_j\}, & i = \bar{B}.\end{cases}$$

By equation 9 and equation 10,

$$x_t(i) = c_{\mathrm{hold}} + \sum_{j=0}^{K}\omega_j^\lambda\psi_{j,t}(i), \qquad \frac{\mathrm{d}x_t(i)}{\mathrm{d}\lambda} = \sum_{j=0}^{K}(\omega_j' - \omega_j)\psi_{j,t}(i) + \sum_{j=0}^{K}\omega_j^\lambda\frac{\mathrm{d}\psi_{j,t}(i)}{\mathrm{d}\lambda}. \tag{58}$$

At almost every λ, the derivative of each median or maximum equals the derivative of one of the arguments attaining its value (when several arguments attain it on a set of positive measure, their derivatives agree almost everywhere there). For each j, choose such an argument, and let $\mathcal{J}_{t,i}$ be the set of j for which it is some $d_t(b_j)$. Since h is fixed, $\mathrm{d}r_j/\mathrm{d}\lambda = 0$. Hence $\psi_{j,t}(i) = d_t(b_j)$ and $\mathrm{d}\psi_{j,t}(i)/\mathrm{d}\lambda = \mathrm{d}d_t(b_j)/\mathrm{d}\lambda$ for $j \in \mathcal{J}_{t,i}$, while $\psi_{j,t}(i) = r_j$ and $\mathrm{d}\psi_{j,t}(i)/\mathrm{d}\lambda = 0$ for $j \notin \mathcal{J}_{t,i}$. By equation 7,

$$\frac{\mathrm{d}d_t(b)}{\mathrm{d}\lambda} = \sum_{c=0}^{b}\rho_c^\lambda\frac{\mathrm{d}x_{t+1}(b+1-c)}{\mathrm{d}\lambda} + \sum_{c=0}^{b}(\rho_c' - \rho_c)x_{t+1}(b+1-c). \tag{59}$$

Substituting equation 59 into equation 58 for $j \in \mathcal{J}_{t,i}$ gives

$$\frac{\mathrm{d}x_t(i)}{\mathrm{d}\lambda} = \sum_{k=1}^{\bar{B}}a_{t,i,k}\frac{\mathrm{d}x_{t+1}(k)}{\mathrm{d}\lambda} + f_t(i), \tag{60}$$

where

$$a_{t,i,k} := \sum_{j\in\mathcal{J}_{t,i}}\sum_{c=0}^{b_j}\omega_j^\lambda\rho_c^\lambda\mathbf{1}\{k = b_j + 1 - c\} \ge 0, \qquad \sum_{k=1}^{\bar{B}}a_{t,i,k} = \sum_{j\in\mathcal{J}_{t,i}}\omega_j^\lambda\sum_{c=0}^{b_j}\rho_c^\lambda \le 1,$$

and

$$f_t(i) := \sum_{j=0}^{K}(\omega_j' - \omega_j)\psi_{j,t}(i) + \sum_{j\in\mathcal{J}_{t,i}}\omega_j^\lambda\sum_{c=0}^{b_j}(\rho_c' - \rho_c)x_{t+1}(b_j + 1 - c).$$

The second term of $f_t(i)$ is at most $L_{\mu_0}\|\rho' - \rho\|_1$ in absolute value, since $0 \le x_{t+1}(k) \le L_{\mu_0}$ and $\sum_j \omega_j^\lambda \le 1$. For the first, both $\mathrm{med}\{d_t(i-1), d_t(i), r\}$ and $\max\{d_t(\bar{B}-1), r\}$ are 1-Lipschitz in

r for fixed d_t, and $r_j = L^h(j) - c_{\text{exp}}$ with $L^h(j) \in [0, c_e]$, so the values $\psi_{j,t}(i)$ lie in an interval of length at most c_e. Since $\sum_j(\omega'_j - \omega_j) = 0$, the first term is at most $c_e\|\omega' - \omega\|_1$ in absolute value. Hence $|f_t(i)| \le F$.

The same coefficients satisfy

$$x_t(i) \ge c_{\text{hold}} + \sum_{k=1}^{\bar{B}} a_{t,i,k} x_{t+1}(k). \tag{61}$$

Indeed, for $j \in \mathcal{J}_{t,i}$, $\psi_{j,t}(i) = d_t(b_j) = \sum_{c=0}^{b_j} \rho_c^\lambda x_{t+1}(b_j + 1 - c)$, so these terms of equation 58 sum to $\sum_k a_{t,i,k} x_{t+1}(k)$. For $j \notin \mathcal{J}_{t,i}$, $\psi_{j,t}(i) = r_j \ge 0$: r_j attaining the median (resp. maximum) implies $r_j \ge d_t(i-1)$ (resp. $r_j \ge d_t(\bar{B}-1)$), which is nonnegative by equation 8.

We now prove by backward induction that

$$\left|\frac{\mathrm{d}x_t(i)}{\mathrm{d}\lambda}\right| \le \frac{F}{c_{\text{hold}}} x_t(i) + G. \tag{62}$$

At $t = T+1$, this follows from $x_{T+1}(i) = g(i) \ge 0$ and $|\mathrm{d}g(i)/\mathrm{d}\lambda| \le G$. If equation 62 holds at $t+1$, then equation 60 and equation 61 give

$$\left|\frac{\mathrm{d}x_t(i)}{\mathrm{d}\lambda}\right| \le \sum_{k=1}^{\bar{B}} a_{t,i,k} \left(\frac{F}{c_{\text{hold}}} x_{t+1}(k) + G\right) + F \le \frac{F}{c_{\text{hold}}}\big(x_t(i) - c_{\text{hold}}\big) + G + F = \frac{F}{c_{\text{hold}}} x_t(i) + G.$$

Thus $|\mathrm{d}x_t(i)/\mathrm{d}\lambda| \le FL_{\mu_0}/c_{\text{hold}} + G$.

Finally, $M_t(b) = c_{\text{exp}} + d_t(b)$ by equation 6, so equation 59 gives

$$\left|\frac{\mathrm{d}M_t(b)}{\mathrm{d}\lambda}\right| \le \frac{FL_{\mu_0}}{c_{\text{hold}}} + G + L_{\mu_0}\|\rho' - \rho\|_1 = a_\omega(\mu_0)\|\omega' - \omega\|_1 + a_C(\mu_0)\|\rho' - \rho\|_1.$$

Integrating over $\lambda \in [0,1]$ proves equation 57. □

C Proofs for Section 5

C.1 Proof sketches

To ease navigation within this appendix, we first provide high-level proof sketches of our instance-dependent and minimax lower bounds. They are proven in Appendices C.3 and C.5 respectively.

Proof sketch of Theorem 3. We first reduce to the long-horizon problem, in which the oracle threshold no longer depends on the round; passing from long-horizon to finite-horizon regret costs only an $O(1)$ additive term (Lemma 16). For a fixed level j with $h_j^\star > \overline{\tau}^\star$, the long-horizon optimal policy accepts level j at every queue state, whereas deferring incurs an extra cost of at least $c_e(h_j^\star - \overline{\tau}^\star)$. We compare $h^\star$ with an alternative instance in which only $h_j^\star$ is moved slightly below $\overline{\tau}^\star$. Although the deferral threshold changes with the instance, we show that deferring level j is strictly beneficial in the long run under the alternative (equation 78), so a uniformly efficient learner must behave differently on the two instances. Since only completed level-j reviews distinguish them (Lemma 17), a change-of-measure argument (Lai & Robbins, 1985) forces order $\log T / \operatorname{kl}(h_j^\star, \overline{\tau}^\star)$ level-j deferrals under $h^\star$. Multiplying by their extra cost $c_e(h_j^\star - \overline{\tau}^\star)$ yields the corresponding term in Theorem 3. Levels with $h_j^\star < \overline{\tau}^\star$ are instead always deferred by the oracle at the empty queue, so they can be reviewed without taking a suboptimal action. □

Proof sketch of Theorem 4. When $\mu = 1$, we move each conditional accuracy h_j a distance δ_j above or below the oracle's accuracy threshold, so that two neighboring instances differing only at level j have different optimal actions for level-j queries. With $\delta_j \asymp \min\{1, (\omega_j T)^{-1/2}\}$, Lemma 17 implies that the corresponding histories remain statistically indistinguishable, so on one of the two instances the learner must take the suboptimal action on a constant fraction of these queries, each costing order $c_e\delta_j$. When $\mu < 1$, perturbing the accuracies moves the threshold they are compared with. We therefore shift all coordinates of each perturbed instance by a common amount until the long-horizon empty-queue threshold coincides with the center of the perturbation (Lemma 21). An additional difficulty is that the two neighboring instances are only guaranteed to have different optimal actions when the queue is empty; by keeping the queue occupied, the learner could reduce the number of rounds on which indistinguishability translates into regret. Lemmas 20 and 22 show that this is not free: if the queue is rarely empty, the learner must defer sufficiently often while it is congested, at a cost of order c_{hold}. Capping δ_j at the holding-cost scale keeps this cost uniformly positive and produces the second term in the minimum, yielding Theorem 4. □

C.2 Common ingredients

We first collect the average-cost and information identities used by both lower-bound proofs. Their proofs and the supporting technical lemmas are given in Subsection C.6.

Throughout this appendix, ω, ρ_C, $\bar{B}$, $(c_e, c_{\mathrm{exp}}, c_{\mathrm{hold}})$ and the review order are fixed; only the conditional accuracy vector h varies. Write $\mathbb{P}_h^\pi$ and $\mathbb{E}_h^\pi$ for probability and expectation under h and algorithm π.

For a function $u : \mathcal{B} \to \mathbb{R}$ and $a \in \{0,1\}$ with $b + a \le \bar{B}$, write $(P_a u)(b) := \sum_{b'} P_a(b,b')u(b') = \mathbb{E}[u((b+a-C_t)^+)]$, and define the expected one-step cost $\ell_h(b,j,a) := c_{\mathrm{hold}}b + c_{\mathrm{exp}}a + L^h(j)(1-a)$. In this notation, the terminal cost satisfies $D(0) = 0$ and

$$D(b) = c_{\mathrm{hold}}b + P_0 D(b), \qquad b \ge 1. \tag{63}$$

We recall that a stationary policy φ is a policy whose action distribution $\varphi(\cdot \mid b, j)$ depends only on the current state (b, j), not on time or past history.

A pair $(g_h, u_h) \in \mathbb{R} \times \mathbb{R}^{\mathcal{B}}$ with $u_h(0) = 0$ solves the average-cost optimality equation (ACOE) for h if

$$g_h + u_h(b) = \sum_{j=0}^{K} \omega_j \min_{a \in \mathcal{A}(b,j)} \{\ell_h(b,j,a) + (P_a u_h)(b)\}, \qquad b \in \mathcal{B}. \tag{64}$$

Given such a pair, write $x_h(q) := u_h(q) - u_h(q-1)$ for $q \in [\bar{B}]$, $d_h(b) := (P_1 - P_0)u_h(b)$ and $\overline{M}^h(b) := c_{\mathrm{exp}} + d_h(b)$ for $b < \bar{B}$; recall $r_j^h = L^h(j) - c_{\mathrm{exp}}$. For a feasible action, define its

average-cost disadvantage by

$$\mathfrak{d}_h(b,j,a) := \ell_h(b,j,a) + (P_a u_h)(b) - \min_{a' \in \mathcal{A}(b,j)} \{\ell_h(b,j,a') + P_{a'} u_h(b)\} \ \geq 0. \tag{65}$$

For $s \geq 0$, let $W_s^h(b)$ be the queue value function with s decision rounds remaining and terminal cost D. Thus $W_0^h = D$ and $W_{T+1-t}^h = U_t^h$. Define $\xi_s^h(q) := W_s^h(q) - W_s^h(q-1)$, so that $x_t^h(q) = \xi_{T+1-t}^h(q)$.

Lemma 16 (Average-cost comparison)**.** *For every $h \in [0,1]^K$, the limits $u_h := \lim_{s\to\infty}[W_s^h - W_s^h(0)]$ and $g_h := \lim_{s\to\infty}[W_{s+1}^h(0) - W_s^h(0)]$ exist. The pair (g_h, u_h) solves equation 64; u_h is nondecreasing and discrete convex, and $u_h \geq D$. Moreover,*

(i) the following increment identities hold:

$$d_h(b) = \sum_{c=0}^{b} \rho_C(c)\, x_h(b+1-c), \tag{66}$$

$$x_h(q) = c_{\text{hold}} + \sum_{j=0}^{K} \omega_j \,\text{med}\{d_h(q-1), d_h(q), r_j^h\}, \qquad 1 \leq q < \bar{B}, \tag{67}$$

$$x_h(\bar{B}) = c_{\text{hold}} + \sum_{j=0}^{K} \omega_j \max\{d_h(\bar{B}-1), r_j^h\}; \tag{68}$$

(ii) $c_{\text{exp}} \leq \overline{M}^h(0) \leq \cdots \leq \overline{M}^h(\bar{B}-1)$, and

$$\overline{M}^h(0) = c_{\text{exp}} + (1-\mu)\, u_h(1); \tag{69}$$

(iii) for every algorithm π,

$$\left| R_T(\pi; h) - \mathbb{E}_h^\pi\Big[\sum_{t=1}^{T} \mathfrak{d}_h(B_t, J_t, A_t)\Big]\right| \leq \text{sp}(D - u_h); \tag{70}$$

(iv) for every stationary policy φ with a stationary distribution λ,

$$g_h \leq \sum_{s,a} \lambda(s)\varphi(a \mid s)\, \ell_h(s,a). \tag{71}$$

Recall $D_{j,t} := \sum_{s=1}^{t} \mathbf{1}\{J_s = j, A_s = 1\}$, the number of self-confidence level j deferrals submitted through round t, and that $N_{j,t}$ is the number of completed self-confidence level j labels available before round t. Let $P_h^{\pi,t}$ denote the law of $\mathcal{F}_t$ under h and algorithm π.

Lemma 17 (Information from completed reviews)**.** *For every $t \geq 0$, any algorithm π, and conditional accuracy vectors $h, \widetilde{h}$ such that $\widetilde{h}_j \in (0,1)$ whenever $h_j \neq \widetilde{h}_j$,*

$$\text{KL}(P_h^{\pi,t} \| P_{\widetilde{h}}^{\pi,t}) = \sum_{j=1}^{K} \mathbb{E}_h^\pi[N_{j,t+1}]\, \text{kl}(h_j, \widetilde{h}_j) \leq \sum_{j=1}^{K} \mathbb{E}_h^\pi[D_{j,t}]\, \text{kl}(h_j, \widetilde{h}_j). \tag{72}$$

Consequently,

$$\text{KL}(P_h^{\pi,t} \| P_{\widetilde{h}}^{\pi,t}) \leq t \sum_{j=1}^{K} \omega_j\, \text{kl}(h_j, \widetilde{h}_j). \tag{73}$$

Define

$$M_\tau := c_e(1-\tau) = c_{\text{exp}} + \frac{1-\mu}{\mu} c_{\text{hold}}, \qquad h^0 := \tau \mathbf{1}. \tag{74}$$

Lemma 18 (Empty-queue threshold bounds)**.** *For every $h \in [0,1]^K$ and $1 \le t \le T$,*

$$M_t^h(0) = c_{\exp} + (1-\mu)x_{t+1}^h(1), \tag{75}$$

and

$$M_\tau = M_T^h(0) \le M_t^h(0) \le \overline{M}^h(0), \qquad M_t^h(0) \longrightarrow \overline{M}^h(0) \quad \text{as } T-t \to \infty. \tag{76}$$

Moreover, $M_t^h(0) \le M_{t-1}^h(0)$ for $2 \le t \le T$. If $\tau \in [0,1]$, then at h^0 all marginal costs in equation 76 equal M_τ, always accepting is optimal, and $u_{h^0} = D$.

In particular, writing $\overline{\tau}(h) := 1 - \overline{M}^h(0)/c_e$, we have $\overline{\tau}(h) \le 1 - M_t^h(0)/c_e \le \tau$ for every h and t, with equality at h^0 when $\tau \in [0,1]$.

C.3 Instance-dependent regret lower bound

We show that every uniformly efficient algorithm must collect order $\log T$ expert labels at each self-confidence level where the oracle accepts, by comparing the true instance with an alternative on which deferral is strictly beneficial.

Proof. For each horizon T, write $\mathbb{P}_h := \mathbb{P}_h^\pi$ and $\mathbb{E}_h := \mathbb{E}_h^\pi$ for probability and expectation under h and π run with horizon T, and use the subscript $\star$ when $h = h^\star$. Let $u^\star := u_{h^\star}$ be the bias from Lemma 16, set $\mathfrak{d}_\star := \mathfrak{d}_{h^\star}$, so that $\overline{M}^\star(0) = c_{\exp} + (P_1 - P_0)u^\star(0)$ and $\overline{\tau}^\star = 1 - \overline{M}^\star(0)/c_e$ by Lemma 18. Finally, let $I := \{j \in [K] : \omega_j > 0,\ h_j^\star > \overline{\tau}^\star\}$.

For $j \in I$, Lemma 16(ii) implies that acceptance is the unique average-cost minimizer at every queue state. For $j \in I$ and $b < \bar{B}$,

$$\mathfrak{d}_{h^\star}(b,j,1) = \overline{M}^\star(b) - L^{h^\star}(j) \ge c_e(h_j^\star - \overline{\tau}^\star).$$

Thus equation 70 gives

$$R_T(\pi; h^\star) \ge c_e \sum_{j \in I} (h_j^\star - \overline{\tau}^\star)\mathbb{E}_\star[D_{j,T}] - C_\star, \qquad C_\star := \operatorname{sp}(D - u^\star). \tag{77}$$

By equation 77, uniform efficiency implies $\mathbb{E}_\star[D_{j,T}] \le [R_T(\pi; h^\star) + C_\star]/[c_e(h_j^\star - \overline{\tau}^\star)] = o(T^\alpha)$ for every $\alpha > 0$ and $j \in I$.

A confusing alternative for self-confidence level j. Fix $j \in I$ and $\varepsilon \in (0, \overline{\tau}^\star)$. We construct an alternative on which deferral at level j is strictly beneficial while all other conditional accuracies are unchanged: set $\widetilde{h}_j := \overline{\tau}^\star - \varepsilon$ and $\widetilde{h}_k := h_k^\star$ for $k \ne j$. We use the transformed costs k_h, gains $\bar{g}_h$, and state values v_h of Lemma 24, which isolate the effect of deferring at level j, and write $\bar{g}_\star := \bar{g}_{h^\star}$ and $v_\star := v_{h^\star}$. Let $\varphi^\star$ be a deterministic stationary minimizing policy for $h^\star$; since $j \in I$, it accepts at level j at every queue state. Define a stationary policy φ' that agrees with $\varphi^\star$ everywhere except that $\varphi'(0,j) = 1$. From any positive queue state, $\bar{B}$ consecutive rounds with $J_t = j$ and $C_t \ge 1$ empty the queue because φ' accepts at level j whenever the queue state is positive. Thus every state reaches the recurrent class containing queue zero with positive probability at least $(\omega_j\mu)^{\bar{B}}$, so this class is unique. Let λ be its stationary distribution; since J_t is drawn independently with law ω and $\omega_j > 0$, $\lambda(0,j) > 0$.

Let $\bar{g}_{\widetilde{h}}(\varphi')$ denote the stationary average transformed cost of φ' under $\widetilde{h}$. Stationarity of λ and equation 111 give

$$\bar{g}_{\widetilde{h}}(\varphi') - \bar{g}_\star = \sum_{s,a} \lambda(s)\varphi'(a \mid s)\big[k_{\widetilde{h}}(s,a) - \bar{g}_\star + \mathbb{E}[v_\star(S_{t+1}) \mid S_t = s, A_t = a] - v_\star(s)\big].$$

For every action used by $\varphi^\star$, the bracket is zero. The only additional action used by φ' is deferral at $(0,j)$, where the bracket equals $\overline{M}^\star(0) - L^{\widetilde{h}}(j) = -c_e\varepsilon$. Therefore $\bar{g}_{\widetilde{h}}(\varphi') = \bar{g}_\star - \lambda(0,j)c_e\varepsilon$. By equation 71 for the transformed problem,

$$\bar{g}_{\widetilde{h}} \le \bar{g}_\star - \lambda(0,j)c_e\varepsilon, \qquad a_\varepsilon := \bar{g}_\star - \bar{g}_{\widetilde{h}} \ge \lambda(0,j)c_e\varepsilon > 0. \tag{78}$$

Regret of under-reviewing at level j under $\widetilde{h}$. Let $r_\varepsilon := L^{\widetilde{h}}(j) - L^{h^\star}(j) = c_e(h_j^\star - \widetilde{h}_j) > 0$, $w := v_{\widetilde{h}} - v_\star$, and $s_w := \mathrm{sp}(w)$. Subtracting equation 111 under the two conditional accuracy vectors gives

$$\mathfrak{d}_{\widetilde{h}}(s,a) - \mathfrak{d}_\star(s,a) = -r_\varepsilon a \mathbf{1}\{k=j\} + a_\varepsilon + \mathbb{E}[w(S_{t+1}) \mid S_t = s, A_t = a] - w(s). \tag{79}$$

Generate $J_{T+1} \sim \omega$ independently after the last decision and set $S_{T+1} = (B_{T+1}, J_{T+1})$. By Lemma 24,

$$\mathcal{M}_t := \sum_{s=1}^{t} \left(w(S_{s+1}) - \mathbb{E}[w(S_{s+1}) \mid S_s, A_s]\right)$$

is a martingale under $\widetilde{h}$, with $|\mathcal{M}_t - \mathcal{M}_{t-1}| \le s_w$. Summing equation 79, telescoping, and using $\mathfrak{d}_\star \ge 0$ gives

$$r_\varepsilon D_{j,T} + \sum_{t=1}^{T} \mathfrak{d}_{\widetilde{h}}(S_t, A_t) \ge T a_\varepsilon - s_w - \mathcal{M}_T. \tag{80}$$

Define $E_T := \{D_{j,T} < a_\varepsilon T/(4r_\varepsilon)\}$. For T such that $s_w \le a_\varepsilon T/4$, equation 80 implies

$$E_T \subseteq \left\{ \sum_{t=1}^{T} \mathfrak{d}_{\widetilde{h}}(S_t, A_t) \ge \frac{a_\varepsilon T}{4} \right\} \cup \left\{ \mathcal{M}_T \ge \frac{a_\varepsilon T}{4} \right\}.$$

By equation 70 and uniform efficiency at $\widetilde{h}$, $\mathbb{E}_{\widetilde{h}}[\sum_{t=1}^{T} \mathfrak{d}_{\widetilde{h}}(S_t, A_t)] \le R_T(\pi; \widetilde{h}) + \mathrm{sp}(D - u_{\widetilde{h}}) = o(T^\alpha)$. Markov's inequality and the Hoeffding–Azuma inequality (Lattimore & Szepesvári, 2020, Exercise 20.6) therefore give, for every $\alpha \in (0,1)$,

$$\mathbb{P}_{\widetilde{h}}(E_T) \le \frac{4\mathbb{E}_{\widetilde{h}}[\sum_{t=1}^{T} \mathfrak{d}_{\widetilde{h}}(S_t, A_t)]}{a_\varepsilon T} + \exp\left\{ -\frac{a_\varepsilon^2 T}{32(1+s_w)^2} \right\} = o(T^{\alpha-1}). \tag{81}$$

Under the true instance,

$$\mathbb{P}_\star(E_T^c) \le \frac{4 r_\varepsilon \mathbb{E}_\star[D_{j,T}]}{a_\varepsilon T} \longrightarrow 0. \tag{82}$$

Information from completed reviews. Lemma 17 gives

$$\mathrm{KL}(P_{h^\star}^{\pi,T} \| P_{\widetilde{h}}^{\pi,T}) \le \mathbb{E}_\star[D_{j,T}]\, \mathrm{kl}(h_j^\star, \widetilde{h}_j).$$

Since $E_T \in \mathcal{F}_T$, the data-processing inequality (Lattimore & Szepesvári, 2020, Exercise 14.10) applies. Moreover, $\mathrm{kl}(p,q) \ge p \log(1/q) - \log 2$ because the binary entropy is at most $\log 2$. Hence

$$\mathbb{E}_\star[D_{j,T}]\, \mathrm{kl}(h_j^\star, \widetilde{h}_j) \ge \mathbb{P}_\star(E_T) \log \frac{1}{\mathbb{P}_{\widetilde{h}}(E_T)} - \log 2.$$

By equation 82, $\mathbb{P}_\star(E_T) \to 1$, while equation 81 gives $\log(1/\mathbb{P}_{\widetilde{h}}(E_T)) \ge (1 - \alpha - o(1)) \log T$. Therefore, for every $\alpha \in (0,1)$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\star[D_{j,T}]}{\log T} \ge \frac{1-\alpha}{\mathrm{kl}(h_j^\star, \overline{\tau}^\star - \varepsilon)}.$$

Letting $\alpha \downarrow 0$ and then $\varepsilon \downarrow 0$, using continuity of Bernoulli relative entropy in its second argument on $(0,1)$, yields

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\star[D_{j,T}]}{\log T} \ge \frac{1}{\mathrm{kl}(h_j^\star, \overline{\tau}^\star)}.$$

Substituting this into equation 77 and summing over $j \in I$ proves equation 2. □

Remark 2. When $\mu = 1$, $\overline{\tau}^\star = 1 - c_{\mathrm{exp}}/c_e$, and Theorem 3 can alternatively be derived from the Graves–Lai lower bound for structured MDPs (Ok et al., 2018, Theorem 1).

C.4 KL−PBD IS ASYMPTOTICALLY OPTIMAL UNDER IMMEDIATE-SERVICE

In this section, we show that KL−PBD matches Theorem 3 when service is immediate ($\mu = 1$).

Throughout this subsection, $C_t \geq 1$ almost surely and $0 < c_{\exp} < c_e$. Thus $\mu = 1$ and $\tau = 1 - c_{\exp}/c_e \in (0, 1)$. We consider KL−PBD with a fixed budget $\phi(T) > 0$. All probabilities and expectations in this subsection are under $h^\star$ and $\pi^{\text{KL−PBD}}$.

By equation 75 with $\mu = 1$, $M_t^h(0) = c_{\exp}$ for every h and t. Thus, after n completed level j reviews, the algorithm defers exactly when $g_{j,n} < \tau$, with ties resolved in favor of acceptance.

Every deferral yields a label before the next round. The sample counts before successive level j deferrals are therefore $0, 1, \ldots, D_{j,T} - 1$. We use the infinite i.i.d. ground-truth label sequences for each self-confidence level from Lemma 8.

Lemma 19 (Erroneous acceptance under immediate service)**.** *For a level j with $0 < h_j^\star < \tau$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{J_t = j, A_t = 0\}\right] \leq \frac{Te^{-\phi(T)}}{\exp\{\mathrm{kl}(\tau, h_j^\star)\} - 1}. \tag{83}$$

If $h_j^\star = 0$, the left-hand side is zero.

Proof. Fix j with $0 < h_j^\star < \tau$. The algorithm accepts only when $g_{j,n} \geq \tau$ and hence $n \geq 1$, since $g_{j,0} = 0$. By the definition of the lower confidence bound and $\phi(T) > 0$, $g_{j,n} \geq \tau$ implies $\widehat{h}_{j,n} > \tau$ and $n\,\mathrm{kl}(\widehat{h}_{j,n}, \tau) \geq \phi(T)$. For every $z \in [\tau, 1]$,

$$\begin{aligned}\mathrm{kl}(z, h_j^\star) - \mathrm{kl}(z, \tau) &= \mathrm{kl}(\tau, h_j^\star) + (z - \tau)\log\left(\frac{\tau(1 - h_j^\star)}{h_j^\star(1 - \tau)}\right) \\ &\geq \mathrm{kl}(\tau, h_j^\star).\end{aligned}$$

If $\phi(T)/n \leq \mathrm{kl}(1, \tau)$, let $z_n \in (\tau, 1]$ solve $\mathrm{kl}(z_n, \tau) = \phi(T)/n$. The Bernoulli Chernoff bound gives

$$\mathbb{P}(g_{j,n} \geq \tau) \leq \mathbb{P}(\widehat{h}_{j,n} \geq z_n) \leq e^{-n\,\mathrm{kl}(z_n, h_j^\star)} \leq e^{-\phi(T) - n\,\mathrm{kl}(\tau, h_j^\star)}.$$

If $\phi(T)/n > \mathrm{kl}(1, \tau)$, the event is empty and the same bound holds. Taking a union bound over sample counts,

$$\mathbb{P}(\exists n \geq 1 : g_{j,n} \geq \tau) \leq \sum_{n=1}^{\infty} e^{-\phi(T) - n\,\mathrm{kl}(\tau, h_j^\star)} = \frac{e^{-\phi(T)}}{\exp\{\mathrm{kl}(\tau, h_j^\star)\} - 1}.$$

Since

$$\sum_{t=1}^{T} \mathbf{1}\{J_t = j, A_t = 0\} \leq T\mathbf{1}\{\exists n \geq 1 : g_{j,n} \geq \tau\},$$

taking expectations proves the bound. If $h_j^\star = 0$, all sample means and lower confidence bounds are zero, so every level j query is deferred. □

Proposition 3 (Asymptotic optimality under immediate service)**.** *For fixed K, any budget sequence satisfying $\phi(T)/\log T \to 1$ makes KL−PBD uniformly efficient on the immediate-service class. If, in addition, $Te^{-\phi(T)}/\log T \to 0$, then, for every fixed immediate-service instance,*

$$\lim_{T\to\infty} \frac{R_T(\pi^{\text{KL−PBD}})}{\log T} = c_e \sum_{\substack{j \in [K] \\ h_j^\star > \tau,\ \omega_j > 0}} \frac{h_j^\star - \tau}{\mathrm{kl}(h_j^\star, \tau)}.$$

In particular, the equality holds for $\phi(T) = \log T + \log\log T$ and $\phi(T) = \log T$ for $T \geq 3$, with arbitrary positive budgets for smaller horizons. If $h_j^\star \geq \tau$ for every $j \in [K]$ with $\omega_j > 0$, the condition $Te^{-\phi(T)}/\log T \to 0$ is unnecessary.

Proof. Fix a level j with $h_j^\star > \tau$ and $\varepsilon \in (0, h_j^\star - \tau)$, and write $d := \mathrm{kl}(h_j^\star - \varepsilon, \tau) > 0$. If $n \geq \phi(T)/d$ and $\widehat{h}_{j,n} \geq h_j^\star - \varepsilon$, then

$$n\,\mathrm{kl}(\widehat{h}_{j,n}, \tau) \geq n\,\mathrm{kl}(h_j^\star - \varepsilon, \tau) \geq \phi(T).$$

Since $\widehat{h}_{j,n} > \tau$, this implies $g_{j,n} \geq \tau$, so the algorithm accepts. Hence each deferral at a sample count $n \geq \phi(T)/d$ requires $\widehat{h}_{j,n} < h_j^\star - \varepsilon$. Since each sample count occurs at most once before a deferral,

$$D_{j,T} \leq \left\lceil \frac{\phi(T)}{d} \right\rceil + \sum_{n=1}^{\infty} \mathbf{1}\{\widehat{h}_{j,n} < h_j^\star - \varepsilon\}.$$

Taking expectations and applying the Bernoulli Chernoff bound gives

$$\begin{aligned}\mathbb{E}[D_{j,T}] &\leq \left\lceil \frac{\phi(T)}{d} \right\rceil + \sum_{n=1}^{\infty} e^{-n\,\mathrm{kl}(h_j^\star - \varepsilon, h_j^\star)} \\ &\leq 1 + \frac{\phi(T)}{\mathrm{kl}(h_j^\star - \varepsilon, \tau)} + \frac{1}{\exp\{\mathrm{kl}(h_j^\star - \varepsilon, h_j^\star)\} - 1} \end{aligned} \tag{84}$$

where the final term is interpreted as zero when the divergence is infinite. In particular, this covers $h_j^\star = 1$.

We next prove uniform efficiency. By Lemma 2, under immediate service,

$$\begin{aligned}R_T(\pi^{\texttt{KL-PBD}}) = c_e &\sum_{\substack{j \in [K] \\ h_j^\star > \tau,\ \omega_j > 0}} (h_j^\star - \tau)\mathbb{E}[D_{j,T}] \\ &+ c_e \sum_{\substack{j \in [K] \\ h_j^\star < \tau,\ \omega_j > 0}} (\tau - h_j^\star)\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{J_t = j, A_t = 0\}\right].\end{aligned}$$

Levels j with $h_j^\star = \tau$ or $\omega_j = 0$ contribute zero regret. If $h_j^\star = 0$, KL-PBD never accepts at level j. By Lemma 19, the second term is at most

$$c_e T e^{-\phi(T)} \sum_{\substack{j \in [K] \\ 0 < h_j^\star < \tau,\ \omega_j > 0}} \frac{\tau - h_j^\star}{\exp\{\mathrm{kl}(\tau, h_j^\star)\} - 1}. \tag{85}$$

If $\phi(T)/\log T \to 1$, then $Te^{-\phi(T)} = \exp\{o(\log T)\} = o(T^\alpha)$ for every $\alpha > 0$, while equation 84 gives $\mathbb{E}[D_{j,T}] = O(\log T)$ for every j with $h_j^\star > \tau$. Thus KL-PBD is uniformly efficient on the immediate-service class.

It remains to identify the asymptotic coefficient. Dividing equation 84 by $\log T$, using $\phi(T)/\log T \to 1$, and then letting $\varepsilon \downarrow 0$ gives, for every j with $h_j^\star > \tau$,

$$\limsup_{T \to \infty} \frac{\mathbb{E}[D_{j,T}]}{\log T} \leq \frac{1}{\mathrm{kl}(h_j^\star, \tau)}.$$

If, in addition, $Te^{-\phi(T)}/\log T \to 0$, then equation 85 is $o(\log T)$. Since K is finite,

$$\limsup_{T \to \infty} \frac{R_T(\pi^{\texttt{KL-PBD}})}{\log T} \leq c_e \sum_{\substack{j \in [K] \\ h_j^\star > \tau,\ \omega_j > 0}} \frac{h_j^\star - \tau}{\mathrm{kl}(h_j^\star, \tau)}.$$

Uniform efficiency and Theorem 3 give the matching lower bound, proving the claimed limit.

If $h_j^\star \geq \tau$ for every $j \in [K]$ with $\omega_j > 0$, the acceptance contribution is identically zero. Hence the same upper coefficient follows from equation 84 under $\phi(T)/\log T \to 1$ alone, and the lower bound again gives equality.

Finally, both $\phi(T) = \log T + \log\log T$ and $\phi(T) = \log T$ satisfy $\phi(T)/\log T \to 1$ and $Te^{-\phi(T)}/\log T \to 0$. □

C.5 WORST-CASE REGRET LOWER BOUND

The main idea of the proof is to construct statistically indistinguishable instances with empty-queue states at which the optimal actions differ, while controlling how often congestion prevents the process from visiting these informative states.

Proof. Fix an algorithm π and write $m := \min\{\tau, 1-\tau\}$. If $\sum_{j=1}^{K} \omega_j = 0$, both sides of equation 3 are zero, so assume otherwise.

We first prove a sharper lower bound for $\mu = 1$. In this setting, the queue stays empty and every deferred label is observed before the next decision. Define the hard family

$$\delta_j := \frac{m}{4} \min\left\{1, \frac{1}{\sqrt{\omega_j T}}\right\}, \qquad h_j^{\sigma} := \tau + \sigma_j \delta_j, \tag{86}$$

with $\delta_j = 0$ when $\omega_j = 0$. Since $B_t = 0$ for every t, the next-state law is the same under both actions, so $M_t^h(0) = c_{\exp} = c_e(1-\tau)$ for every h and $t \le T$, and $c_{\mathrm{hold}} B_t = D(B_{T+1}) = 0$ along every trajectory. Resolving ties in favor of acceptance, the oracle accepts at level j when $\sigma_j = +1$ and defers when $\sigma_j = -1$, and Lemma 2 gives

$$R_T(\pi; h^{\sigma}) = \sum_{j=1}^{K} R_{T,j}(\pi; h^{\sigma}),$$

$$R_{T,j}(\pi; h^{\sigma}) := c_e \delta_j \sum_{t=1}^{T} \mathbb{P}_{h^{\sigma}}^{\pi}\big(J_t = j,\ A_t \neq \mathbf{1}\{\sigma_j = -1\}\big).$$

Fix j with $\omega_j > 0$ and neighboring sign vectors $\sigma, \sigma^{(j)}$ with $\sigma_j = +1$ and $\sigma_j^{(j)} = -1$. Since the two conditional accuracy vectors differ only in coordinate j, Lemma 17 gives, for every $t \le T$,

$$\mathrm{KL}\Big(P_{h^{\sigma}}^{\pi,t-1} \Big\| P_{h^{\sigma^{(j)}}}^{\pi,t-1}\Big) \le \omega_j (t-1)\, \mathrm{kl}(\tau+\delta_j, \tau-\delta_j) \le \frac{32\delta_j^2}{9m^2}\, \omega_j T \le \frac{2}{9}.$$

For the middle inequality, every $u \in [\tau-\delta_j, \tau+\delta_j]$ satisfies $\min\{u, 1-u\} \ge 3m/4$, so Taylor's theorem in the first argument of the Bernoulli relative entropy gives

$$\mathrm{kl}(\tau+\delta_j, \tau-\delta_j) \le \frac{(2\delta_j)^2}{2} \sup_{u \in [\tau-\delta_j, \tau+\delta_j]} \frac{1}{u(1-u)} \le \frac{32\delta_j^2}{9m^2},$$

and $\delta_j^2 \omega_j T \le m^2/16$ by equation 86. Pinsker's inequality (Lattimore & Szepesvári, 2020, Eq. (14.12)) therefore gives $\mathrm{TV}(P_{h^{\sigma}}^{\pi,t-1}, P_{h^{\sigma^{(j)}}}^{\pi,t-1}) \le \sqrt{(2/9)/2} = 1/3$.

Let $\alpha_t := \pi_t(A_t = 1 \mid \mathcal{F}_{t-1}, J_t = j)$ be the algorithm's conditional probability of deferral after the arrival of a query at level j. Since J_t is independent of $\mathcal{F}_{t-1}$, for $\sigma_j = +1$,

$$\begin{aligned}
\mathbb{P}_{h^{\sigma}}^{\pi}(J_t = j, A_t = 1) + \mathbb{P}_{h^{\sigma^{(j)}}}^{\pi}(J_t = j, A_t = 0) &= \omega_j \big[1 + \mathbb{E}_{h^{\sigma}}^{\pi}[\alpha_t] - \mathbb{E}_{h^{\sigma^{(j)}}}^{\pi}[\alpha_t]\big] \\
&\ge \omega_j \big(1 - \mathrm{TV}(P_{h^{\sigma}}^{\pi,t-1}, P_{h^{\sigma^{(j)}}}^{\pi,t-1})\big) \\
&\ge \frac{2\omega_j}{3},
\end{aligned}$$

using $0 \le \alpha_t \le 1$ and that α_t is $\mathcal{F}_{t-1}$-measurable. Summing over t gives $R_{T,j}(\pi; h^{\sigma}) + R_{T,j}(\pi; h^{\sigma^{(j)}}) \ge 2c_e\delta_j\omega_j T/3$. Pairing the 2^{K-1} sign vectors with identical σ_{-j} and summing over j gives

$$2^{-K} \sum_{\sigma} R_T(\pi; h^{\sigma}) \ge \frac{c_e}{3} \sum_{j=1}^{K} \delta_j \omega_j T = \frac{c_e m}{12} \sum_{j=1}^{K} \min\{\omega_j T, \sqrt{\omega_j T}\}.$$

The supremum over h is at least this average, proving the $\mu = 1$ sharpening stated after Theorem 4.

Assume from now on that $\mu < 1$. If $c_{\mathrm{hold}} = 0$, the right-hand side of equation 3 is zero, so there is nothing to prove. We therefore take $c_{\mathrm{hold}} > 0$.

Self-centered heterogeneous family. We next recenter each sign pattern so that its empty-queue action gaps remain exactly controlled despite the queue-dependent threshold. For $\omega_j > 0$, set

$$\delta_j := \frac{1}{24} \min\left\{ \mu^2 m \min\left\{ 1, \frac{1}{\sqrt{\omega_j T}} \right\}, \frac{\mu c_{\mathrm{hold}}}{c_e} \right\}, \tag{87}$$

and set $\delta_j = 0$ if $\omega_j = 0$. Write $\delta_{\max} := \max_j \delta_j$.

For each $\sigma \in \{-1, +1\}^K$, Lemma 21 provides a unique center a_σ, with $S_\sigma^- := \sum_{j:\sigma_j=-1} \omega_j \delta_j$. Set

$$h_j^\sigma := \tau - a_\sigma + \sigma_j \delta_j. \tag{88}$$

Write $u_\sigma := u_{h^\sigma}$, $\mathbb{P}_\sigma^s := \mathbb{P}_{h^\sigma}^{\pi,s}$, $\mathbb{P}_\sigma := \mathbb{P}_{h^\sigma}^\pi$ and $\mathbb{E}_\sigma := \mathbb{E}_{h^\sigma}^\pi$. By Lemma 21,

$$\overline{\tau}(h^\sigma) = \tau - a_\sigma, \tag{89}$$

so $\overline{M}^{h^\sigma}(0) - L^{h^\sigma}(j) = \sigma_j c_e \delta_j$. Moreover,

$$0 \le a_\sigma \le \frac{1-\mu}{\mu} S_\sigma^-. \tag{90}$$

Testing neighboring instances. We now show that flipping one sign changes the observations too little to identify the optimal empty-queue action reliably. By equation 90 and equation 87, every hard parameter lies within $m/12$ of τ. The inequality $\log x \le x - 1$ gives $\mathrm{kl}(p, q) \le (p-q)^2/[q(1-q)]$; since $q(1-q) \ge (3m/4)^2$ on the hard family, $\mathrm{kl}(p, q) \le 16(p-q)^2/(9m^2)$. Using the neighboring-center bound equation 110 from Lemma 21, for neighboring sign vectors with $\sigma_j = +1$ and $\sigma_j^{(j)} = -1$, write $\Delta a := a_{\sigma^{(j)}} - a_\sigma$ and $\Omega := \sum_{k=1}^K \omega_k$. Then

$$\begin{aligned} \sum_{k=1}^K \omega_k |h_k^\sigma - h_k^{\sigma^{(j)}}|^2 &= 4\omega_j \delta_j^2 + 4\omega_j \delta_j \Delta a + \Omega(\Delta a)^2 \\ &\le \frac{4}{\mu^2} \omega_j \delta_j^2. \end{aligned}$$

Lemma 17 and equation 87 therefore give $\mathrm{KL}(P_\sigma^{t-1} \| P_{\sigma^{(j)}}^{t-1}) \le \mu^2/81$ uniformly over $t \le T$. Pinsker's inequality (Lattimore & Szepesvári, 2020, Eq. (14.12)) then yields

$$\mathrm{TV}(P_\sigma^{t-1}, P_{\sigma^{(j)}}^{t-1}) \le \frac{\mu}{12}, \qquad t \le T. \tag{91}$$

Empty-queue occupancy and average-cost disadvantage. We combine equation 92 and equation 91: either the queue is often empty, forcing decision errors, or keeping it occupied requires costly deferrals. Let $N_0(T) := \sum_{t=1}^T \mathbf{1}\{B_t = 0\}$. For each hard instance and level j, define

$$\mathcal{D}_{\sigma,j} := \mathbb{E}_\sigma \left[\sum_{t=1}^T \mathfrak{d}_{h^\sigma}(B_t, J_t, A_t) \mathbf{1}\{J_t = j\} \right], \qquad \mathcal{D}_\sigma := \sum_{j=1}^K \mathcal{D}_{\sigma,j}.$$

Let r be the conditional service median in Lemma 20. By equation 102, equation 89, and equation 87, for $j \in [K]$ and $r \le b < \bar{B}$,

$$\mathfrak{d}_{h^\sigma}(b, j, 1) \ge \frac{c_{\mathrm{hold}}}{2} - c_e \delta_{\max} \ge \frac{11}{24} c_{\mathrm{hold}} =: g_{\mathrm{bar}}. \tag{92}$$

Let $H_T := \sum_{t=1}^T \mathbf{1}\{B_t \ge r, A_t = 1\}$. Lemma 22 and equation 92 give

$$\mathbb{E}_\sigma[N_0(T)] \ge \frac{\mu T}{2} - 2\mathbb{E}_\sigma[H_T] \ge \frac{\mu T}{2} - \frac{48}{11 c_{\mathrm{hold}}} \mathcal{D}_\sigma. \tag{93}$$

Pair σ and $\sigma^{(j)}$. For the positive sign, every level j deferral has disadvantage at least $c_e \delta_j$; for the negative sign, accepting a level-j query at an empty queue has disadvantage exactly $c_e \delta_j$. With

$\alpha_t := \pi_t(A_t = 1 \mid \mathcal{F}_{t-1}, J_t = j)$ as in the case $\mu = 1$, and since $\mathbf{1}\{B_t = 0\}\alpha_t$ and $\mathbf{1}\{B_t = 0\}(1-\alpha_t)$ are $\mathcal{F}_{t-1}$-measurable and $[0,1]$-valued, the variational characterization of total variation gives

$$\mathbb{E}_\sigma[\mathbf{1}\{B_t = 0\}\alpha_t] + \mathbb{E}_{\sigma^{(j)}}[\mathbf{1}\{B_t = 0\}(1-\alpha_t)] \geq \frac{\mathbb{P}_\sigma(B_t = 0) + \mathbb{P}_{\sigma^{(j)}}(B_t = 0)}{2} - \mathrm{TV}(P_\sigma^{t-1}, P_{\sigma^{(j)}}^{t-1}).$$

Summing this inequality and using equation 91 yields

$$\frac{1}{2}\left(\mathcal{D}_{\sigma,j} + \mathcal{D}_{\sigma^{(j)},j}\right) \geq \frac{c_e \delta_j \omega_j}{2}\left[\frac{\mathbb{E}_\sigma[N_0(T)] + \mathbb{E}_{\sigma^{(j)}}[N_0(T)]}{2} - \frac{\mu T}{12}\right].$$

Averaging over all neighboring pairs and summing over j, with $S_\delta := \sum_{j=1}^K \omega_j \delta_j$ and $\overline{\mathcal{D}} := 2^{-K}\sum_\sigma \mathcal{D}_\sigma$, gives

$$\overline{\mathcal{D}} \geq \frac{5\mu T}{24} c_e S_\delta - \frac{24 c_e S_\delta}{11 c_{\mathrm{hold}}}\overline{\mathcal{D}}.$$

Since $c_e S_\delta \leq \mu c_{\mathrm{hold}}/24$, we conclude that

$$\overline{\mathcal{D}} \geq \frac{55}{288}\,\mu T c_e S_\delta. \tag{94}$$

From disadvantage to finite-horizon regret. It remains to control the terminal bias correction without introducing any dependence on the capacity $\bar{B}$. For a general conditional accuracy vector h, write $\mathcal{D}_h(\pi) := \mathbb{E}_h^\pi[\sum_{t=1}^T \mathfrak{d}_h(B_t, J_t, A_t)]$. We only need a one-sided terminal comparison. By Lemma 16, $u_h \geq D$. Because $B_1 = 0$, equation 100 shows that a stationary average-cost minimizer has finite-horizon cost at most $T g_h$, while an arbitrary algorithm has cost $T g_h + \mathcal{D}_h(\pi) + \mathbb{E}_h^\pi[(D - u_h)(B_{T+1})]$. Consequently,

$$R_T(\pi; h) \geq \mathcal{D}_h(\pi) - \mathbb{E}_h^\pi[(u_h - D)(B_{T+1})]. \tag{95}$$

For the hard instance h^σ, Lemma 23 gives

$$0 \leq u_\sigma(b) - D(b) \leq \frac{2 c_e S_\sigma^-}{\mu}\left\lceil \frac{b}{r} \right\rceil.$$

Pathwise, $\lceil B_{T+1}/r \rceil \leq 1 + H_T$, because a deferral below r cannot raise the queue above r, while each deferral at or above r increases this coarse queue state by at most one. Together with equation 92, equation 95 yields

$$R_T(\pi; h^\sigma) \geq \frac{9}{11}\mathcal{D}_\sigma - \frac{2 c_e S_\delta}{\mu}.$$

Averaging over the hard family and using equation 94,

$$\sup_h R_T(\pi; h) \geq c_e S_\delta \left(\frac{5\mu T}{32} - \frac{2}{\mu}\right). \tag{96}$$

Independently, the first decision gives a complementary bound valid for every T. Consider the family $h_j^\sigma = \tau + \sigma_j m/4$. By equation 75, Lemma 18 at h^0, and the adjacent-increment estimate equation 37 from the proof of Proposition 2, $|M_1^{h^\sigma}(0) - M_\tau| \leq (1-\mu)c_e m/4$. Since $|L^{h^\sigma}(j) - M_\tau| = c_e m/4$, opposite signs in coordinate j induce opposite optimal actions at $B_1 = 0$, each with gap at least $\mu c_e m/4$. Before the first action the algorithm has no accuracy-dependent observations, so averaging each sign pair forces an error probability at least $1/2$ conditional on the arrival of a query at level j. Summing over j gives

$$\sup_h R_T(\pi; h) \geq \frac{\mu c_e m}{8}\sum_{j=1}^K \omega_j. \tag{97}$$

Finally, define

$$\mathcal{L}_T := \sum_{j=1}^K \min\left\{\mu^3 c_e m \min\{\omega_j T, \sqrt{\omega_j T}\}, \mu^2 c_{\mathrm{hold}} \omega_j T\right\}.$$

By equation 87, $\mathcal{L}_T = 24\mu T c_e S_\delta$. Therefore equation 96 gives

$$\sup_h R_T(\pi; h) \geq \mathcal{L}_T \left(\frac{5}{768} - \frac{1}{12\mu^2 T} \right), \tag{98}$$

while equation 97 and $\mathcal{L}_T \leq \mu^3 c_e m T \sum_j \omega_j$ give

$$\sup_h R_T(\pi; h) \geq \frac{\mathcal{L}_T}{8\mu^2 T}. \tag{99}$$

If $\mu^2 T \geq 32$, equation 98 is at least $\mathcal{L}_T/256$; if $\mu^2 T \leq 32$, equation 99 is at least $\mathcal{L}_T/256$. Since π was arbitrary, taking the infimum over algorithms proves equation 3. □

Remark 3 (Worst case over ω). The function $f(x) = \min\{x, \sqrt{x}\}$ is nondecreasing and concave on $[0, \infty)$, so Jensen's inequality gives $\sum_{j=1}^K f(\omega_j T) \leq K f(T/K) = \min\{T, \sqrt{KT}\}$, with equality when $\omega_0 = 0$ and $\omega_j = 1/K$ for every $j \in [K]$.

C.6 Technical lemmas

C.6.1 Proofs of the common ingredients

Proof of Lemma 16. By Lemma 3, each ξ_s^h is nonnegative and nondecreasing in q. Its update is given by equation 7, equation 9, and equation 10.

Write $g_D(q) := D(q) - D(q-1)$ and $d_0(b) := \sum_{c=0}^b \rho_C(c) g_D(b+1-c)$. Differencing equation 63 gives $g_D(q) = c_{\text{hold}} + d_0(q-1)$. Since D is discrete convex, d_0 is nondecreasing. The median or maximum in the increment update is therefore at least $d_0(q-1)$, giving $\xi_1^h(q) \geq g_D(q) = \xi_0^h(q)$. The update is coordinatewise nondecreasing, so $\xi_{s+1}^h \geq \xi_s^h$ for every $s \geq 0$.

By equation 39, $\xi_s^h(q) \leq q c_{\text{hold}}/\mu + (c_e - c_{\text{exp}})_+$, uniformly in s. Hence $\xi_s^h(q) \uparrow x_h(q)$, and $W_s^h(b) - W_s^h(0) \to u_h(b) := \sum_{q=1}^b x_h(q)$. In particular, $u_h \geq D$ because $\xi_0^h = g_D$.

Let $\mathcal{T}_h$ be the queue Bellman operator. With $v_s^h := W_s^h - W_s^h(0)$ and $g_s^h := W_{s+1}^h(0) - W_s^h(0)$, translation invariance gives $g_s^h + v_{s+1}^h = \mathcal{T}_h v_s^h$ and $g_s^h = (\mathcal{T}_h v_s^h)(0)$. The operator is continuous because the state and action spaces are finite. Thus $g_s^h \to g_h := (\mathcal{T}_h u_h)(0)$, and passing to the limit yields equation 64. Passing the increment identities to the limit proves (i) and preserves monotonicity and discrete convexity.

For (ii), equation 66 shows that $d_h(b)$ is nonnegative and nondecreasing in b. At the empty queue, $d_h(0) = \rho_C(0) x_h(1) = (1-\mu) u_h(1)$, proving equation 69.

For (iii), using equation 64, independence of J_t from $\mathcal{F}_{t-1}$, and the tower property,

$$\mathbb{E}_h^\pi \left[\sum_{t=1}^T \ell_h(B_t, J_t, A_t) + D(B_{T+1}) \right] = T g_h + \mathbb{E}[u_h(B_1)] + \mathbb{E}_h^\pi[(D - u_h)(B_{T+1})] + \mathbb{E}_h^\pi \left[\sum_{t=1}^T \mathfrak{d}_h(B_t, J_t, A_t) \right]. \tag{100}$$

A stationary policy attaining the minima in equation 64 has zero disadvantage. Therefore every algorithm has cost at least $T g_h + \mathbb{E}[u_h(B_1)] + \min_b (D - u_h)(b)$, while that stationary policy has cost at most the same expression with the minimum replaced by the maximum. The finite-horizon optimum lies between these two quantities, which proves equation 70.

For (iv), evaluating the minimum in equation 64 at the action distribution of φ gives $g_h + u_h(b) \leq \sum_{j=0}^K \omega_j \sum_a \varphi(a \mid b, j) \{\ell_h(b, j, a) + (P_a u_h)(b)\}$ for every b. Averaging under λ, stationarity cancels the two u_h terms and leaves equation 71. □

Proof of Lemma 17. For $s \geq 1$, set $\Delta N_{j,s} := N_{j,s+1} - N_{j,s}$, order the labels completed in round s by submission time, and expose them one at a time. Conditional on $\mathcal{F}_{s-1}$, J_s, A_s, C_s, and the

labels already exposed in round s, each newly exposed level j label has law $\mathrm{Bernoulli}(h_j)$ under h and $\mathrm{Bernoulli}(\widetilde{h}_j)$ under $\widetilde{h}$; the remaining observations have the same conditional law under the two instances. The KL chain rule (Lattimore & Szepesvári, 2020, Exercise 14.12, Eq. (14.17)) therefore gives

$$\begin{aligned}\mathrm{KL}(P_h^{\pi,t}\|P_{\widetilde{h}}^{\pi,t}) &= \sum_{s=1}^{t}\mathbb{E}_h^{\pi}\left[\sum_{j=1}^{K}\Delta N_{j,s}\,\mathrm{kl}(h_j,\widetilde{h}_j)\right]\\ &= \sum_{j=1}^{K}\mathbb{E}_h^{\pi}[N_{j,t+1}]\,\mathrm{kl}(h_j,\widetilde{h}_j).\end{aligned}$$

Since $N_{j,t+1}\le D_{j,t}\le\sum_{s=1}^{t}\mathbf{1}\{J_s=j\}$ pathwise, $\mathbb{E}_h^{\pi}[N_{j,t+1}]\le\mathbb{E}_h^{\pi}[D_{j,t}]\le t\omega_j$, proving equation 72 and equation 73. □

Proof of Lemma 18. Equation 75 follows from equation 7 at $b=0$. Since $x_{t+1}^h(1)=\xi_{T-t}^h(1)$, the monotone convergence proved in Lemma 16 gives both the time monotonicity and the limit, using equation 69. Also, equation 63 gives $D(1)=c_{\mathrm{hold}}+(1-\mu)D(1)$, so $\xi_0^h(1)=c_{\mathrm{hold}}/\mu$ and $M_T^h(0)=M_\tau$.

Suppose now that $\tau\in[0,1]$. At h^0, $L^{h^0}(j)=M_\tau$ for every $j\in[K]$. By Proposition 1 and equation 76, $M_t^{h^0}(b)\ge M_\tau$ whenever deferral is feasible. Hence always accepting is optimal for every horizon. Under this policy, the difference in expected cost between starting from queue state b and from zero is exactly $D(b)$. Thus $W_s^{h^0}(b)-W_s^{h^0}(0)=D(b)$ for every s. Taking the limit gives $u_{h^0}=D$, and the finite-horizon empty-queue marginals all equal M_τ. □

C.6.2 Lemmas for the worst-case lower bound

For the following four lemmas, assume $0<\mu<1$, $c_{\mathrm{hold}}>0$, and $\tau\in(0,1)$, and write $m:=\min\{\tau,1-\tau\}$. Choose an integer $r\ge1$ that is a median of C_t conditional on $C_t>0$, which exists since $\mu>0$, so that

$$p_-:=\mathbb{P}(1\le C_t\le r)\ge\frac{\mu}{2},\qquad p_+:=\mathbb{P}(C_t\ge r)\ge\frac{\mu}{2}. \tag{101}$$

Let

$$\mathcal{N}:=\left\{h:\|h-h^0\|_\infty<\min\left\{\frac{m}{2},\frac{c_{\mathrm{hold}}}{4c_e}\right\}\right\}.$$

Starting from $B_1=b$ under a stationary policy φ, write $\mathbb{E}_b^{\varphi}$ for expectation and let $\tau_0:=\inf\{s\ge0:B_{s+1}=0\}$ be the number of rounds until the queue first becomes empty.

Lemma 20 (Local queue estimates)**.** *Fix $h\in\mathcal{N}$. For $r\le b<\bar{B}$,*

$$\overline{M}^h(b)-\overline{M}^h(0)\ge\frac{c_{\mathrm{hold}}}{2}, \tag{102}$$

every stationary minimizing policy accepts at b, and $\mathfrak{d}_h(b,j,1)\ge c_{\mathrm{hold}}/4$ for $j\in[K]$. Under every stationary minimizing policy, and under the policy that always accepts,

$$\mathbb{E}_b[\tau_0]\le\frac{\lceil b/r\rceil}{p_+}\le\frac{2\lceil b/r\rceil}{\mu},\qquad b\ge1. \tag{103}$$

For every stationary policy φ with $\mathbb{E}_b^{\varphi}[\tau_0]<\infty$ for every b,

$$u_h(b)\le\mathbb{E}_b^{\varphi}\left[\sum_{t=1}^{\tau_0}\big(\ell_h(B_t,J_t,A_t)-g_h\big)\right], \tag{104}$$

with equality when φ selects an ACOE-minimizing action at every state.

Proof. For the solution u_h from Lemma 16, equation 66, monotonicity of $(x_h(q))_q$, and equation 76 give, for every $b\ge r$ with $b<\bar{B}$,

$$\overline{M}^h(b)-\overline{M}^h(0)\ge\sum_{c=1}^{b}\rho_C(c)x_h(1)\ge\frac{c_{\mathrm{hold}}}{\mu}\mathbb{P}(1\le C_t\le b)\ge\frac{c_{\mathrm{hold}}}{2},$$

which is equation 102. Because $\overline{M}^h(0) \geq M_\tau$ and $\|h - h^0\|_\infty < c_{\mathrm{hold}}/(4c_e)$, we have $L^h(j) \leq M_\tau + c_{\mathrm{hold}}/4$. Thus every stationary minimizing policy for $h \in \mathcal{N}$ accepts every query at $b \geq r$. For $j \in [K]$ and $r \leq b < \bar{B}$, we also have $\overline{M}^h(b) - L^h(j) \geq c_{\mathrm{hold}}/4$.

From queue state one the queue then stays at most r until it hits zero, so under any stationary minimizing policy

$$\mathbb{E}_1[\tau_0] \leq \frac{1}{p_+} \leq \frac{2}{\mu}. \tag{105}$$

More generally, positive-service events of size at least r reduce any queue state by at least r until it enters $\{0, \dots, r\}$. This proves equation 103 under every stationary minimizing policy for $h \in \mathcal{N}$. The same drainage bounds hold for the policy that always accepts.

Let φ be a stationary policy with $\mathbb{E}_b^\varphi[\tau_0] < \infty$ for every b. Evaluating the minimum in equation 64 at φ, summing over rounds $1, \dots, \tau_0 \wedge s$, and letting $s \to \infty$ by dominated convergence gives equation 104, with equality when φ selects an ACOE-minimizing action at every state. □

Lemma 21 (Self-centered family)**.** *Fix $T \geq 1$, let δ_j be given by equation 87, with $\delta_j = 0$ when $\omega_j = 0$, and write $\delta_{\max} := \max_j \delta_j$. Let r and $\mathcal{N}$ be as in Lemma 20.*

For $\sigma \in \{-1, +1\}^K$, let $S_\sigma^- := \sum_{j:\sigma_j=-1} \omega_j \delta_j$ and $A_\sigma := (1-\mu)S_\sigma^-/\mu$. There is a unique $a_\sigma \in [0, A_\sigma]$ for which the vector $h_j^\sigma := \tau - a_\sigma + \sigma_j \delta_j$ satisfies $\overline{\tau}(h^\sigma) = \tau - a_\sigma$. It belongs to $\mathcal{N}$. If $\sigma_j = +1$ and $\sigma^{(j)}$ is obtained by flipping the jth sign, then $0 \leq a_{\sigma^{(j)}} - a_\sigma \leq 2(1-\mu)\omega_j\delta_j/\mu$.

Proof. We need the ordered stability from Proposition 2, applied to the values with s rounds remaining: for $h' \leq h$ coordinatewise and every $s \geq 0$,

$$0 \leq \left[W_s^{h'}(1) - W_s^{h'}(0)\right] - \left[W_s^h(1) - W_s^h(0)\right] \leq c_e \|h - h'\|_\infty. \tag{106}$$

For $a \geq 0$ such that $h - a\mathbf{1} \in [0,1]^K$, let $s \to \infty$ with $h' := h - a\mathbf{1}$. Using equation 69 yields

$$0 \leq \overline{\tau}(h) - \overline{\tau}(h - a\mathbf{1}) \leq (1-\mu)a. \tag{107}$$

Define, for $a \in [0, A_\sigma]$, $F_\sigma(a) := \tau - \overline{\tau}(\tau\mathbf{1} - a\mathbf{1} + \sigma \odot \delta)$ where $\odot$ denotes componentwise multiplication. Since $A_\sigma \leq \delta_{\max}/\mu$, for $a \in [0, A_\sigma]$ we have

$$a + \delta_{\max} \leq \frac{\mu m}{24} + \frac{\mu^2 m}{24} \leq \frac{m}{12}, \qquad a + \delta_{\max} \leq \frac{c_{\mathrm{hold}}}{24c_e} + \frac{\mu c_{\mathrm{hold}}}{24c_e} \leq \frac{c_{\mathrm{hold}}}{12c_e},$$

so every conditional accuracy vector in this interval belongs to $\mathcal{N}$. Equation 107 gives $|F_\sigma(a) - F_\sigma(a')| \leq (1-\mu)|a - a'|$, so F_σ is a contraction.

We next check that F_σ maps $[0, A_\sigma]$ into itself. Fix a in this interval, write $h = \tau\mathbf{1} - a\mathbf{1} + \sigma \odot \delta$, and set $z := F_\sigma(a)$. Set $w_h := (d_h(1) - d_h(0))/c_e \geq 0$ when $\bar{B} \geq 2$ and $w_h := +\infty$ when $\bar{B} = 1$. By equation 69 and the definition of z, $d_h(0) = (1-\mu)u_h(1) = (1-\mu)c_{\mathrm{hold}}/\mu + c_e z$, hence $r_j^h - d_h(0) = c_e(a - \sigma_j\delta_j - z)$. Since $\mathrm{med}\{d_h(0), d_h(1), r\} - d_h(0) = \min\{(r - d_h(0))_+, d_h(1) - d_h(0)\}$, equation 67 at $q = 1$ (equation 68 when $\bar{B} = 1$) gives

$$\mu z = (1-\mu) \sum_{j=1}^K \omega_j \min\{(a - \sigma_j\delta_j - z)_+, w_h\}. \tag{108}$$

In particular $z \geq 0$. If $z > A_\sigma$, then $z > a$ and only negative coordinates can contribute to the right-hand side, each by at most δ_j. Thus $\mu z \leq (1-\mu)S_\sigma^- = \mu A_\sigma$, a contradiction. Since F_σ maps the closed interval $[0, A_\sigma]$ into itself and is a contraction, it has a unique fixed point $a_\sigma \in [0, A_\sigma]$.

Set $h_j^\sigma := \tau - a_\sigma + \sigma_j\delta_j$, as in equation 88, and write $g_\sigma := g_{h^\sigma}$, $u_\sigma := u_{h^\sigma}$. The fixed-point identity gives equation 89, so $\overline{M}^{h^\sigma}(0) - L^{h^\sigma}(j) = \sigma_j c_e \delta_j$. Moreover, setting $z = a_\sigma$ in equation 108 yields equation 90.

We need one sharper comparison between neighboring centers. Let $\sigma_j = +1$ and let $\sigma^{(j)}$ be obtained by flipping the jth sign. At a common value of a, the conditional accuracy vector for $\sigma^{(j)}$ is coordinatewise no larger than that for σ. Applying equation 106 and letting $s \to \infty$ gives

$F_{\sigma^{(j)}}(a) \geq F_\sigma(a)$. Both maps are nondecreasing contractions. Starting the iteration of $F_{\sigma^{(j)}}$ at a_σ therefore produces a nondecreasing sequence whose limit is its fixed point, so $a_{\sigma^{(j)}} \geq a_\sigma$. Write $\Delta a := a_{\sigma^{(j)}} - a_\sigma$. Evaluating equation 64 at $b = 0$, where $u_\sigma(0) = P_0 u_\sigma(0) = 0$ and $P_1 u_\sigma(0) = \overline{M}^{h^\sigma}(0) - c_{\exp}$, and using equation 89 gives

$$g_\sigma = \Omega(M_\tau + c_e a_\sigma) - c_e \sum_{k:\sigma_k=+1} \omega_k \delta_k, \qquad \Omega := \sum_{k=1}^{K} \omega_k. \tag{109}$$

Let φ_σ be a stationary minimizing policy under h^σ. Starting from queue state one, the queue stays at most r until its first hitting time τ_0 of zero, so equation 105 gives $\mathbb{E}_1^{\varphi_\sigma}[\tau_0] \leq 2/\mu$. From equation 109, $g_{\sigma^{(j)}} - g_\sigma = c_e \Omega \Delta a + c_e \omega_j \delta_j$. For the same queue state and action rule φ_σ, the conditional one-step cost difference satisfies

$$\mathbb{E}\big[\ell_{h^{\sigma^{(j)}}}(B_t, J_t, A_t) - \ell_{h^\sigma}(B_t, J_t, A_t) \,\big|\, B_t\big] \leq c_e \Omega \Delta a + 2 c_e \omega_j \delta_j,$$

because acceptance losses increase by $c_e \Delta a$ at every self-confidence level and by an additional $2c_e\delta_j$ at level j, while deferral costs do not change. Hence

$$\mathbb{E}\big[(\ell_{h^{\sigma^{(j)}}} - g_{\sigma^{(j)}}) - (\ell_{h^\sigma} - g_\sigma) \,\big|\, B_t\big] \leq c_e \omega_j \delta_j.$$

By equation 104 at $b = 1$, with equality for h^σ under φ_σ and as an inequality for $h^{\sigma^{(j)}}$ under the same φ_σ, subtracting the two relations yields

$$u_{\sigma^{(j)}}(1) - u_\sigma(1) \leq c_e \omega_j \delta_j \, \mathbb{E}_1^{\varphi_\sigma}[\tau_0] \leq \frac{2c_e}{\mu} \omega_j \delta_j.$$

Since equation 89 and equation 69 give $c_e \Delta a = (1-\mu)\{u_{\sigma^{(j)}}(1) - u_\sigma(1)\}$, we obtain

$$0 \leq \Delta a \leq \frac{2(1-\mu)}{\mu} \omega_j \delta_j. \tag{110}$$

□

Lemma 22 (Empty-queue occupancy)**.** *Fix a conditional accuracy vector h and an algorithm π, and let r be as in Lemma 20. Write $N_0(T) := \sum_{t=1}^{T} \mathbf{1}\{B_t = 0\}$ and $H_T := \sum_{t=1}^{T} \mathbf{1}\{B_t \geq r, A_t = 1\}$. Then $\mathbb{E}_h^\pi[N_0(T)] \geq \mu T/2 - 2\mathbb{E}_h^\pi[H_T]$.*

Proof. Write $p_+ := \mathbb{P}(C_t \geq r)$, as in equation 101, and drop the sub- and superscripts of $\mathbb{E}_h^\pi$. With $W_t := \lceil (B_t - r)_+ / r \rceil$, pathwise $W_{t+1} - W_t \leq \mathbf{1}\{B_t \geq r, A_t = 1\} - \mathbf{1}\{B_t > r, A_t = 0, C_t \geq r\}$. Thus

$$p_+ \, \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{B_t > r, A_t = 0\}\right] \leq \mathbb{E}[H_T].$$

A round with $B_t + A_t \leq r$ and $C_t \geq r$ leaves the next queue empty; the number of rounds with $B_t + A_t > r$ is at most the count in the preceding display plus H_T. Since $B_1 = 0$, the number of empty next-round queues $B_{t+1} = 0$ is at most $N_0(T)$. Using equation 101 therefore gives $\mathbb{E}[N_0(T)] \geq \mu T/2 - 2\mathbb{E}[H_T]$, as claimed. □

Lemma 23 (Terminal comparison for the hard family)**.** *For every h^σ from Lemma 21 and every $b \in \mathcal{B}$, write $u_\sigma := u_{h^\sigma}$ and let r be as in Lemma 20. Then $0 \leq u_\sigma(b) - D(b) \leq (2c_e S_\sigma^- / \mu)\lceil b/r \rceil$.*

Proof. Let φ^0 be the stationary policy that always accepts and τ_0 its first hitting time of queue state zero. By equation 103, $\mathbb{E}_b^{\varphi^0}[\tau_0] \leq 2\lceil b/r \rceil / \mu$. Moreover, equation 104 gives

$$u_\sigma(b) \leq \mathbb{E}_b^{\varphi^0}\left[\sum_{t=1}^{\tau_0} (\ell_{h^\sigma}(B_t, J_t, 0) - g_{h^\sigma})\right], \qquad D(b) = \mathbb{E}_b^{\varphi^0}\left[\sum_{t=1}^{\tau_0} (\ell_{h^0}(B_t, J_t, 0) - g_{h^0})\right],$$

where the second identity uses Lemma 18, since φ^0 is a stationary minimizing policy at h^0. Moreover, equation 109 gives, at every queue state b,

$$\sum_{j=0}^{K} \omega_j \Big[\ell_{h^\sigma}(b, j, 0) - g_\sigma - \big(\ell_{h^0}(b, j, 0) - g_{h^0}\big)\Big] = c_e S_\sigma^-.$$

Subtracting the two stopped-cost representations and using the last display gives $u_\sigma(b) - D(b) \le c_e S_\sigma^- \, \mathbb{E}_b^{\varphi^0}[\tau_0]$; together with the bound on $\mathbb{E}_b^{\varphi^0}[\tau_0]$ and $u_\sigma \ge D$ from Lemma 16, this proves the claim. □

C.6.3 Lemmas for the instance-dependent lower bound

Lemma 24 (Transformed disadvantages)**.** *For $h \in [0,1]^K$, set $k_h(b,j,a) := c_{\mathrm{hold}} b + a\{c_{\exp} - L^h(j)\}$ and $\bar{g}_h := g_h - \sum_{k=0}^K \omega_k L^h(k)$. Then $(\bar{g}_h, u_h)$ solves equation 64 with ℓ_h replaced by k_h. The disadvantages values in equation 65 are unchanged, and equation 71 holds with the transformed gains. Define $v_h(b,j) := \min_{a \in \mathcal{A}(b,j)}\{k_h(b,j,a) + (P_a u_h)(b)\} - \bar{g}_h$. For a state $s = (b,k)$ and an action $a \in \mathcal{A}(b,k)$,*

$$\mathfrak{d}_h(s,a) = k_h(s,a) - \bar{g}_h + \mathbb{E}[v_h(S_{t+1}) \mid S_t = s, A_t = a] - v_h(s). \tag{111}$$

Moreover, for any instance, any algorithm, and any bounded $w : \mathcal{B} \times \{0,\dots,K\} \to \mathbb{R}$, with $J_{T+1} \sim \omega$ independently after the last decision and $S_{T+1} := (B_{T+1}, J_{T+1})$, the process $\mathcal{M}_t := \sum_{s=1}^t (w(S_{s+1}) - \mathbb{E}[w(S_{s+1}) \mid S_s, A_s])$, $0 \le t \le T$, is a martingale with respect to $\mathcal{H}_t := \mathcal{F}_t \vee \sigma(J_{t+1})$, with $|\mathcal{M}_t - \mathcal{M}_{t-1}| \le \mathrm{sp}(w)$ for $1 \le t \le T$.

Proof. Since $\ell_h(b,j,a) = k_h(b,j,a) + L^h(j)$ and $L^h(j)$ does not depend on the action, subtracting $\sum_{k=0}^K \omega_k L^h(k)$ from equation 64 gives the transformed ACOE for $(\bar{g}_h, u_h)$. The minimizing actions and disadvantages are unchanged, and equation 71 holds with the transformed gains.

For the martingale property, $\mathcal{M}_t$ is $\mathcal{H}_t$-measurable. Since $B_{t+1} = (B_t + A_t - C_t)^+$ with C_t independent of $(\mathcal{H}_{t-1}, A_t)$, and $J_{t+1} \sim \omega$ is independent of $(\mathcal{H}_{t-1}, A_t, C_t)$, the conditional law of S_{t+1} given $\mathcal{H}_{t-1}$ and A_t depends only on (S_t, A_t); the tower property therefore gives $\mathbb{E}[\mathcal{M}_t - \mathcal{M}_{t-1} \mid \mathcal{H}_{t-1}] = 0$. The increment bound follows from $|w(S_{t+1}) - \mathbb{E}[w(S_{t+1}) \mid S_t, A_t]| \le \mathrm{sp}(w)$. □

D Additional related work

D.1 Alternative self-confidence signals

Our experiments use the self-consistency score, i.e. the empirical mass of the largest answer class. This choice is natural when multiple responses are sampled, and answer consistency has previously been used as a signal to decide whether to escalate queries in LLM cascades (Yue et al., 2024). However, the deferral and learning components of our model do not depend on this particular definition of the score. They only require that queries be partitioned into levels $J_t \in [K]$, each with an unknown conditional accuracy $h_j^\star = \Pr(Y_t = 1 \mid J_t = j)$ that can be learned from selectively observed expert feedback. Thus, other scalar self-confidence measures can be substituted for self-consistency and discretized into levels before applying the same deferral framework. In particular, neither the deferral rule nor our analysis requires the self-confidence signal itself to be calibrated or the conditional accuracy to be monotone in that signal.

Sample-based self-confidence. Let $\mathcal{C}_t$ denote the set of answer classes represented among the G sampled responses, with empirical frequencies

$$q_{t,c} := \frac{1}{G} \sum_{g=1}^{G} \mathbf{1}\{\widehat{R}_{t,g} \in c\}, \qquad c \in \mathcal{C}_t.$$

Writing $q_{t,(1)} \geq q_{t,(2)} \geq \cdots$ for the ordered frequencies, with $q_{t,(2)} = 0$ if $|\mathcal{C}_t| = 1$, the self-consistency score is simply $q_{t,(1)}$. It therefore retains only the empirical mass of the most frequent answer class.

A natural alternative is to use the entropy of the empirical answer distribution,

$$H_t := -\sum_{c \in \mathcal{C}_t} q_{t,c} \log q_{t,c}.$$

which takes all sampled answer classes into account. For tasks with canonical answers, the classes can be defined directly by the extracted answer, yielding an entropy over answer labels. For open-ended tasks, responses can instead be grouped according to semantic equivalence, giving the discrete variant of semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024). Semantic entropy has also been considered directly as an uncertainty signal for LLM routing (Zhang et al., 2025).

Another simple alternative is the margin between the two most frequent answer classes,

$$q_{t,(1)} - q_{t,(2)},$$

which measures how strongly the dominant answer is separated from its closest competitor. Margin-based self-confidence scores have been used to control escalation in language-model cascades (Ramírez et al., 2024). The statistic above applies the same principle directly to the empirical distribution of sampled answer classes.

For open-ended generation, self-confidence can also be estimated without first assigning responses to hard equivalence classes. Kernel Language Entropy (KLE) (Nikitin et al., 2024), for example, constructs a kernel from pairwise semantic similarities between sampled responses and measures uncertainty through its von Neumann entropy. Semantic voting (Jiang et al., 2026) also aggregates pairwise semantic similarities, using them to select responses for model self-improvement.

Model-internal self-confidence. When model logits are available, self-confidence can instead be estimated from the predictive distributions produced during generation. For a generated response $\widehat{R}_t = (\widehat{r}_{t,1}, \ldots, \widehat{r}_{t,L_t})$, let

$$p_{t,\ell} := \Pr(\widehat{r}_{t,\ell} \mid X_t, \widehat{r}_{t,<\ell})$$

denote the probability assigned by the model to the generated token at position ℓ. For example, Gupta et al. (2024) consider the sequence log-probability, which they refer to as the *Chow-Sum*,

$$S_t^{\mathrm{sum}} := \sum_{\ell=1}^{L_t} \log p_{t,\ell},$$

as a measure of model confidence. Since this score is affected by sequence length, they also consider the length-normalized *Chow-Average* score

$$S_t^{\text{avg}} := \frac{1}{L_t} \sum_{\ell=1}^{L_t} \log p_{t,\ell}.$$

These and other model-internal scores can be incorporated into our model by discretizing the resulting scalar self-confidence signal into levels and learning the corresponding conditional accuracies online.

These alternatives modify the self-confidence signal supplied to the deferral layer, but not the central learning problem studied in this work. In all cases, the system must learn from delayed and selectively observed feedback which score levels are sufficiently reliable to accept, while simultaneously accounting for the state-dependent cost of expert intervention.

D.2 Probabilistic calibration

Calibration is the property that characterizes the relationship between a model's confidence scores and the probability that its predictions are correct: a confidence score is calibrated if predictions assigned confidence p are correct with probability p. Classical post-hoc calibration methods include histogram binning (Zadrozny & Elkan, 2001), Platt scaling (Platt, 1999), isotonic regression (Robertson et al., 1988; Zadrozny & Elkan, 2002). In our problem, the conditional accuracies are learned from delayed, selectively observed human feedback, and learning is evaluated through deferral regret rather than calibration error.

D.3 Additional learning-to-defer formulations

Charusaie et al. (2022) study learning predictors that complement humans and develop active learning schemes to reduce the number of expert predictions required during training. In our model, expert feedback is instead acquired during deployment to learn the conditional accuracies of a fixed LLM.

Learning to defer with multiple experts. Building on the single-expert setting, learning to defer has been generalized to multiple heterogeneous experts whose competencies may vary across inputs. In this setting, the learner must determine not only whether to defer, but also which expert should handle each instance. Early multi-expert formulations jointly learned a predictive model and an allocation rule while accounting for heterogeneous expert competencies and, in some cases, expert biases and fairness considerations (Keswani et al., 2021; Hemmer et al., 2022). Subsequent work developed more principled formulations for expert selection, including calibrated and consistent surrogate objectives (Verma et al., 2023). A complementary line of work has focused on the theoretical foundations of multi-expert deferral. Mao et al. (2023) study the two-stage setting, in which a predictor is fixed and a deferral function is learned to select among experts, while Mao et al. (2024a) develop surrogate losses with realizable H-consistency, H-consistency, and Bayes-consistency guarantees. More recently, Mao et al. (2025) provide a unified treatment of single-stage and two-stage multi-expert deferral and establish stronger consistency guarantees for both settings. Other extensions consider learning with limited expert annotations (Hemmer et al., 2023), adaptation to previously unseen experts (Tailor et al., 2024), regression (Mao et al., 2024b), imbalanced expert data (Cortes et al., 2026), and Top-k deferral, where multiple experts may be selected for the same instance (Montreuil et al., 2026).

D.4 LLM routing

FrugalGPT constructs LLM cascades to reduce monetary cost while maintaining predictive quality (Chen et al., 2024). RouteLLM learns routers between stronger and weaker models using preference data (Ong et al., 2025). Jitkrittum et al. (2026) introduce UniRoute, a model-routing framework that represents LLMs through their prediction-error profiles on representative prompts, thus enabling cost-aware routing to previously unseen models without retraining the router. These methods allocate computation among automated models. We first sample from a fixed LLM, then decide whether to defer to a capacity-constrained human expert whose completed reviews supply ground-truth labels.

D.5 Structured admission control

Queue-aware learning has also been studied through structured admission-control policies. In particular, UCRL-AC (Weber et al., 2024) exploits the structure of multi-class admission control to obtain sharper worst-case regret guarantees than generic MDP approaches, while SALMUT (Roy et al., 2021) assumes that the optimal policy has an ordered multi-threshold structure and directly learns the corresponding vector of class-specific thresholds using stochastic approximation. The learning problems differ substantially from ours. In Weber et al. (2024), arrival rates are unknown and are learned passively from exogenously observed arrivals, while the class-dependent reward model is known. Roy et al. (2021) allows the transition dynamics to be unknown, but learns from state, reward, and transition observations generated by interaction with the system. In our model, by contrast, the confidence level-dependent cost of accepting the LLM answer is itself unknown since it depends on the unknown conditional accuracy vector, and it is learned only through queries deferred to the expert. Thus, obtaining information about prediction quality requires consuming expert capacity and increasing queue congestion, with the resulting feedback additionally delayed by expert service.

D.6 Active learning and oracle queries

Active learning studies how a learner can selectively acquire costly labels while maintaining predictive performance (Dasgupta, 2011; Hanneke, 2014). Online variants require these query decisions to be made sequentially under partial feedback (Hao et al., 2018). Réveillard et al. (2026) study online classification in which costly expert consultations provide labeled examples used to improve subsequent predictions. In our model, the LLM remains fixed, and expert feedback instead estimates conditional accuracies to guide deferral. Russo et al. (2024) study a limited budget of best-action queries, which reveal the identity of the best action before the current decision. This information differs from our completed-review labels, which reveal whether the LLM response was correct without revealing the optimal deferral decision.

D.7 Human–AI decision support

Related work studies how the relationship between human and AI confidence affects the usefulness of AI advice (Corvelo Benz & Rodriguez, 2023; Corvelo Benz & Gomez Rodriguez, 2025). Benz et al. (2026) study learning decision policies through repeated interactions under human-alignment and full feedback. In our model, an accepted LLM response provides no correctness feedback, and expert review is limited by queue capacity.

E DATASETS

E.1 SYNTHETIC DATASETS

In this section, we describe the two synthetic instances used in Appendix G. Both have $K = 10$ self-confidence levels and are shown in Figure 5.

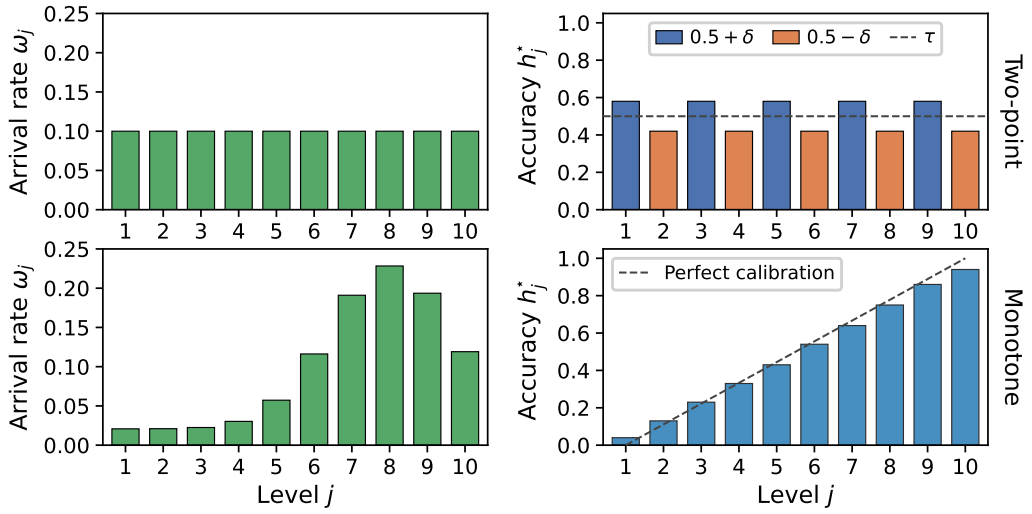


Figure 5: Top: two-point instance with $\delta = 0.08$. Bottom: monotone instance (dashed line: perfect calibration, i.e., $h_j^\star = p_j$ with $p_j = (j-1)/9$).

Two-point instance. The first instance uses the two-point family $h_j^\star \in \{0.5-\delta, 0.5+\delta\}$, $j \in [K]$, with uniform arrival rates $\omega_j = 1/K$ (top panel of Figure 5). This construction is motivated by the hard instances used in the proof of our minimax lower bound (Theorem 4): under immediate service ($\mu = 1$), uniform arrivals and $c_{\text{exp}}/c_e = 1/2$, the oracle threshold is $\tau = 1/2$, and if $K \leq T$, these instances take $h_j^\star \in \{\tau - \delta, \tau + \delta\}$ with $\delta = \frac{1}{8}\sqrt{K/T}$. Appendix G.1 uses this instance in that setting only, with δ as a free parameter.

Monotone instance. The second instance combines increasing conditional accuracies $h^\star = (.04, .13, .23, .33, .43, .54, .64, .75, .86, .94)$ with arrival rates concentrated on the high levels (bottom panel of Figure 5). Since the scores are $p_j = (j-1)/9$, they are nearly calibrated: this instance models an LLM that is slightly underconfident at the three lowest levels and overconfident at the others, by at most 0.06. Since the conditional accuracies span most of $[0, 1]$, the oracle threshold falls between the accuracies of two levels for a wide range of service rates, costs and queue sizes. This ensures the optimal actions do not become trivial to learn when we vary these parameters, unless deferral is never worthwhile ($c_{\text{exp}} \geq c_e$). We use this instance in Appendix G.2, where we vary the service rate μ, the expert cost c_{exp} and the holding cost c_{hold}, with the conditional accuracies held fixed.

E.2 REAL-WORLD DATASETS

In this section, we describe the datasets used in our real-world experiments. We start from three question-answering datasets with verifiable answers:

- GSM8K (Cobbe et al., 2021) contains grade-school math problems with numerical answers. We use all $8{,}792$ questions from its train and test splits.
- NuminaMath-TIR (Li et al., 2024) contains competition-style mathematics problems. We use its first $10{,}000$ training questions, from which we remove 157 exact duplicates and 659 questions that also appear in GSM8K, leaving $9{,}184$.

- MedMCQA is a large-scale medical dataset derived primarily from medical entrance and professional examinations (Pal et al., 2022). It contains multiple-choice questions with four options. We use all 187,005 questions of its train and validation splits.

We use Qwen3-4B-Instruct-2507, Qwen3-0.6B, and Qwen2.5-0.5B-Instruct (Yang et al., 2025a; Qwen et al., 2025), and sample 20 answers per question. Answers are sampled with vLLM at temperature 0.6 and top-p 0.95, without top-k truncation, with up to 2,048 generated tokens per answer and a maximum model sequence length (prompt plus output) of 4,096 tokens.

Answer extraction and grading. A final answer is extracted from each generation by taking the last complete `<answer>` block. If none is present, we use a terminal `\boxed{}` expression or an explicit final-answer declaration. Otherwise, no answer is returned. Answers are grouped into equivalence classes, either by mathematical equivalence with Math-Verify (Hugging Face, 2025) for the math datasets, which handles fractions, radicals, tuples, sets and symbolic expressions, or by option letter for MedMCQA. The same equivalence decides whether $R_t \equiv R_t^\star$. Ties between the largest classes are broken by first occurrence. Generations without an extractable answer count in the denominator of the score but cast no vote, so a question on which no generation yields an answer has score 0 and $Y_t = 0$. Questions whose ground-truth answer could not be parsed, or whose grading timed out, are removed: 58 NuminaMath questions for the Qwen3 models, and 85 NuminaMath and 12 GSM8K questions for Qwen2.5-0.5B.

Instances. Figure 6 shows $h^\star$ and ω for all (dataset, LLM) pairs, from which the instances are built as in Section 6. The regimes of Figure 2 extend to Qwen2.5-0.5B: its conditional accuracy is above the score on the math datasets, especially on GSM8K, and varies little with the score on MedMCQA. Across all three datasets, the self-consistency score is generally not calibrated, i.e. $h^\star \neq p$. Moreover, $h^\star$ generally is not monotone in the self-consistency level j. This motivates learning $h^\star$ rather than assuming the score is calibrated.

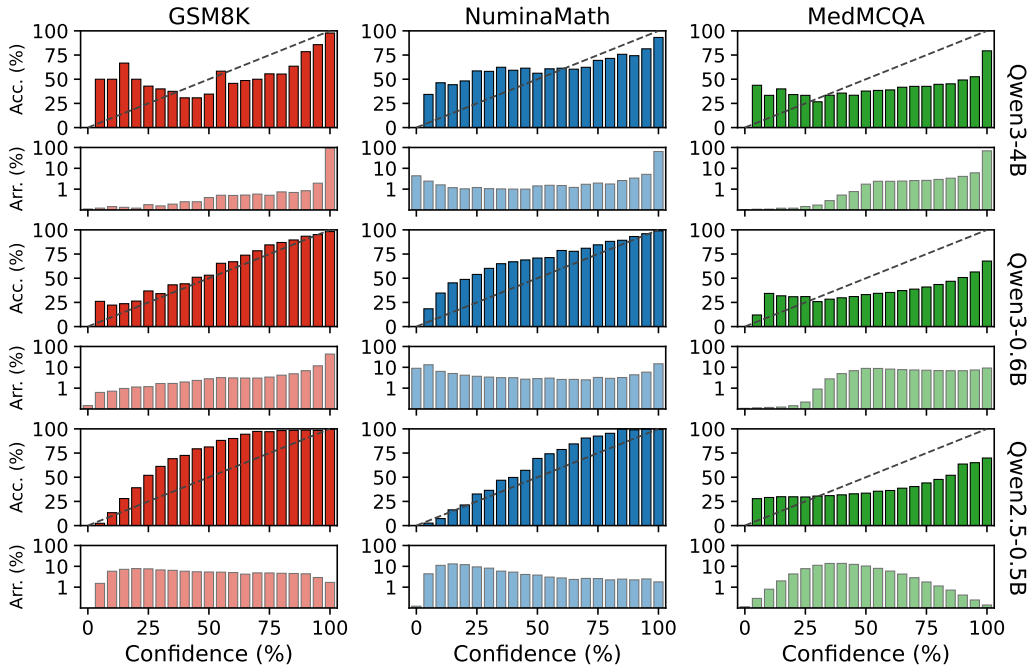


Figure 6: Conditional accuracy (top) and arrival rates (bottom) of the self-consistency score. The dashed line represents a perfectly calibrated score.

F BASELINES AND IMPLEMENTATION DETAILS

The costs $(c_e, c_{\exp}, c_{\text{hold}})$, the service distribution ρ_C and the queue capacity $\bar{B}$ are known to all algorithms. In the synthetic experiments, ω is known. In the real-world experiments it is unknown, and every learning algorithm that needs it uses the empirical frequencies $\hat{\omega}_{j,t} = O_{j,t}/t$, where $O_{j,t} = \sum_{s \le t} \mathbf{1}\{J_s = j\}$ counts the arrivals at self-confidence level j up to round t. Let $N_{j,t}$ and $C_{j,t}$ be the number of level-j labels received before the decision of round t and the number of correct ones among them, and $\widehat{h}_{j,N_{j,t}} = (C_{j,t}/N_{j,t})\mathbf{1}\{N_{j,t} > 0\}$ their empirical mean.

Plug-in algorithms (`KL-PBD`, `KL-PBD+`, `GBD`, `NBD`). These four algorithms share the decision rule of `KL-PBD`, $A_t = \mathbf{1}\{c_e(1 - h_{J_t,t}) > M_t^{h_t}(B_t)\}$, and differ only in the conditional accuracy vector h_t plugged into the marginal cost of deferral $M_t^{h_t}$:

- `KL-PBD`: the KL lower confidence bound $\underline{h}_{j,t}$ with confidence budget $\phi(T) = \log T + \log\log T$;
- `KL-PBD+`: the same bound with the level-dependent budget $\log(O_{j,t}/N_{j,t})$ in place of $\phi(T)$, in the spirit of the KL-UCB+ family of algorithms (Garivier & Cappé, 2011; Achab et al., 2018);
- `GBD` (Greedy Bellman Deferral): the empirical mean $\widehat{h}_{j,N_{j,t}}$;
- `NBD` (Naive Bellman Deferral): the surrogate $h = p$, i.e. the score is taken as calibrated and nothing is learned about $h^\star$.

With known ω, the finite-horizon recursion is re-solved whenever a new label arrives (never for `NBD`). With unknown ω, it is re-solved at every round with the current $(h_t, \hat{\omega}_t)$, for `NBD` included. The marginal costs are computed exactly by backward induction over the remaining horizon, with the terminal draining cost $D(b)$ of Section 3.

Fixed-threshold policies (`FCT`, `FQT`). Fixed Confidence Threshold defers iff $p_{J_t} < \tau^{\text{conf}}$; Fixed Queue Threshold defers iff $B_t < \tau_{J_t}^{\text{queue}}$ for a threshold vector $\tau^{\text{queue}} \in \{0, \dots, \bar{B}\}^K$. Both are tuned once before the run to minimize the expected cost under the surrogate $h = p$, computed exactly by dynamic programming. When ω is unknown the uniform distribution $\omega_j = 1/K$ is used in place of ω. Since only the ordering of $\{p_1, \dots, p_K\}$ matters for `FCT`, its $K+1$ candidate policies are enumerated exhaustively. For `FQT`, exhaustive search over $\{0, \dots, \bar{B}\}^K$ is infeasible, so we use coordinate descent: each threshold τ_j^{queue} is optimized over $\{0, \dots, \bar{B}\}$ with the others fixed, cycling over $j = 1, \dots, K$ until the vector no longer changes. We run this from several deterministic initializations and keep the best result, which need not be the global optimum.

`SALMUT-SF`. We adapt `SALMUT` (Roy et al., 2021), which defers a level-j query iff $B_t < \tau_j$, with one queue threshold $\tau_j \in [0, \bar{B}]$ per level, initialized at $\tau_j = 0$. In the setting of Roy et al. (2021), h would be known. In our case, we estimate it with $\widehat{h}_{j,N_{j,t}}$. Let $b = B_t$, $j = J_t$, $a = A_t$ and $b' = B_{t+1}$. At the end of round t, after service and label processing, the relative-value critic, initialized at $V(b) = 0$, is updated at the pre-decision queue state b, including when $b = \bar{B}$, with step size $\alpha_{n_b} = (\lfloor n_b/100 \rfloor + 2)^{-0.6}$ where n_b counts the updates at state b:

$$V(b) \leftarrow V(b) + \alpha_{n_b}\big[c_{\text{hold}} b + c_{\exp} a + c_e(1 - \widehat{h}_{j,N_{j,t+1}})(1 - a) + V(b') - V(0) - V(b)\big].$$

For the actor, we retain `SALMUT`'s sigmoid interpolation $f(b, \tau) = 1/(1 + e^{-(b - \tau - 1/2)})$. Note that unlike Roy et al. (2021), the transition probabilities are assumed known in our setting, so they do not need to be estimated. Therefore, after the critic update, for $b < \bar{B}$, we estimate the advantage function via the updated critic, $\hat{Q}_1 - \hat{Q}_0 = c_{\exp} - c_e(1 - \widehat{h}_{j,N_{j,t+1}}) + (P_1 V)(b) - (P_0 V)(b)$, where $(P_a V)(b) = \sum_{b'} P_a(b, b') V(b')$, and for $b < \bar{B}$ we update the threshold of the arriving level as

$$\tau_j \leftarrow \Pi_{[0,\bar{B}]}\Big[\tau_j - \tfrac{10}{t+100}\, f(b, \tau_j)\big(1 - f(b, \tau_j)\big)(\hat{Q}_1 - \hat{Q}_0)\Big].$$

We do not use SALMUT's ordering constraint on the thresholds, since conditional accuracies need not be ordered in the score. When the threshold rule accepts, we force a feasible deferral with probability $\varepsilon_t = \min\{.5,\, 2/\sqrt{t}\}$; without it a level can stop producing the labels needed to revise an optimistic threshold due to the selective nature of the feedback in our problem (this failure mode is explored in Appendix H.5). The same hyperparameters are used in every instance.

Queue-unaware bandit algorithms (`KL-UCB-4P`, `KL-UCB+-4P`). Under immediate service ($\mu = 1$), the queue stays empty and $M_t(0) = c_{\exp}$. The problem then reduces to a separate learning task for each self-confidence level: at each round the level $J_t = j$ is revealed and the learner faces a two-armed bandit with one known arm. Indeed, one can equivalently subtract the cost $c_e(1 - Y_t)$ of accepting from both actions, so that accepting has reward 0, while deferring has reward $c_e(1 - Y_t) - c_{\exp}$, with unknown mean $c_e(1 - h_j^\star) - c_{\exp}$, observed only when the query is deferred. This is close to the profitable-bandit problem of Achab et al. (2018), played independently at each level, which motivates a comparison with their algorithms. We adapt their algorithms to our setting, so that they defer iff $c_e \bar{q}_{j,t} \geq c_{\exp}$, ignoring the queue except when it is full, where $\bar{q}_{j,t}$ is the KL upper confidence bound on the LLM error probability $1 - h_j^\star$, computed from $1 - \widehat{h}_{j,N_{j,t}}$ with budget $\log O_{j,t} + 3 \log\log O_{j,t}$ for `KL-UCB-4P` and $\log(O_{j,t}/N_{j,t})$ for `KL-UCB+-4P`, using the level clock $O_{j,t}$ in place of t, with unlabeled levels having $\bar{q}_{j,t} = 1$ so that they are deferred. Equivalently, up to ties, they follow the decision rule of `KL-PBD` with $M_t^{h_t}(B_t)$ replaced by $c_{\exp}$, its value under immediate service: `KL-UCB+-4P` is then `KL-PBD+`, and `KL-UCB-4P` is `KL-PBD` with budget $\log O_{j,t} + 3 \log\log O_{j,t}$ in place of $\phi(T)$. When $\mu < 1$, labels arrive only once a deferred query is served, and the algorithms use the labels available at each round.

Simulation protocol. For every run, all algorithms see the same arrivals (J_t), latent labels (Y_t) and service capacities (C_t), and keep separate first-in, first-out (FIFO) queues. A label completed after the decision of round t is available from round $t + 1$. We compute the regret as $\hat{R}_T = \sum_{t \leq T,\, B_t < \bar{B}} |L^\star(J_t) - M_t^\star(B_t)| \mathbf{1}\{A_t \neq A_t^\star\}$ (Lemma 2), averaged over independent runs with pointwise 95% normal confidence intervals. The code used for our experiments is provided in the supplementary files.

G SYNTHETIC EXPERIMENTS

G.1 EFFECT OF K AND OF THE ORACLE GAP ON KL-PBD REGRET

In this experiment, we examine how the number of levels, the separation between the accuracies and the oracle threshold, and the arrival distribution ω affect KL-PBD's regret. We use the two-point instance of Appendix E.1 (top row of Figure 5) with $\mu = 1$, $c_e = 1$ and $c_{\text{exp}} = 0.5$, so that the oracle threshold is $\tau = 1/2$: the oracle defers the levels j with $h_j^\star = 0.5 - \delta$ and accepts those with $h_j^\star = 0.5 + \delta$. All experiments use $T = 2 \cdot 10^5$ and 100 runs.

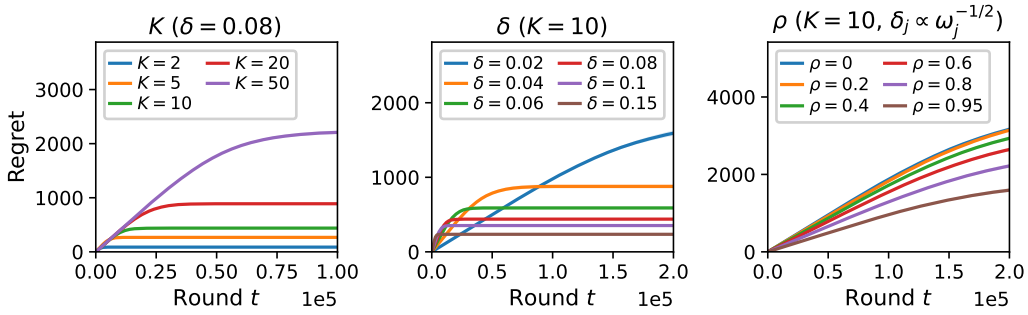


Figure 7: Regret of KL-PBD on the two-point instance for different K with $\delta = 0.08$ (left, shown up to $t = 10^5$) and different δ with $K = 10$ (center), and for different arrival distributions with $K = 10$ and $h_j^\star = 0.5 + \sqrt{\phi(T)/(2\omega_j T)}$ (right). 95% confidence intervals over 100 runs are displayed as bands but are too small to be visible.

Figure 7 shows that regret increases with K, since the learner must acquire more expert feedback to identify the correct action in a larger number of levels. The dependence on δ reflects the usual gap-dependent learning trade-off. Larger gaps make an incorrect deferral more costly initially, but they are statistically easier to distinguish from the oracle threshold, so the learner identifies the optimal action sooner and the regret curves flatten earlier. Conversely, when δ is small, the conditional accuracies lie closer to the threshold and require substantially more feedback to distinguish, leading to larger regret over the range of the figure.

Under immediate service, KL-PBD defers level j until $N_{j,t}\,\mathrm{kl}(\widehat{h}_{j,N_{j,t}}, 1/2) \geq \phi(T)$. A level with $h_j^\star = 0.5 + \delta$ therefore needs about $\phi(T)/(2\delta^2)$ labels before it is accepted, each deferral costing δ, while a level with $h_j^\star = 0.5 - \delta$ is deferred, as by the oracle. Each level above the threshold thus contributes about $\min\{\delta\,\omega_j T,\ \phi(T)/(2\delta)\}$ to the regret.

Table 1 complements Figure 7 by reporting the final regret at $T = 2 \cdot 10^5$. The results show that regret increases significantly with K, confirming the additional learning burden associated with a larger number of levels. At $\delta = 0.08$, every oracle decision is learned well before T, and each level above the threshold adds about $\phi(T)/(2\delta) \approx 90$ to the regret. For every value of K except $K = 50$, regret decreases as δ increases; we explain the exception below.

K	$\delta = 0.02$	$\delta = 0.04$	$\delta = 0.06$	$\delta = 0.08$	$\delta = 0.10$	$\delta = 0.15$
2	334 ± 23	179 ± 13	121 ± 9	87 ± 7	71 ± 5	50 ± 3
5	1085 ± 51	555 ± 21	357 ± 13	267 ± 12	209 ± 9	138 ± 6
10	1592 ± 36	878 ± 28	589 ± 20	439 ± 13	353 ± 12	235 ± 7
20	1948 ± 11	1757 ± 44	1181 ± 29	889 ± 17	706 ± 15	462 ± 11
50	1997 ± 2	3564 ± 30	2966 ± 42	2225 ± 34	1789 ± 27	1168 ± 14

Table 1: Final regret R_T of KL-PBD at $T = 2 \cdot 10^5$ for different K and δ, mean $\pm$ 95% confidence interval half-width over 100 runs.

Table 2 further illustrates the non-monotone dependence of finite-horizon regret on the oracle gap δ. For very small δ, distinguishing which side of the oracle threshold the accuracy is requires many observations, but the regret incurred by an incorrect decision is itself only $O(\delta)$, so the regret is at

δ	0.0025	0.005	0.01	0.015	$0.0192^{\dagger}$	0.025
R_T	250 ± 0	500 ± 0	998 ± 3	1442 ± 16	1596 ± 32	1396 ± 43
δ	0.03	0.04	0.06	0.08	0.10	0.15
R_T	1189 ± 41	878 ± 28	589 ± 20	439 ± 13	353 ± 12	235 ± 7

Table 2: Final regret R_T of KL−PBD at $T = 2 \cdot 10^5$ for $K = 10$ and different δ, mean $\pm$ 95% confidence interval half-width over 100 runs. †: $\delta^\star = \sqrt{\phi(T)K/(2T)} \approx 0.019$.

most of order δT. In this case, KL−PBD does not have enough time to learn the oracle decisions, and $R_T \approx \delta T/2$. For larger δ, the levels are statistically easier to distinguish and the regret decreases roughly as $K\phi(T)/(4\delta)$. Consequently, the largest regret is observed near the crossover between these two regimes, $\delta^\star = \sqrt{\phi(T)K/(2T)} \approx 0.019$ for $K = 10$. Since $\delta^\star$ grows with K ($\delta^\star \approx 0.043$ for $K = 50$), the regret for $K = 50$ is lower at $\delta = 0.02$ than at $\delta = 0.04$ (Table 1).

The right panel examines the effect of non-uniform arrivals with $K = 10$. We parameterize

$$\omega_1 = \frac{1}{K} + \rho\left(1 - \frac{1}{K}\right), \qquad \omega_j = \frac{1-\rho}{K}, \quad j = 2, \dots, K,$$

so that $\rho = 0$ corresponds to uniform arrivals and increasing ρ concentrates probability mass on a single level. We set $h_j^\star = 0.5 + \delta_j$ with $\delta_j = \sqrt{\phi(T)/(2\omega_j T)}$ for every level. Unlike in the two-point instance, all levels lie above the threshold: a level below it is deferred by both KL−PBD and the oracle and adds no regret, so this makes every level contribute. Level j receives about $\omega_j T$ arrivals and needs about $\phi(T)/(2\delta_j^2) = \omega_j T$ labels before it is accepted, so with this gap KL−PBD starts accepting level j at about the horizon. For uniform arrivals, $\delta_j = \delta^\star$. Each level then contributes about $\delta_j \omega_j T = \sqrt{\phi(T)\,\omega_j T/2}$ to the regret. Regret is largest under the uniform distribution and decreases as the arrival mass becomes more concentrated. This ordering is consistent with the dependence on ω in Theorem 2, whose leading term is $\widetilde{O}\big(\sqrt{T}\sum_{j=1}^K \sqrt{\omega_j}\big)$ and is maximized when $\omega_j = 1/K$ for $j \in [K]$. Quantitatively, $R_T/(\sqrt{T}\sum_{j=1}^K \sqrt{\omega_j})$ stays between 2.21 and 2.29 for all ρ.

G.2 Effect of service rate and cost parameters

In this experiment we use the monotone instance (bottom row of Figure 5) in order to evaluate how KL−PBD is affected by the expert service rate. We set the queue capacity to $\bar{B} = 20$, the LLM error cost to $c_e = 2$, the expert-review cost to $c_{\text{exp}} = 0.25$, and the holding cost to $c_{\text{hold}} = 0.1$. We vary the expert service rate over $\mu \in \{0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$, where in each round the expert completes a pending query with probability μ.

For each service rate, we run the system for $T = 50{,}000$ rounds and average results over 20 independent runs. We report the average queue size, the cumulative regret relative to the oracle, the response accuracy, i.e., the fraction of queries answered correctly, counting deferred queries as correct, and the average threshold, i.e., the time average of $\widehat{\tau}_t(B_t)$.

Table 3: Finite-horizon KL-PBD performance across service rates μ with $\bar{B} = 20$. Results are mean $\pm$ 95% confidence interval half-width over 20 runs.

Service rate μ	Average queue size	Cumulative regret R_T	Response accuracy	Average threshold
0.20	0.74 ± 0.01	194.06 ± 14.96	0.7710 ± 0.0010	0.2663 ± 0.0024
0.40	0.70 ± 0.01	233.59 ± 30.80	0.8445 ± 0.0008	0.5130 ± 0.0030
0.50	0.74 ± 0.00	115.49 ± 7.42	0.8856 ± 0.0005	0.5716 ± 0.0013
0.60	0.55 ± 0.01	135.36 ± 11.75	0.9070 ± 0.0006	0.6579 ± 0.0018
0.80	0.37 ± 0.00	301.17 ± 19.26	0.9603 ± 0.0011	0.7632 ± 0.0010
1.00	0.00 ± 0.00	80.61 ± 10.91	0.9937 ± 0.0001	0.8750 ± 0.0000

Table 3 shows that increasing the expert service rate generally makes deferral more attractive. As μ increases, deferred queries are processed more quickly, reducing the congestion cost associated with sending additional jobs to the expert. Consistent with this behavior, the average threshold increases from 0.2663 at $\mu = 0.2$ to $0.8750 = 1 - c_{\text{exp}}/c_e$ at $\mu = 1.0$, while response accuracy improves

substantially from 0.7710 to 0.9937. The average queue size remains below one across all service rates, falls from 0.74 at $\mu = 0.5$ to 0.37 at $\mu = 0.8$, and reaches zero at $\mu = 1.0$, where deferred jobs are cleared immediately.

Cumulative regret remains non-monotone in μ, ranging from 80.61 to 301.17 across the tested service rates. In particular, regret is relatively high at $\mu = 0.4$ and reaches its maximum at $\mu = 0.8$, despite the corresponding improvements in response accuracy and queue length. This reflects the fact that each service rate induces a different control problem and oracle policy. At $\mu = 1.0$, the experiment attains both its highest response accuracy, 0.9937, and its lowest cumulative regret, 80.61.

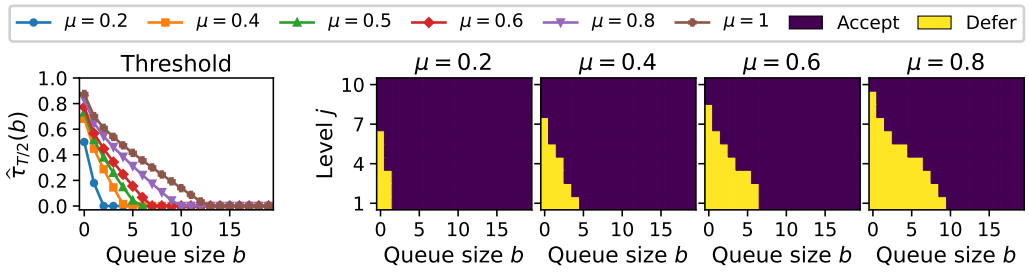


Figure 8: KL-PBD threshold $\widehat{\tau}_{T/2}(b)$ as a function of queue size b, averaged over the runs (left), and learned KL-PBD decisions, most frequent over the runs (right), at $t = T/2$ for different service rates μ.

Figure 8 (right) shows the policy learned by KL-PBD at $t = T/2$. It exhibits a clear threshold structure: within each level, the action changes from defer to accept once the queue size exceeds a level-dependent boundary, producing the staircase-shaped frontier visible across all panels. As the expert service rate μ increases, the defer region expands toward larger queue sizes, since a faster expert clears deferred requests more quickly.

Figure 8 (left) shows how the KL-PBD threshold $\widehat{\tau}_t(b)$ varies with the current queue size for different expert service rates. For every value of μ, the threshold decreases as the queue grows, reflecting the increasing congestion cost of sending additional queries to the expert. Once the queue becomes sufficiently large, the threshold reaches zero, and no query is deferred. Increasing the service rate shifts the threshold curve upward and extends it over larger queue sizes. When the expert is slow, as for $\mu = 0.2$, the threshold falls rapidly and reaches zero at a small queue size. In contrast, faster service allows the system to tolerate substantially more queued queries before the threshold becomes small.

We then fix $\mu = 0.5$ and vary the holding cost $c_{\text{hold}} \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$ with $c_{\text{exp}} = 0.25$, and the expert-review cost $c_{\text{exp}} \in \{0, 0.25, 1, 3\}$ with $c_{\text{hold}} = 0.05$, over 10 independent runs of $T = 50{,}000$ rounds. In these sweeps, the thresholds of KL-PBD are recomputed after every 50 changes of the lower bounds.

Table 4: KL-PBD performance across holding and expert-review costs with $\mu = 0.5$. Results are mean $\pm$ 95% confidence interval half-width over 10 runs.

c_{hold}	c_{exp}	Mean queue	Defer (%)	Accuracy (%)	Average threshold	Regret
0.005	0.25	2.840 ± 0.057	49.62 ± 0.18	91.27 ± 0.11	0.635 ± 0.003	138.20 ± 8.81
0.010	0.25	2.106 ± 0.059	48.68 ± 0.29	90.87 ± 0.13	0.638 ± 0.005	183.55 ± 14.20
0.050	0.25	1.042 ± 0.027	45.20 ± 0.24	89.48 ± 0.12	0.604 ± 0.004	145.86 ± 18.98
0.100	0.25	0.751 ± 0.007	43.52 ± 0.17	88.56 ± 0.10	0.571 ± 0.003	118.58 ± 13.78
0.500	0.25	0.229 ± 0.005	22.90 ± 0.39	81.06 ± 0.11	0.429 ± 0.003	222.55 ± 33.59
0.050	0.00	1.221 ± 0.010	48.72 ± 0.18	89.97 ± 0.10	0.611 ± 0.003	71.57 ± 10.97
0.050	1.00	0.264 ± 0.011	18.71 ± 0.79	80.20 ± 0.36	0.441 ± 0.001	432.32 ± 57.30
0.050	3.00	0.000 ± 0.000	0.00 ± 0.00	67.83 ± 0.12	0.000 ± 0.000	0.00 ± 0.00

Table 4 shows that the holding cost and expert-review cost affect KL−PBD in distinct ways. With $c_{\text{exp}} = 0.25$, increasing c_{hold} from 0.005 to 0.5 reduces the mean queue from 2.84 to 0.23 and the deferral rate from 49.62% to 22.90%. As fewer queries are routed to the expert, response accuracy decreases from 91.27% to 81.06%. The average threshold decreases overall from 0.635 to 0.429. Regret is non-monotone because changing c_{hold} also changes the oracle policy.

The expert-review-cost sweep exhibits a sharper transition. With $c_{\text{hold}} = 0.05$, increasing c_{exp} from 0 to 1 reduces the mean queue from 1.22 to 0.26, the deferral rate from 48.72% to 18.71%, and response accuracy from 89.97% to 80.20%; the average threshold correspondingly decreases from 0.611 to 0.441. For $c_{\text{exp}} \geq 3$, KL−PBD does not defer, and the accuracy of 67.83% is the LLM-only accuracy. This is due to the choice of cost parameters: the expected cost $c_e(1 - h^\star_{J_t}) \leq c_e = 2$ of accepting the LLM response is below the expert-review cost, so both KL−PBD and the oracle accept every query, yielding zero regret.

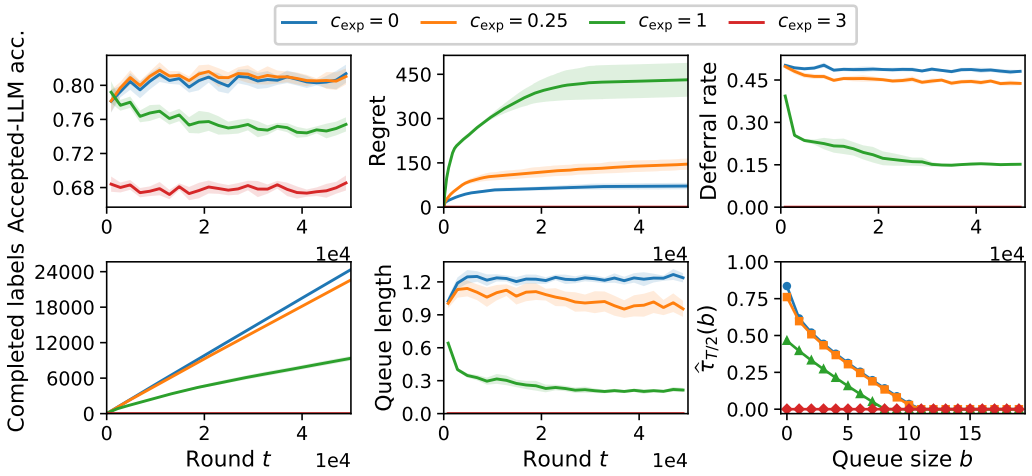


Figure 9: **Effect of the expert-review cost on finite-horizon KL−PBD.** We vary $c_{\text{exp}} \in \{0, 0.25, 1, 3\}$, with $T = 50{,}000$, $\mu = 0.5$, $\bar{B} = 20$, $c_e = 2$, and $c_{\text{hold}} = 0.05$. From the upper-left to the lower-right, the panels show accepted-LLM accuracy, cumulative regret, empirical deferral rate, cumulative completed expert labels, average queue size, and the deferral threshold $\widehat{\tau}_{T/2}(b)$ at $t = T/2$ as a function of queue size b.

Figure 9 demonstrates how the expert-review cost c_{exp} governs the operating regime of KL−PBD. For $c_{\text{exp}} \in \{0, 0.25\}$, expert review remains sufficiently inexpensive that the policy defers a large fraction of arriving queries, thereby maintaining a high accepted-LLM accuracy. As c_{exp} increases to 1, the policy becomes substantially more selective: the deferral rate decreases, more queries are answered directly by the LLM, and the accuracy of the accepted LLM responses decreases because LLM answers are also accepted at lower confidence levels. For $c_{\text{exp}} \geq 3$, expert review is no longer used, and the accepted-LLM accuracy equals the LLM accuracy.

The cumulative-regret trajectories show that the intermediate-cost regime is the most challenging. At $c_{\text{exp}} = 1$, expert review is still optimal in a substantial set of level–queue states, but less frequent deferrals slow the acquisition of informative expert feedback. Consequently, regret is substantially larger than for $c_{\text{exp}} \in \{0, 0.25\}$. The second row makes the underlying mechanism explicit. Increasing c_{exp} reduces the number of completed expert labels, lowers the average queue size, and shifts the threshold $\widehat{\tau}_{T/2}(b)$ downward.

Figure 10 illustrates the behavior shown in Table 4 over time: as c_{hold} increases, maintaining backlog becomes more expensive, so the policy becomes increasingly selective and a larger fraction of queries is answered directly by the LLM. It also shows that the reduction in expert use limits the feedback available for learning: at low holding costs, roughly half of all queries, about 2.5×10^4, yield completed expert labels, compared with only about 1.1×10^4 at $c_{\text{hold}} = 0.5$. The threshold curves provide the state-level view: for every c_{hold}, $\widehat{\tau}_{T/2}(b)$ decreases with backlog, with the decline becoming markedly steeper as the holding cost increases. In particular, the threshold reaches zero

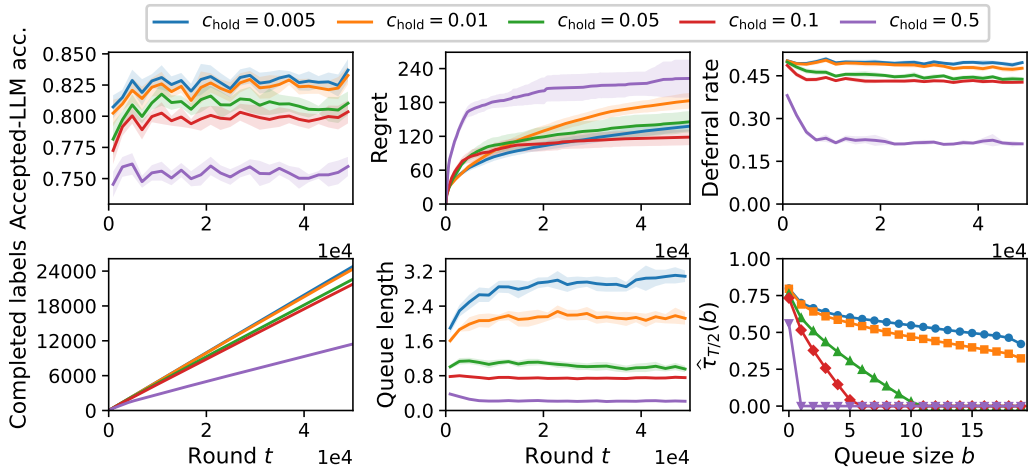


Figure 10: **Effect of the holding cost on KL−PBD.** We vary $c_{\text{hold}} \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$ with $T = 50{,}000$, $\mu = 0.5$, $B = 20$, $c_e = 2$, and $c_{\text{exp}} = 0.25$. From the upper-left to the lower-right, the panels show accepted-LLM accuracy, cumulative regret, empirical deferral rate, cumulative completed expert labels, average queue size, and the deferral threshold $\widehat{\tau}_{T/2}(b)$ at $t = T/2$ as a function of queue size b.

at moderate backlog for $c_{\text{hold}} = 0.1$ and almost immediately for $c_{\text{hold}} = 0.5$, showing that higher holding costs make the deferral rule increasingly sensitive to congestion.

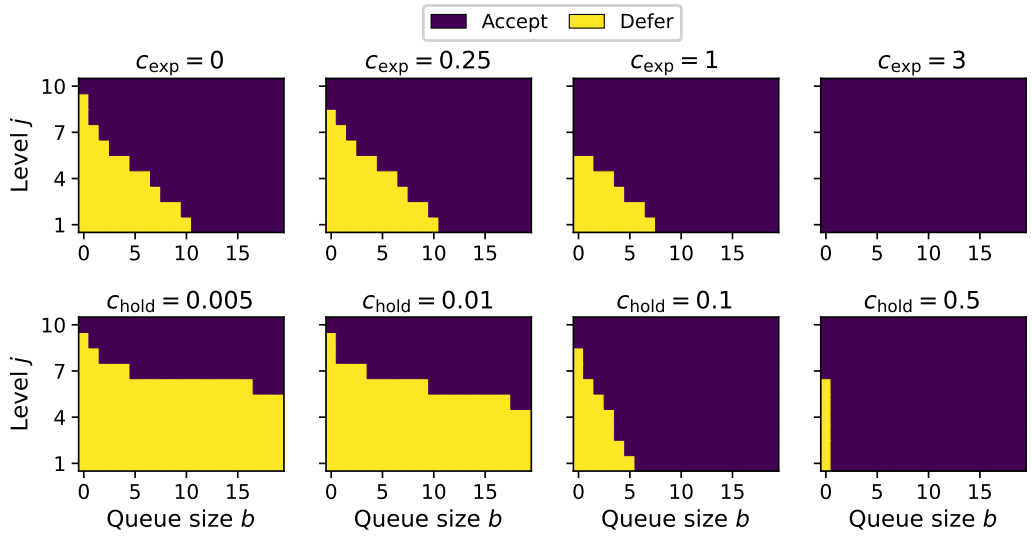


Figure 11: Learned KL−PBD decisions for varying c_{exp} and c_{hold}, for $t = T/2$ (most frequent decision over the runs).

Figure 11 shows how the learned KL−PBD deferral policy changes as the expert-review and holding costs increase. In both sweeps, higher costs make deferral less attractive, so the deferral region progressively shrinks: the yellow defer region contracts while the purple accept region expands, indicating that the policy relies increasingly on the LLM and sends fewer queries to the expert.

H ADDITIONAL REAL-WORLD EXPERIMENTS AND DETAILS

This section extends the experiments of Section 6. We first report the regret of all algorithms on the instances of Section 6 and on additional Qwen2.5-0.5B instances, and in a separate experiment, we report how the final regret grows with the horizon. We then vary the service rate to show the benefit of queue awareness, and compare the policy learned by KL−PBD with the oracle policy. Finally, we show how selective feedback makes GBD fail, which illustrates the need for pessimism. Unless stated otherwise, the setting is that of Section 6. Implementation details are given in Appendix F.

H.1 REGRET OF ALL ALGORITHMS

Table 5 and Figure 12 report the regret of all algorithms described in Appendix F on all instances. KL−PBD+ has the lowest regret in 8 of the 9 instances. The exception is Qwen2.5-0.5B on Numina-Math, where the score is nearly calibrated and NBD (102) is still below KL−PBD+ (115) at $T = 10^5$; a longer horizon is needed for KL−PBD+ to overtake NBD in this instance, as we show in Table 6. The queue-unaware algorithms KL−UCB−4P and KL−UCB+−4P incur 24 to 450 times the regret of KL−PBD+ on eight instances. KL−UCB+−4P is only competitive on Qwen3-4B GSM8K, where the LLM accuracy averages 95%, which makes deferral rare and the queue almost always empty. FCT is the worst of the three $h = p$ surrogate algorithms on every instance.

Table 5: Final regret at $T = 10^5$, mean ± 95% confidence half-width over 100 runs.

Model	Dataset	KL−PBD	KL−PBD+	GBD	SALMUT−SF	KL−UCB−4P	KL−UCB+−4P	NBD	FCT	FQT
Qwen3-4B	GSM8K	100.4 ± 3.3	35.4 ± 2.7	425.5 ± 20.5	531.0 ± 26.7	223.6 ± 5.9	71.3 ± 3.8	255.6 ± 1.3	756.3 ± 2.6	477.2 ± 1.9
Qwen3-4B	NuminaMath	424.4 ± 6.6	108.0 ± 4.2	1057.3 ± 82.0	1460.2 ± 41.0	7646.6 ± 101.8	4985.1 ± 156.7	231.9 ± 0.9	542.3 ± 1.2	195.4 ± 0.6
Qwen3-4B	MedMCQA	99.9 ± 2.9	40.2 ± 2.5	2236.7 ± 128.8	678.8 ± 13.8	7610.5 ± 191.7	5247.8 ± 87.7	2579.0 ± 4.0	5307.1 ± 5.4	4349.8 ± 5.3
Qwen3-0.6B	GSM8K	287.9 ± 7.0	93.5 ± 3.5	1468.3 ± 110.7	1173.0 ± 61.8	4654.6 ± 60.3	2227.5 ± 63.9	146.1 ± 2.6	2228.8 ± 3.8	185.5 ± 1.0
Qwen3-0.6B	NuminaMath	451.9 ± 7.3	159.4 ± 5.0	1264.1 ± 287.7	3007.0 ± 68.9	23600.1 ± 40.1	21539.7 ± 57.8	528.7 ± 1.3	8414.0 ± 27.7	1040.3 ± 2.3
Qwen3-0.6B	MedMCQA	106.0 ± 4.5	75.0 ± 5.8	2443.2 ± 218.0	306.0 ± 15.4	21834.8 ± 18.9	21831.9 ± 18.8	1501.1 ± 3.2	10439.5 ± 4.1	4070.2 ± 4.7
Qwen2.5-0.5B	GSM8K	314.1 ± 6.3	116.0 ± 3.8	1931.4 ± 317.9	1961.7 ± 43.2	20567.7 ± 109.6	17884.8 ± 107.5	1108.8 ± 2.4	3349.7 ± 20.2	1423.1 ± 2.4
Qwen2.5-0.5B	NuminaMath	298.5 ± 5.3	115.4 ± 3.6	1141.0 ± 296.9	1973.4 ± 34.0	23363.2 ± 34.0	22291.6 ± 54.5	102.3 ± 0.3	9785.0 ± 24.3	588.7 ± 1.7
Qwen2.5-0.5B	MedMCQA	100.6 ± 1.9	46.5 ± 2.2	2520.2 ± 168.7	161.6 ± 9.0	21041.6 ± 18.2	21039.9 ± 18.2	1365.7 ± 1.7	10027.7 ± 7.3	1003.3 ± 1.6

H.2 REGRET AS A FUNCTION OF THE HORIZON

Figure 13 and Table 6 report the final regret when the horizon varies in $T \in \{10^4, 5{\cdot}10^4, 10^5, 2{\cdot}10^5\}$, averaged over 24 runs per horizon. The regret of GBD and of the algorithms using the $h = p$ surrogate grows linearly in T, by a factor of about 10 from $T = 10^4$ to 10^5, whereas that of KL−PBD and KL−PBD+ grows by a factor between 1.6 and 4.4. SALMUT−SF lies in between, with a factor between 3.4 and 6.8.

H.3 SERVICE RATE AND QUEUE-UNAWARE ALGORITHMS

We rerun the Qwen3-4B MedMCQA instance with service rates $\mu \in \{.2, .35, .5, .7, 1\}$, $T = 5 \cdot 10^4$ and 24 runs, with FCT and FQT retuned for each μ (Figure 14 and Table 7). At $\mu = 1$ the queue stays empty and $M_t(0) = c_{\exp}$. The regret of queue-aware and queue-unaware algorithms then coincides pairwise, KL−PBD+ with KL−UCB+−4P (41.9) and NBD with FCT and FQT (2,511), while KL−PBD and KL−UCB−4P differ in regret only due to the difference in their confidence budgets (113 against 158). For $\mu < 1$, KL−UCB−4P and KL−UCB+−4P keep deferring regardless of the queue: at $\mu = .2$ it is full in 41 to 56% of the rounds, and their regret exceeds 200 times that of KL−PBD+. The regret of KL−PBD and KL−PBD+ varies by less than a factor 2 across μ. The regret of NBD grows with μ, since a faster expert makes deferral optimal at more levels, which the surrogate $h = p$ misses. FCT and FQT follow the same trend up to $\mu = .7$, with an additional cost from their restricted form, which vanishes at $\mu = 1$: the optimal rule under $h = p$ is then a fixed confidence cutoff, which both can represent, so they recover the decisions of NBD.

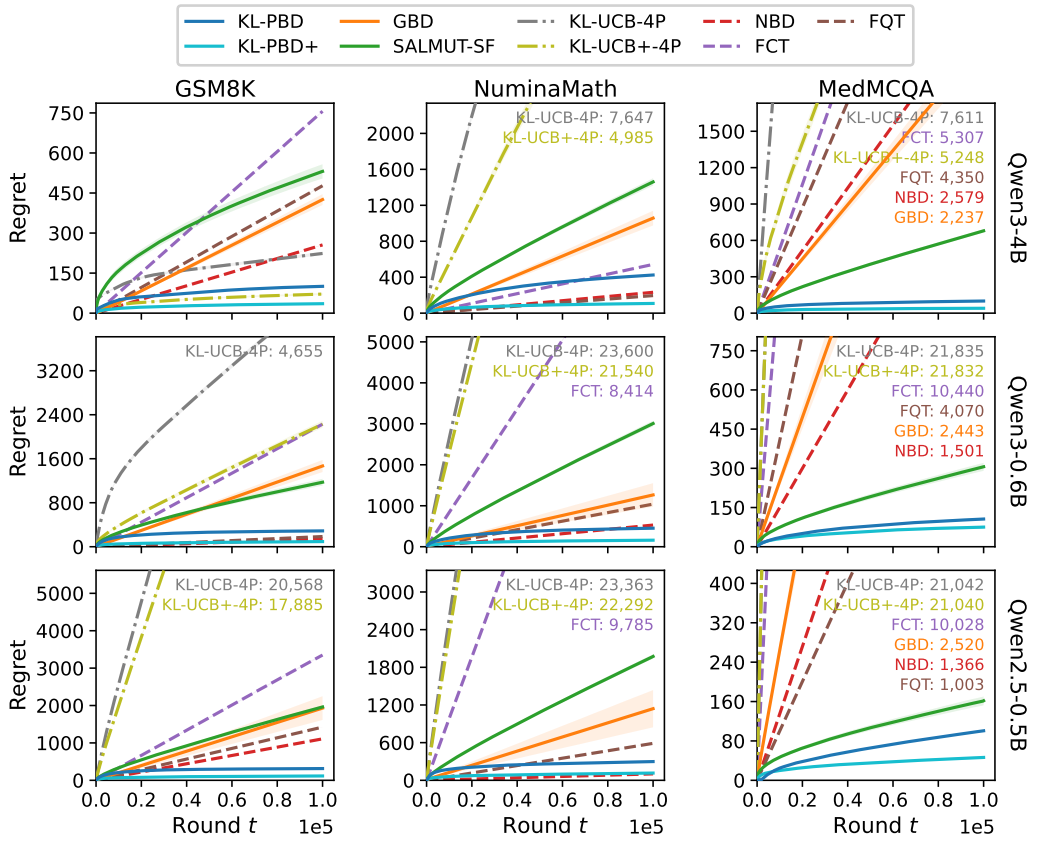


Figure 12: Regret of all algorithms, $T = 10^5$, 100 runs. Truncated curves are listed with their final regret.

Table 6: Final regret of the horizon sweep, mean ± 95% confidence half-width over 24 runs.

Model	Dataset	T	KL-PBD	KL-PBD+	GBD	SALMUT-SF	KL-UCB-4P	KL-UCB+-4P	NBD	FCT	FQT
Qwen3-4B	GSM8K	10k	43.4 ± 3.0	17.1 ± 1.8	39.9 ± 3.8	152.8 ± 5.5	97.6 ± 5.6	35.0 ± 3.9	26.4 ± 0.8	76.0 ± 1.6	48.9 ± 0.9
		50k	77.5 ± 4.8	28.7 ± 3.4	206.3 ± 20.2	357.4 ± 20.8	168.8 ± 10.8	60.4 ± 5.0	128.9 ± 1.9	379.2 ± 3.7	240.8 ± 2.9
		100k	98.2 ± 6.9	34.2 ± 4.3	414.6 ± 40.4	513.6 ± 33.7	221.6 ± 10.9	79.1 ± 6.7	255.7 ± 2.8	754.8 ± 4.1	478.5 ± 3.9
		200k	110.8 ± 9.6	38.3 ± 5.2	831.4 ± 81.0	746.8 ± 52.9	312.0 ± 8.9	99.4 ± 10.0	510.0 ± 3.3	1510.5 ± 5.7	954.4 ± 5.2
Qwen3-4B	NuminaMath	10k	130.4 ± 4.3	48.8 ± 3.5	108.0 ± 13.6	250.6 ± 18.7	1201.9 ± 25.1	560.0 ± 38.4	22.2 ± 0.8	53.7 ± 0.8	19.3 ± 0.4
		50k	314.2 ± 9.8	88.6 ± 7.5	530.3 ± 69.4	813.0 ± 63.7	4416.2 ± 167.6	2504.3 ± 183.6	115.2 ± 1.2	270.2 ± 1.9	97.0 ± 0.8
		100k	420.9 ± 14.4	105.2 ± 9.5	1058.1 ± 138.5	1423.0 ± 112.6	7610.0 ± 187.2	4847.4 ± 311.0	231.9 ± 2.1	541.8 ± 2.6	195.3 ± 1.2
		200k	545.1 ± 17.4	120.2 ± 12.2	2113.5 ± 278.3	2554.9 ± 210.0	13963.5 ± 193.3	9537.6 ± 470.1	465.0 ± 3.1	1085.5 ± 2.8	391.6 ± 1.8
Qwen3-4B	MedMCQA	10k	52.1 ± 3.6	27.8 ± 3.7	233.5 ± 23.9	144.4 ± 17.1	2512.0 ± 96.9	1000.8 ± 137.4	257.9 ± 2.9	531.2 ± 4.1	434.8 ± 3.5
		50k	83.5 ± 5.4	38.8 ± 4.5	1180.6 ± 123.7	407.1 ± 20.3	5484.8 ± 371.2	3073.6 ± 185.5	1290.9 ± 5.0	2656.6 ± 8.3	2176.5 ± 9.0
		100k	100.2 ± 5.6	43.5 ± 5.5	2360.2 ± 247.8	680.5 ± 20.9	7911.0 ± 395.9	5541.9 ± 205.0	2579.9 ± 8.1	5311.9 ± 13.6	4352.6 ± 13.7
		200k	115.2 ± 7.5	46.8 ± 6.4	4717.8 ± 492.6	1190.8 ± 20.8	12638.8 ± 424.3	10127.5 ± 206.9	5155.7 ± 11.9	10622.1 ± 16.4	8703.8 ± 18.7
Qwen3-0.6B	GSM8K	10k	148.9 ± 4.4	51.6 ± 4.1	161.7 ± 22.6	235.9 ± 16.0	1272.3 ± 56.5	367.5 ± 32.6	10.8 ± 1.7	225.2 ± 2.8	18.6 ± 0.5
		50k	243.6 ± 9.7	77.3 ± 5.2	796.3 ± 117.2	645.9 ± 50.8	2951.3 ± 93.1	1217.7 ± 86.3	68.5 ± 4.7	1115.9 ± 5.6	93.0 ± 1.6
		100k	281.0 ± 13.3	92.2 ± 6.4	1583.5 ± 235.0	1053.5 ± 81.6	4689.9 ± 144.5	2174.1 ± 138.4	144.4 ± 6.2	2227.4 ± 5.5	184.7 ± 1.8
		200k	317.6 ± 13.1	113.6 ± 8.6	3153.0 ± 471.6	1755.3 ± 130.5	7439.0 ± 290.3	4039.5 ± 220.3	297.8 ± 6.7	4453.4 ± 9.6	369.6 ± 3.1
Qwen3-0.6B	NuminaMath	10k	188.2 ± 5.3	80.1 ± 4.5	140.7 ± 56.9	429.8 ± 17.2	2638.2 ± 21.5	2297.8 ± 31.9	54.4 ± 1.4	840.0 ± 19.2	104.1 ± 1.5
		50k	372.5 ± 10.8	130.1 ± 8.3	657.6 ± 284.0	1600.3 ± 75.5	12100.9 ± 62.0	10916.9 ± 95.9	265.6 ± 2.5	4212.3 ± 49.2	518.8 ± 3.4
		100k	451.3 ± 14.7	154.9 ± 9.5	1302.7 ± 568.8	2901.5 ± 121.0	23607.2 ± 76.8	21467.3 ± 128.5	529.2 ± 2.5	8430.3 ± 72.0	1039.3 ± 4.3
		200k	533.0 ± 16.7	179.5 ± 8.9	2583.8 ± 1132.9	5352.0 ± 140.2	45998.7 ± 119.2	42326.4 ± 161.1	1052.1 ± 3.6	16824.7 ± 83.6	2075.5 ± 4.8
Qwen3-0.6B	MedMCQA	10k	33.0 ± 2.2	31.8 ± 3.1	254.1 ± 52.3	75.7 ± 5.8	2177.1 ± 14.4	2176.2 ± 14.5	151.1 ± 2.2	1042.4 ± 3.3	407.0 ± 3.8
		50k	77.3 ± 7.8	58.6 ± 5.7	1251.6 ± 263.9	194.0 ± 23.8	10903.5 ± 32.8	10902.6 ± 33.0	746.2 ± 5.1	5214.8 ± 7.0	2028.9 ± 8.6
		100k	100.0 ± 7.3	71.9 ± 7.9	2500.1 ± 527.2	310.2 ± 37.4	21829.0 ± 43.5	21828.1 ± 43.8	1499.8 ± 6.2	10436.3 ± 10.7	4068.7 ± 11.0
		200k	137.5 ± 8.0	88.3 ± 9.6	4986.9 ± 1060.1	511.4 ± 60.3	43671.7 ± 54.5	43670.7 ± 54.7	3004.1 ± 9.2	20877.6 ± 15.3	8144.1 ± 14.6
Qwen2.5-0.5B	GSM8K	10k	179.5 ± 4.3	72.3 ± 3.8	181.7 ± 55.0	339.6 ± 25.5	2576.5 ± 21.3	2008.0 ± 37.7	111.4 ± 1.6	334.3 ± 13.2	141.2 ± 2.3
		50k	288.3 ± 13.4	108.2 ± 9.1	867.6 ± 270.1	1113.6 ± 70.8	11065.3 ± 71.7	9117.7 ± 159.8	552.6 ± 4.0	1660.7 ± 32.3	708.4 ± 4.6
		100k	324.3 ± 15.4	122.4 ± 10.3	1727.4 ± 539.4	1949.2 ± 104.2	20651.6 ± 189.5	17814.5 ± 290.7	1108.4 ± 6.0	3339.7 ± 40.0	1422.7 ± 5.7
		200k	351.4 ± 16.3	137.1 ± 11.0	3442.9 ± 1078.1	3515.2 ± 144.5	38398.9 ± 281.5	35172.1 ± 529.0	2218.8 ± 7.5	6670.6 ± 57.6	2850.4 ± 7.4
Qwen2.5-0.5B	NuminaMath	10k	156.4 ± 6.3	62.3 ± 3.5	111.8 ± 34.8	292.8 ± 11.0	2574.9 ± 18.1	2321.4 ± 20.7	10.2 ± 0.2	992.4 ± 9.1	58.6 ± 1.1
		50k	251.7 ± 9.8	96.0 ± 4.8	497.9 ± 178.8	1081.2 ± 33.7	11909.0 ± 44.4	11218.1 ± 64.4	51.2 ± 0.4	4919.7 ± 37.2	293.1 ± 2.9
		100k	301.2 ± 9.5	115.3 ± 6.9	983.2 ± 357.0	1956.1 ± 47.5	23287.7 ± 72.8	22275.3 ± 117.0	102.3 ± 0.5	9813.7 ± 47.3	589.1 ± 4.1
		200k	363.9 ± 10.4	138.4 ± 9.6	1954.2 ± 715.0	3648.0 ± 66.5	45885.3 ± 140.9	44312.7 ± 181.2	204.3 ± 0.8	19570.7 ± 60.9	1177.4 ± 5.0
Qwen2.5-0.5B	MedMCQA	10k	23.9 ± 2.3	21.0 ± 3.2	256.9 ± 34.9	51.9 ± 4.3	2096.5 ± 13.5	2096.1 ± 13.5	133.7 ± 1.4	1002.7 ± 4.0	99.8 ± 0.9
		50k	65.2 ± 3.2	37.5 ± 4.9	1230.6 ± 176.7	119.6 ± 11.9	10502.6 ± 32.6	10501.6 ± 32.8	679.3 ± 2.8	5012.3 ± 10.4	500.6 ± 2.1
		100k	104.3 ± 3.6	49.5 ± 5.8	2444.1 ± 351.8	178.9 ± 18.1	21033.0 ± 41.9	21031.6 ± 42.3	1362.7 ± 3.8	10027.7 ± 15.4	1001.6 ± 3.4
		200k	160.5 ± 6.2	63.2 ± 6.9	4870.1 ± 706.8	275.3 ± 27.7	42084.4 ± 54.6	42082.6 ± 55.1	2732.4 ± 6.3	20040.8 ± 19.8	2008.3 ± 6.4

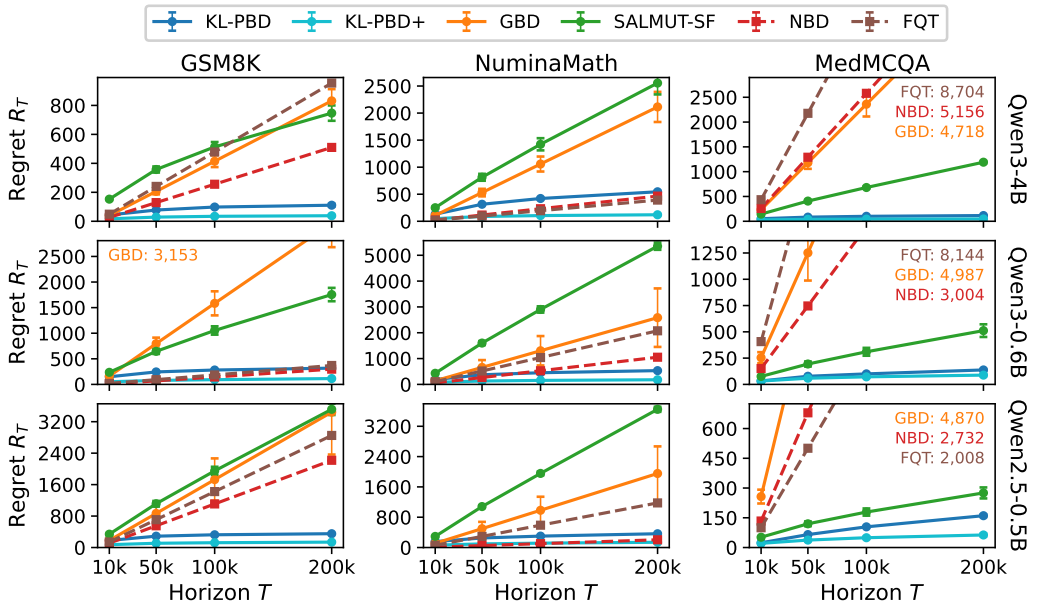


Figure 13: Final regret versus horizon T, 24 runs per point. Truncated curves are listed with their final regret.

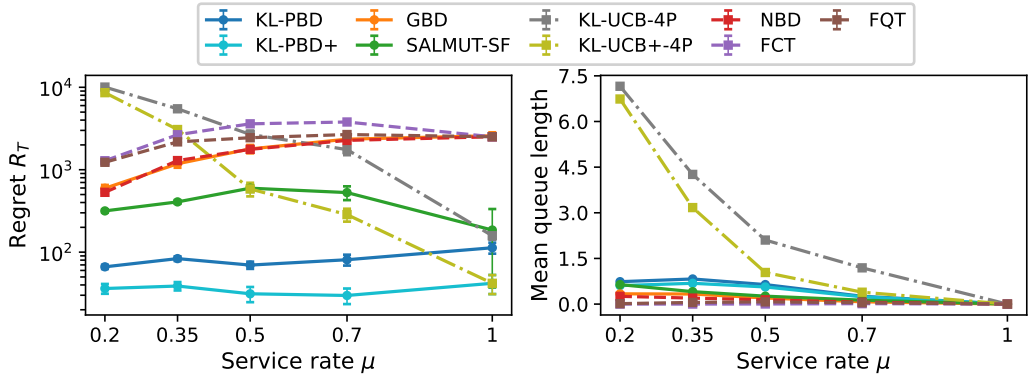


Figure 14: Service-rate sweep, $T = 5 \cdot 10^4$, 24 runs: final regret (left) and mean queue length (right).

Table 7: Final regret of the service-rate sweep, mean $\pm$ 95% confidence half-width over 24 runs.

Algorithm	$\mu = .2$	$\mu = .35$	$\mu = .5$	$\mu = .7$	$\mu = 1$
KL-PBD	66.4 $\pm$ 3.9	83.5 $\pm$ 5.4	69.7 $\pm$ 7.0	80.9 $\pm$ 12.2	113.0 $\pm$ 17.2
KL-PBD+	36.2 $\pm$ 4.9	38.8 $\pm$ 4.5	31.2 $\pm$ 6.5	29.8 $\pm$ 6.4	41.9 $\pm$ 10.9
GBD	595.6 $\pm$ 57.9	1180.6 $\pm$ 123.7	1788.5 $\pm$ 207.5	2347.8 $\pm$ 249.4	2584.3 $\pm$ 259.7
SALMUT-SF	316.4 $\pm$ 13.8	407.1 $\pm$ 20.3	597.1 $\pm$ 58.6	528.1 $\pm$ 100.4	185.4 $\pm$ 148.1
KL-UCB-4P	10060.2 $\pm$ 228.5	5484.8 $\pm$ 371.2	2679.8 $\pm$ 302.3	1754.2 $\pm$ 252.2	158.2 $\pm$ 16.9
KL-UCB+-4P	8605.3 $\pm$ 110.2	3073.6 $\pm$ 185.5	584.5 $\pm$ 109.6	285.3 $\pm$ 50.8	41.9 $\pm$ 10.9
NBD	531.8 $\pm$ 3.7	1290.9 $\pm$ 5.0	1767.2 $\pm$ 5.5	2268.8 $\pm$ 6.6	2510.5 $\pm$ 7.2
FCT	1285.3 $\pm$ 4.7	2656.6 $\pm$ 8.3	3604.9 $\pm$ 11.1	3793.1 $\pm$ 13.9	2510.6 $\pm$ 7.2
FQT	1230.9 $\pm$ 4.7	2176.5 $\pm$ 9.0	2444.1 $\pm$ 8.7	2671.8 $\pm$ 9.1	2510.6 $\pm$ 7.2

H.4 Learned threshold and decisions

The top row of Figure 15 shows the threshold and decisions learned by KL−PBD for Qwen3-0.6B at $t = T/2$, from one run for each μ, in the format of Figure 4: the threshold $\widehat{\tau}_t(b) = 1 - M_t^{h_t}(b)/c_e$ (top) and the decisions $\mathbf{1}\{\underline{h}_{j,t} < \widehat{\tau}_t(b)\}$ for $\mu = 0.35$ (bottom). The learned thresholds are within 0.09 of the oracle ones for every μ. The learned decisions differ from the oracle's due to extra deferrals caused by KL−PBD's pessimism. For example, at $b = 2$ on MedMCQA, KL−PBD defers queries with scores up to 30% whereas the oracle accepts those with scores higher than 5%. The difference is larger for Qwen3-4B (middle and bottom rows of Figure 15). Since this LLM is more accurate, it is deferred less, so by $t = T/2$ KL−PBD has collected fewer labels (2,649 on GSM8K against 9,901 for Qwen3-0.6B). Its lower confidence bounds are consequently looser, and it keeps deferring levels at intermediate queue sizes where the oracle no longer does ($b = 3$ to 6 on GSM8K, $b = 2$ to 5 on MedMCQA). On GSM8K, the levels still deferred at $b = 7$ are those with at most 20 labels, which produces the isolated cells in the bottom row.

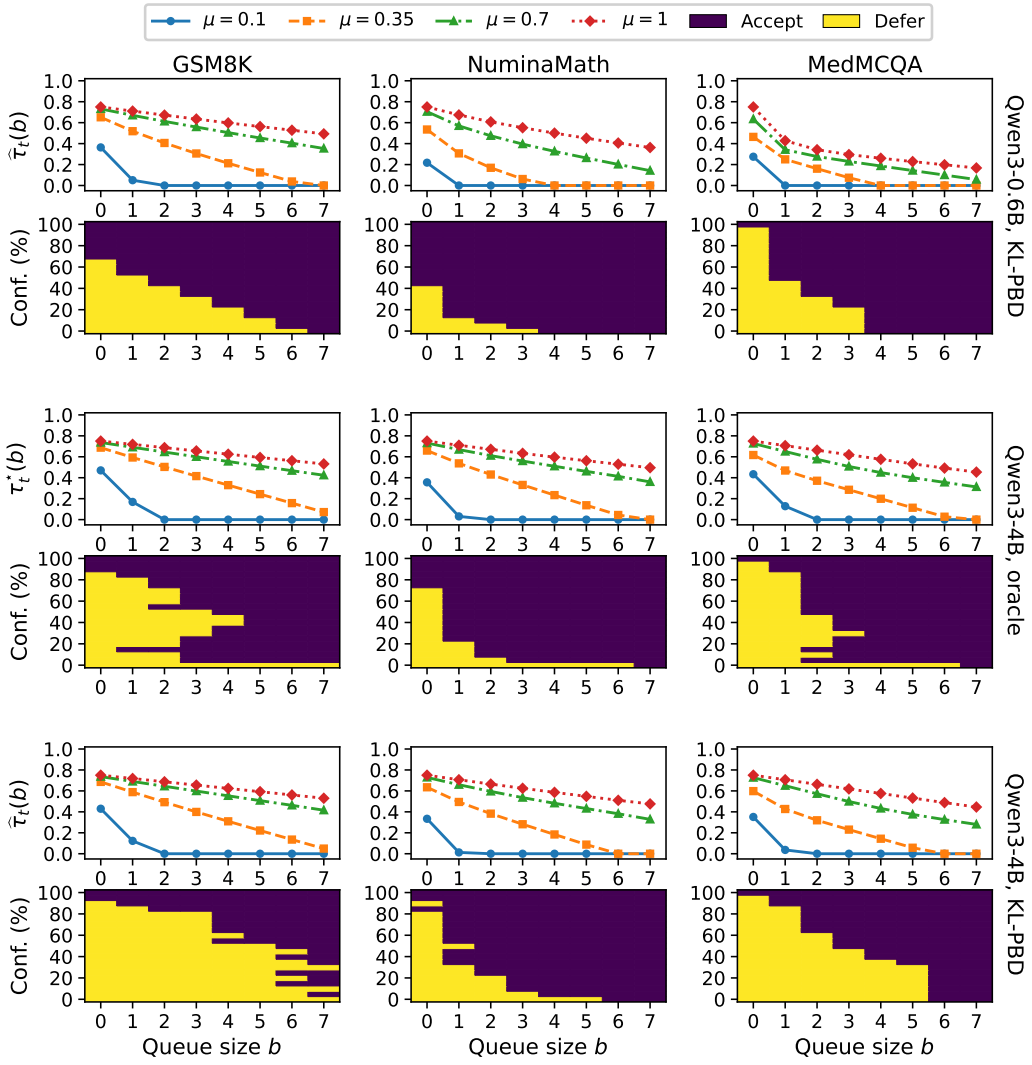


Figure 15: Thresholds for varying μ and decisions for $\mu = 0.35$ at $t = T/2$: KL−PBD for Qwen3-0.6B (top), oracle (middle) and KL−PBD (bottom) for Qwen3-4B, one run per μ.

H.5 THE IMPORTANCE OF PESSIMISM

GBD uses the same decision rule as KL−PBD, except that it uses the empirical mean $\widehat{h}_{j,N_{j,t}}$ instead of the lower confidence bound $\underline{h}_{j,t}$. Figure 16 shows, for every instance, the final regret of each of the 100 runs against the run's average deferral rate. For GBD the two are strongly negatively correlated, with correlation $\rho \in [-.95, -.85]$ across the nine instances: runs that defer less accumulate more regret. This is the selective-feedback failure of greedy learning: an early optimistic estimate of the conditional accuracy of a level stops its deferrals, hence no more feedback for this level is received, and the estimate is never corrected. On Qwen2.5-0.5B NuminaMath, for instance, the deferral rate of GBD ranges from .15 to .31 and its regret from 11 to 10,658. KL−PBD+, which keeps deferring a level while its lower confidence bound is below the current threshold, shows neither the spread nor the correlation.

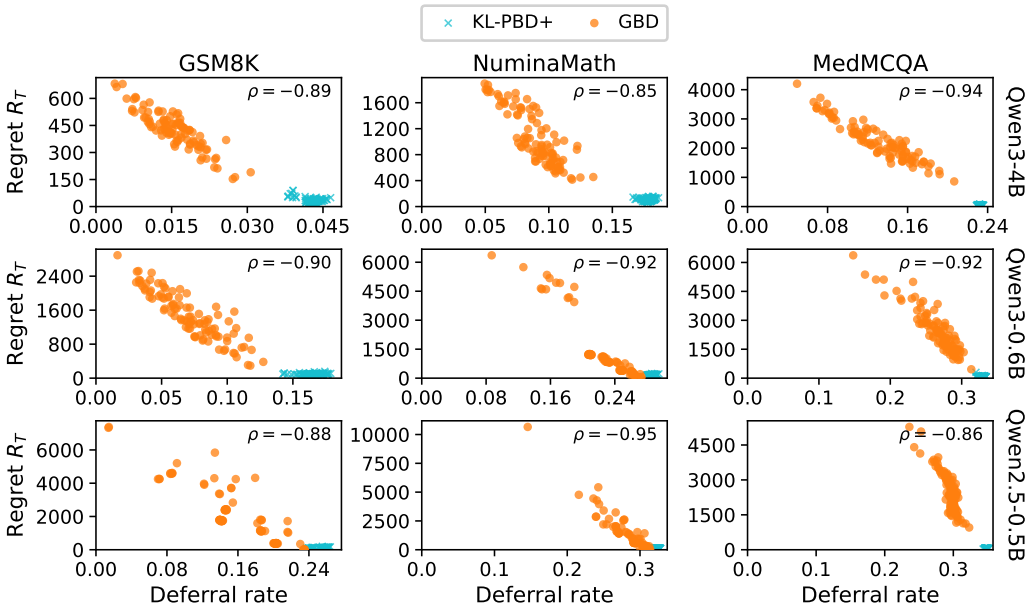


Figure 16: Final regret against average deferral rate, one point per run, $T = 10^5$, and ρ the correlation for GBD.