DELEGATE-UCB: REGRET-AUDIT TRADEOFFS IN ONLINE HUMAN-AI DELEGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Organizations adopting AI must learn which tasks to automate, yet the evidence needed to make that decision can itself be costly. We study online delegation in established human–AI teams, where human assignments provide routine performance feedback but independent assessments of AI outputs require paid audits. Notably, an audit can evaluate an AI candidate even when a human performs the task, allowing the organization to learn about future automation without changing current execution. We formulate this problem as a contextual bandit with fixed human and AI capabilities and prove that every no-audit policy suffers linear worst-case regret. We propose DELEGATE-UCB, which combines optimistic allocation with uncertainty-driven audits and targets human-route audits using overlapping human–AI confidence intervals. Under a noisy linear-residual model over T rounds, it achieves $\widetilde{O}(T^{2/3})$ expected total regret, including audit costs, matching the minimax horizon dependence asymptotically for fixed problem parameters. A known positive lower bound on the machine's suboptimality gap improves the upper bound to $\widetilde{O}(\sqrt{T})$. With a fixed organizational tolerance ϵ, the same algorithm instead achieves polylogarithmic ϵ-insensitive regret, with explicit dependence on tolerance and audit cost. Both refinements are obtained by adjusting the audit threshold to the required precision. A two-context construction further demonstrates the value of audit placement: human-route audits acquire information early enough to prevent subsequent routing errors. We examine these audit settings through controlled experiments, showing how audit placement and precision affect the balance between routing regret and review cost. We further evaluate DELEGATE-UCB in a matched-horizon ImageNet16H replay and show that DELEGATE-UCB can reduce total regret relative to other benchmark auditing strategies.

1 INTRODUCTION

Many recurring organizational workflows, from triage and screening to exception handling and escalation, now involve both human specialists and machine agents. This is especially salient with frontier language models, which can act as low-cost generalists but whose reliability varies substantially across task types, domains, and failure modes. In these organizational settings, delegation is rarely a one-time choice about whether to automate, but rather a sequential learning problem. Tasks (with observable context x_t) arrive sequentially, and different agents have comparative advantage on different regions of the task space. Although the machine enters with pretrained capabilities, the organization must learn how those capabilities translate to its own tasks, and how the machine compares with the expertise and cost of its human specialists.

Consider an organization introducing a pretrained machine into an established team. Its employees already operate within routine performance-evaluation procedures. We study workflows in which a human assignment yields a noisy assessment of that employee's deliverable, whereas machine execution produces an output but no independent assessment of its quality. A supervisor's rating of an employee's report, for example, evaluates that report, not an AI-generated alternative. Assessing the latter requires a separate, paid review.

Correspondingly, we use *audit* to refer to a costly, standardized form of selective review: the organization pays for a trusted, high-quality assessment that can be converted into residual feedback for machine calibration. The organization chooses whether to purchase this feedback as part of its

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.

task-allocation decision. An audit can evaluate the machine candidate even when a human completes the task, providing a shadow evaluation of the unused alternative without changing the current handler or its payoff. Routing determines who acts now, and auditing determines what the organization learns about future AI use. The objective therefore combines routing loss relative to the context-wise optimal handler with the audit charge c_R.

In workflows such as medical triage or legal compliance, audit often requires scarce domain experts, creating substantial labor costs and latency bottlenecks due to bounded human attention. The organization therefore cannot audit passively or uniformly, and audit must be treated as a limited investment budget. The learner must decide, online, which cases are worth auditing, even though the value of an audit depends on an unknown human–machine delegation frontier.

This entanglement distinguishes our setting from standard learning-to-defer (L2D) and model-routing formulations, where the central task is typically a terminal decision of who should produce the final prediction. In contrast, in our setting, the learner makes a composite decision: $a_t = (j_t, q_t)$, where j_t is the primary handler (i.e., who performs the task) and $q_t \in \{0, 1\}$ is the audit decision.

Contributions. *Feedback model and necessity.* We formulate this organizational problem through a contextual bandit model.Human agents are modeled as imperfect specialists with unknown but fixed context/task-dependent competence. For horizon T, any no-audit policy suffers linear regret on one of two observationally indistinguishable instances. With Gaussian audit noise of scale $\sigma > 0$, we further bound the routing loss incurred under a limited audit budget. Balancing this identification error with audit expenditure yields the interior $\Omega((c_R\sigma^2)^{1/3}T^{2/3})$ minimax lower bound.

Audit precision and regret. We propose DELEGATE-UCB, which routes tasks optimistically and audits machine-routed tasks when residual uncertainty is at least a chosen threshold. On human-routed tasks, it uses the same threshold but audits only when the machine's confidence interval overlaps that of the selected human. The threshold sets the audit precision, while confidence overlap excludes human-route audits when the current comparison is already resolved in the human's favor. For d-dimensional contexts and H humans, DELEGATE-UCB attains expected total regret $\widetilde{O}(d\sqrt{HT} + d^{2/3}c_R^{1/3}T^{2/3} + c_R)$, where $\widetilde{O}$ suppresses logarithmic factors. For fixed d and H, the machine-calibration term matches the lower bound in T and c_R up to logarithmic factors. A known positive lower bound on the machine's suboptimality gap improves the regret upper bound to $\widetilde{O}(\sqrt{T})$. A uniform margin is nevertheless restrictive for continuous context distributions. Correspondingly, we consider a more practical setting where the organization instead chooses an indifference band ϵ, and show that DELEGATE-UCB achieves polylogarithmic ϵ-insensitive regret, including full audit costs. These refinements change only the audit threshold. The upper bounds also extend to a class of admissible threshold-based audit rules.

1.1 RELATED WORK

Learning to defer and multi-expert routing. Learning to defer (L2D) studies prediction systems in which a model either predicts directly or passes an instance to a human expert (Madras et al., 2018). Subsequent work develops consistent estimators for the single-expert setting (Mozannar & Sontag, 2020), post-hoc deferral rules for a fixed predictor (Narasimhan et al., 2022), and exact algorithms for jointly learning a classifier and a deferral rule (Mozannar et al., 2023). The multiple-expert setting introduces the additional problem of choosing which expert to consult, motivating surrogate losses and consistency guarantees for expert routing (Verma et al., 2023; Mao et al., 2023; 2025). Other extensions account for dependence between human and model predictions (Wei et al., 2024) and adapt deferral policies to previously unseen experts (Tailor et al., 2024).

Online deferral and feedback acquisition. Online formulations study delegation as tasks arrive and feedback becomes available. Reid et al. (2024) formulate decision deferral as a contextual bandit with budget constraints and partial feedback. Lykouris & Weng (2024) account for congestion and delays in human review, while Montreuil et al. (2026a) study online L2D with bandit feedback and a varying expert pool. Related work considers the acquisition of additional expert input: Pugnana et al. (2025) learn when to request human feedback for an enriched predictor, and Montreuil et al. (2026b) jointly select an expert and the additional information provided to it.

The cost of feedback is explicit in Tucker et al. (2023), who study bandits in which observing the selected arm's reward requires payment and establish a regret-cost tradeoff with $T^{2/3}$ horizon

dependence. Selective-sampling and preference-query methods also control the number of expert labels or action comparisons requested during learning (Sekhari et al., 2023b;a).

Human-AI collaboration with bandit feedback. Bandit-feedback models study how to exploit human–machine complementarity when outcomes are observed only for actions taken. Gao et al. (2021) learn collaboration policies from logged bandit feedback, allocating decisions between humans and an algorithm, including settings with multiple human decision-makers. Gao et al. (2023) study complementary policies for human–AI teams under partial outcome feedback. Cao et al. (2025) introduce HR-Bandit, which incorporates human expertise into linear recourse bandits where the learner chooses both actions and feature modifications.

Adaptive model routing. Model-routing methods select among models with different performance and cost profiles. Poon et al. (2026) study online multi-LLM selection under unstructured prompt evolution and develop a contextual-bandit algorithm with sublinear myopic regret. Wei et al. (2025) learn LLM-routing policies from bandit feedback and allow users to adjust the cost–performance tradeoff at inference time.

2 Problem Formulation

We study sequential delegation over rounds $t = 1, \dots, T$. Each task arrives with an observable context $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ and a predictable pretrained-machine score $b_0(x_t)$. The organization chooses a composite action $a_t = (j_t, q_t)$: $j_t \in \mathcal{J} = \{0, 1, \dots, H\}$ is the primary handler (machine 0 or one of H humans), and $q_t \in \{0, 1\}$ purchases a trusted audit at cost $c_R > 0$. Handler j has a nonnegative assignment cost $c_j \geq 0$, for example, representing routine evaluation on human routes. The primary handler executes the task, while audit is a separate costly information-acquisition action. An audit can assess the machine candidate whether the task is assigned to a human or to the machine. All costs are assumed finite, with $C_{\max} := \max\{c_0, \dots, c_H, c_R\} < \infty$. Routing and audit are both chosen before current task noise is observed (Figure 1).

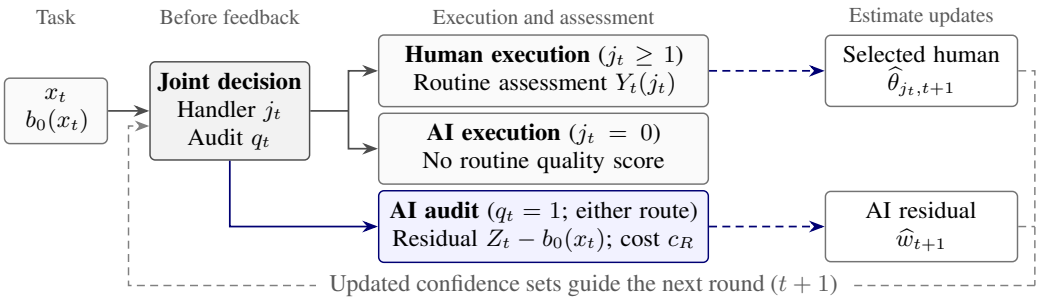


Figure 1: Delegation and feedback. Both decisions precede observation of current feedback. Routine human assessments and paid AI audits update separate estimates (dashed arrows). Under a human route, audit evaluates a separately generated AI candidate without replacing the handler.

Human $j \geq 1$ has a fixed unknown parameter $\theta_j^\star \in \mathbb{R}^d$ and mean utility $\mu_j(x) = x^\top \theta_j^\star - c_j$. When human j handles task t, the routing policy observes a routine performance assessment $Y_t(j) = x_t^\top \theta_j^\star + \eta_{j,t}$, where $\eta_{j,t}$ is assessment noise. The machine's expected output quality is $b_0(x) + x^\top w^\star$, where $w^\star \in \mathbb{R}^d$ is a fixed unknown residual parameter. Its mean utility is therefore $\mu_0(x) = b_0(x) + x^\top w^\star - c_0$. The routing policy receives a machine-quality assessment only when an audit is purchased. An audit evaluates the machine's candidate under either human or machine execution and returns a handler-independent assessment $Z_t = b_0(x_t) + x_t^\top w^\star + \xi_t$, where ξ_t captures variation in the assessed machine quality around its conditional mean. Both human and machine assessments satisfy the conditional sub-Gaussian noise condition in Assumption 1. The observed residual $Z_t - b_0(x_t)$ is used to estimate $w^\star$. The revealed history $\mathcal{F}_t$ consists of the contexts, pretrained scores, chosen actions, and quality assessments observed through round t. Any output-derived routing predictors are included in the pre-decision inputs x_t or $b_0(x_t)$.

Remark 1. We do not treat frontline human outputs as machine-calibration labels. In many organizational workflows, human outputs are produced to resolve the immediate case rather than to train a

machine. Different humans may have different expertise, quality standards, documentation habits, and levels of effort. The recorded output may vary in format, granularity, and rationale, and privacy or compliance constraints may prevent directly reusing the operational record for model updating. Similarly, an unaudited machine output is not treated as feedback for updating the machine. The machine's own prediction is not an independent assessment of its correctness. Using it as a target would amount to self-training on the model's current belief and could reinforce systematic residual errors. Hence, the machine residual statistics are updated only through audited residuals $Z_t - b_0(x_t)$.

Comparative advantage. Human and machine agents may have comparative advantages in different regions of the task space, so their relative performance need not admit a global ranking. For agents i, j, the gap $\Delta_{ij}(x) = \mu_i(x) - \mu_j(x)$ may therefore change sign with x. Audits are useful where learning the residual could change future routing, particularly near the frontier between the machine and the best human.

We make the following assumptions throughout the rest of this work.

Assumption 1 (Regularity)**.** For finite constants B_0, S, σ, the contexts, pretrained scores, and unknown parameters satisfy $\|x_t\|_2 \le 1$, $|b_0(x_t)| \le B_0$, $\|w^\star\|_2 \le S$, and $\left\|\theta_j^\star\right\|_2 \le S$ for every t and $j \ge 1$. Moreover, with $\mathcal{G}_t = \sigma(\mathcal{F}_{t-1}, x_t, b_0(x_t), j_t, q_t)$, all $\eta_{j,t}$ and ξ_t are conditionally mean-zero and σ-sub-Gaussian given $\mathcal{G}_t$. Contexts and actions may adapt to the observed history but not to current unobserved noise.

The bounded-feature and bounded-parameter conditions, together with conditionally sub-Gaussian noise, follow the stochastic linear-bandit formulation of Abbasi-Yadkori et al. (2011). The unit-norm bound can be obtained by rescaling bounded task features, and the bound on $b_0(x)$ accommodates pretrained quality scores reported on a finite scale. The noise condition includes conditionally Gaussian errors with zero mean and variance at most σ^2.

Assumption 2 (Audited-residual realizability)**.** When $q_t = 1$, $Z_t - b_0(x_t) = x_t^\top w^\star + \xi_t$. Conditional on x_t, the residual target has the same mean under every primary handler j_t.

Audits evaluate the fixed AI's candidate using the same quality criteria, whether the task is assigned to a human or to the machine. For a given task context, Assumption 2 takes the expected audit score to be the same across primary handlers.

Benchmark and regret. The oracle knows $w^\star$ and the human parameters $\theta_1^\star, \dots, \theta_H^\star$, routes to $j_t^\star \in \arg\max_j \mu_j(x_t)$, and never audits because audit has no informational value to it. We consider admissible policies π that map the revealed history $\mathcal{F}_{t-1}$ and current $(x_t, b_0(x_t))$ to a distribution over $\mathcal{J} \times \{0, 1\}$. We evaluate a policy using the handlers' mean utilities and the cost of purchased audits. For policy π on instance $\mathcal{I}$, let $N_{\mathrm{aud}}(T) = \sum_{t=1}^T q_t$ denote the number of purchased audits. Its pathwise pseudo-regret is

$$R_\pi(T) = \underbrace{\sum_{t=1}^T [\mu_{j_t^\star}(x_t) - \mu_{j_t}(x_t)]}_{R_{\pi,\mathrm{route}}(T)} + c_R N_{\mathrm{aud}}(T).$$

Additionally, we define $\mathrm{Reg}_\pi^{\mathcal{I}}(T) = \mathbb{E}_{\mathcal{I}}[R_\pi(T)]$ and $\mathrm{Reg}_{\pi,\mathrm{route}}^{\mathcal{I}}(T) = \mathbb{E}_{\mathcal{I}}[R_{\pi,\mathrm{route}}(T)]$, where expectations follow the history law induced by π on $\mathcal{I}$, and we omit the instance index when fixed.

3 Identification Requires Audit

Audits provide information about machine quality, but their cost also contributes to regret. In this section, we first show that policies that never audit incur linear worst-case regret. When audits are available, we derive a lower bound on routing regret under a limited audit budget and then incorporate audit costs to obtain a minimax lower bound on total regret.

Theorem 1 (No-audit impossibility)**.** *For every $\gamma \in (0, S]$, there exist two instances $\mathcal{I}_+(\gamma)$ and $\mathcal{I}_-(\gamma)$ with one human of known utility and one machine such that every possibly randomized, history-dependent policy π that never audits satisfies*

$$\max_{\nu \in \{-1,+1\}} \mathrm{Reg}_\pi^{\mathcal{I}_\nu(\gamma)}(T) \ge \gamma T/2.$$

Thus, knowing human utility is not enough to guarantee sublinear regret when the policy receives no feedback about machine quality.

We construct a family of paired instances that yields regret lower bounds both without audits and under a limited audit budget. For $\nu \in \{-1, +1\}$ and $\Delta \in (0, S]$, let $\mathcal{I}_\nu(\Delta)$ have $d = H = 1$, $x_t = 1$, zero execution costs and base score, audit cost $c_R > 0$, known human utility 0, and machine residual $w^\star = \nu\Delta$. Human assessments are noiseless, whereas each audit returns $\nu\Delta + \xi_t$ with $\xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Under this feedback protocol, no unaudited observation depends on ν, although the optimal handler does. This family removes estimation of human utilities and leaves only the information supplied by audits.

Under any no-audit policy, the instances $\mathcal{I}_+(\gamma)$ and $\mathcal{I}_-(\gamma)$ induce the same distribution of observed histories, even though the machine is optimal in the former and the human in the latter. The expected regrets in the two instances therefore sum to γT, so any no-audit policy incurs expected regret of at least $\gamma T/2$ on one of them.

Theorem 2 (Audit-budget lower bound). *Assume $S > 0$ and $\sigma > 0$. For an audit budget $B \in [0, T]$, set $\Delta_B = \min\{S, \sigma/(4\sqrt{B+1})\}$. Any policy satisfying $\max_\nu \mathbb{E}_{\mathcal{I}_\nu(\Delta_B)}[N_{\text{aud}}(T)] \le B$ obeys*

$$\max_\nu \operatorname{Reg}^{\mathcal{I}_\nu(\Delta_B)}_{\pi,\text{route}}(T) \ge \frac{3}{8}\Delta_B T \ge \frac{3T}{32}\min\left\{S, \frac{\sigma}{\sqrt{B+1}}\right\}.$$

A hard cap of B audits also satisfies the expected-budget condition. Only audits provide information about the sign of the machine residual, and the choice of Δ_B keeps the two instances difficult to distinguish within that budget. Appendix B.2 gives the full proof.

Corollary 1 (Cost-sensitive minimax lower bound). *For $S > 0$ and $\sigma > 0$, there is a universal $c > 0$ such that*

$$\inf_\pi \sup_{\mathcal{I}\in\mathfrak{I}_{\text{cal}}} \operatorname{Reg}^{\mathcal{I}}_\pi(T) \ge c \inf_{0\le B\le T}\left\{T\min\left(S, \frac{\sigma}{\sqrt{B+1}}\right) + c_R B\right\},$$

where $\mathfrak{I}_{\text{cal}} = \{\mathcal{I}_\nu(\Delta) : \nu \in \{-1, +1\}, 0 < \Delta \le S\}$. If the optimizing audit level is interior and the radius does not bind, namely $1 \lesssim (\sigma T/c_R)^{2/3} \lesssim T$ and $(c_R\sigma^2/T)^{1/3} \lesssim S$, the lower bound becomes $\Omega((c_R\sigma^2)^{1/3}T^{2/3})$.

Balancing $\sigma T/\sqrt{B}$ with $c_R B$ gives $B^\star \asymp (\sigma T/c_R)^{2/3}$: audits reduce identification error at the square-root rate but are purchased linearly. The hard family embeds in every $d, H \ge 1$ through one informative coordinate and dominated humans, establishing the calibration lower bound without resolving the dependence on d or H.

4 DELEGATE-UCB: ROUTING AND SELECTIVE AUDIT

DELEGATE-UCB maintains separate ridge models for human assessments $Y_s(j)$ and audited machine residuals $Z_s - b_0(x_s)$. For a ridge parameter $\lambda > 0$, write $A_\lambda := 1 + 1/\lambda$ and define the corresponding design matrices and parameter estimates by

$$\begin{aligned} V_{j,t} &= \lambda I + \sum_{s<t:j_s=j} x_s x_s^\top, & \widehat{\theta}_{j,t} &= V_{j,t}^{-1}\sum_{s<t:j_s=j} x_s Y_s(j), \\ V_t &= \lambda I + \sum_{s<t} q_s x_s x_s^\top, & \widehat{w}_t &= V_t^{-1}\sum_{s<t} q_s x_s [Z_s - b_0(x_s)]. \end{aligned} \tag{1}$$

Write $b_{j,t}$ and b_t for the human and residual response sums in (1). For $A \succ 0$, use $\|x\|_A = (x^\top A x)^{1/2}$, and set $\nu_t(x) = \|x\|_{V_t^{-1}}$. For $j \in \mathcal{J}$, let $\beta_{j,t}$ denote the confidence radius in Lemma 2 with failure probability $(T+1)^{-2}/(H+1)$, where $\beta_{0,t}$ is the radius for estimating the machine residual parameter $w^\star$. The lower and upper utility bounds are

$$\begin{aligned} U_{j,t}(x) &= x^\top\widehat{\theta}_{j,t} + \beta_{j,t}\|x\|_{V_{j,t}^{-1}} - c_j, & L_{j,t}(x) &= x^\top\widehat{\theta}_{j,t} - \beta_{j,t}\|x\|_{V_{j,t}^{-1}} - c_j, \\ U_{0,t}(x) &= b_0(x) + x^\top\widehat{w}_t + \beta_{0,t}\nu_t(x) - c_0, & L_{0,t}(x) &= b_0(x) + x^\top\widehat{w}_t - \beta_{0,t}\nu_t(x) - c_0. \end{aligned} \tag{2}$$

Bounds without an argument are evaluated at x_t. Let $h_t \in \arg\max_{j\geq 1} U_{j,t}(x_t)$ be the human with the largest upper confidence bound. Routing compares h_t with the machine:

$$j_t = 0 \text{ if } U_{0,t}(x_t) \geq U_{h_t,t}(x_t), \qquad j_t = h_t \text{ otherwise.} \tag{3}$$

With audit threshold $\tau > 0$, the audit decision is

$$q_t = \mathbf{1}\{\nu_t(x_t) \geq \tau \text{ and } (j_t = 0 \text{ or } [L_{0,t}, U_{0,t}] \cap [L_{h_t,t}, U_{h_t,t}] \neq \varnothing)\}. \tag{4}$$

Machine assignments provide no quality feedback on their own, so the policy audits them whenever $\nu_t(x_t) \geq \tau$ to update the residual estimate for subsequent routing decisions. The threshold τ controls the precision of this information by leaving contexts with smaller residual uncertainty unaudited. These audits are retained even when the machine is confidently preferred on the current task, because observations of the shared residual parameter can also inform decisions at other contexts.

On human-routed tasks, confidence overlap identifies opportunities to learn about machine quality while a human performs the current task. Since a human route implies $U_{h_t,t} > U_{0,t}$, the overlap condition reduces to $L_{h_t,t} \leq U_{0,t}$.

Routine human assessments update the selected human's estimate, while audited residuals update the machine residual estimate, as summarized in Algorithm 1.

Algorithm 1 DELEGATE-UCB with confidence-overlap audit

Require: Confidence radii $\{\beta_{j,t}\}$ and threshold $\tau > 0$
Initialize $V_{j,1} = \lambda I, b_{j,1} = 0$ $(j \geq 1)$ and $V_1 = \lambda I, b_1 = 0$
for $t = 1, \ldots, T$ **do**
  Observe $(x_t, b_0(x_t))$; compute all confidence intervals and $\nu_t(x_t)$
  $h_t \in \arg\max_{j\geq 1} U_{j,t}(x_t)$
  $j_t \leftarrow 0$ if $U_{0,t}(x_t) \geq U_{h_t,t}(x_t)$; otherwise $j_t \leftarrow h_t$
  $o_t \leftarrow \mathbf{1}\{[L_{0,t}, U_{0,t}] \cap [L_{h_t,t}, U_{h_t,t}] \neq \varnothing\}$
  $q_t \leftarrow \mathbf{1}\{\nu_t(x_t) \geq \tau \text{ and } (j_t = 0 \text{ or } o_t = 1)\}$
  Execute (j_t, q_t) and carry all human statistics forward
  If $j_t \geq 1$, observe $Y_t(j_t)$ and update
  $(V_{j_t,t+1}, b_{j_t,t+1}) \leftarrow (V_{j_t,t} + x_t x_t^\top, b_{j_t,t} + x_t Y_t(j_t))$
  Carry (V_t, b_t) forward; if $q_t = 1$, observe Z_t and update
  $(V_{t+1}, b_{t+1}) \leftarrow (V_t + x_t x_t^\top, b_t + x_t[Z_t - b_0(x_t)])$
end for

4.1 THEORETICAL ANALYSES

We define an audit rule as τ-admissible if it never audits below τ and audits every machine-routed context with $\nu_t(x_t) \geq \tau$. The audit rule in Equation (4) is τ-admissible.

Let $L_T = \log(1+T/(\lambda d))$, $\Lambda_T = A_\lambda d L_T$, $\delta_T = (T+1)^{-2}$, $\beta_T = \sigma\sqrt{dL_T + 2\log((H+1)/\delta_T)} + \sqrt{\lambda}S$, and $M = 2(B_0 + S + C_{\max}) + c_R$. Here β_T bounds all model radii through round T. We use $R_\tau(T)$ and $\mathrm{Reg}_\tau(T)$ to denote the pathwise and expected total regret of DELEGATE-UCB with threshold τ. The complete proof and the high-probability bounds are included in Appendix B.

Proposition 1 (Audit count). *For $\tau > 0$, under the context bound in Assumption 1, every τ-admissible rule satisfies*

$$N_{\mathrm{aud}}(T) \leq \min\left\{T, \frac{A_\lambda d}{\tau^2}\log\left(1 + \frac{T}{\lambda d}\right)\right\}.$$

Each audit has $\nu_t(x_t) \geq \tau$ and updates the residual design, so elliptical potential gives $N_{\mathrm{aud}}(T)\tau^2 \leq \sum_t q_t \nu_t(x_t)^2 \leq \Lambda_T$. Repeated audits shrink uncertainty in measured directions below the threshold.

On human-routed rounds, confidence overlap bounds the true utility gap between the machine and the best human. Let $\mu_H^\star(x) = \max_{j\geq 1} \mu_j(x)$ be the best human utility at context x. On an audited human-routed round, write $s_{0,t} = \beta_{0,t}\nu_t(x_t)$ and $s_{h_t,t} = \beta_{h_t,t}\|x_t\|_{V_{h_t,t}^{-1}}$. On the simultaneous confidence event, Lemma 5 in Appendix B.3 shows

$$\mu_H^\star(x_t) - \mu_0(x_t) \leq 2s_{0,t} + 2s_{h_t,t},$$
$$\mu_0(x_t) - \mu_H^\star(x_t) \leq 2s_{h_t,t}.$$

These bounds show that human-route audits are confined to contexts where the utility difference between the machine and the best human is controlled by the current confidence widths. The overlap condition skips a human-route audit whenever the selected human's lower confidence bound exceeds the machine's upper bound.

Optimistic routing controls loss through the chosen handler's confidence radius (Lemma 4 in Appendix B.3). Human assignments supply their own observations, so the accumulated estimation loss can be bounded along the sequence of updates to each human's estimate. All paid audits, including those on human-routed tasks, update the shared residual design used to control the loss on audited machine assignments. For a τ-admissible rule, unaudited machine assignments instead satisfy $\nu_t(x_t) < \tau$, so their cumulative routing loss is at most $2\beta_T T\tau$ on the confidence event. Setting $\tau_B := \sqrt{\Lambda_T/B}$ enforces a hard audit cap B and makes unaudited residual uncertainty of order $B^{-1/2}$.

Theorem 3 (Budgeted regret)**.** *Under Assumptions 1 and 2, run* DELEGATE-UCB *with target $B \in [1, T]$ and threshold τ_B. Then $N_{\mathrm{aud}}(T) \le B$ and*

$$\mathrm{Reg}_{\tau_B}(T) \le \widetilde{O}\left(d\sqrt{HT} + \frac{dT}{\sqrt{B}}\right) + c_R B + \frac{M}{T}.$$

The first term is the cost of learning human performance. The residual term $\widetilde{O}(dT/\sqrt{B})$ captures the routing loss left by limited machine feedback, whereas $c_R B$ bounds the cost of acquiring that feedback. Optimizing this tradeoff gives the following bound.

Corollary 2 (Optimized budget)**.** *Choosing $B_T^\star = \min\{T, \max\{1, \lceil(\beta_T^2 A_\lambda d L_T T^2/c_R^2)^{1/3}\rceil\}\}$ gives*

$$\mathrm{Reg}_{\tau_{B_T^\star}}(T) \le \widetilde{O}\left(d\sqrt{HT} + d^{2/3} c_R^{1/3} T^{2/3} + c_R\right) + \frac{M}{T}.$$

For fixed d, H, when neither optimizer is clipped and the lower-bound residual-radius constraint does not bind, this matches the calibration lower bound in T and c_R up to logarithmic factors.

4.2 KNOWN MARGINS AND ORGANIZATIONAL TOLERANCE

The worst-case rate reflects the need to distinguish arbitrarily small utility differences. We consider two distinct ways to limit this burden: a known margin bounds the utility gap away from zero whenever the machine is suboptimal, whereas an organizational tolerance makes near-optimal assignments acceptable.

A known lower bound on the machine's suboptimality gap changes the required audit resolution.

Theorem 4 (Known-margin regret)**.** *Under Assumptions 1–2, suppose a known $\underline{\Delta} > 0$ satisfies $\mu_H^\star(x) - \mu_0(x) \ge \underline{\Delta}$ whenever $\mu_0(x) < \mu_H^\star(x)$, for every admissible context x. For $\beta_T > 0$, run* DELEGATE-UCB *with $\tau_\Delta = \underline{\Delta}/(2\beta_T)$. Then $N_{\mathrm{aud}}(T) \le \min\{T, 4\beta_T^2 \Lambda_T/\underline{\Delta}^2\}$ and*

$$\mathrm{Reg}_{\tau_\Delta}(T) \le \widetilde{O}\left(d\sqrt{HT} + \frac{d^2}{\underline{\Delta}} + \frac{c_R d^2}{\underline{\Delta}^2}\right) + \frac{M}{T}.$$

This one-sided condition separates suboptimal machine assignments from the optimum, while allowing ties and arbitrarily small machine advantages. Let $g_t = \max_{j \in \mathcal{J}} \mu_j(x_t) - \mu_{j_t}(x_t)$. On the confidence event, unaudited machine routes have $g_t < \underline{\Delta}$ and are therefore optimal. For fixed problem parameters and a fixed positive margin, expected machine-routing regret and audit expenditure have polylogarithmic upper bounds in T. Human-route gaps remain unrestricted, so the total-regret bound retains a $\widetilde{O}(\sqrt{T})$ term, even when $H = 1$.

When utilities vary continuously on a connected context support, a change in the sign of $\mu_0(x) - \mu_H^\star(x)$ creates arbitrarily small positive gaps on tasks where the machine is suboptimal. An organization can instead choose an indifference band $\epsilon > 0$ to specify which utility differences warrant further assessment, without requiring a positive margin in the environment. Define its ϵ-insensitive total regret by

$$R_\pi^{(\epsilon)}(T) = \sum_{t=1}^{T} [g_t - \epsilon]_+ + c_R N_{\mathrm{aud}}(T),$$

where $[z]_+ = \max\{z, 0\}$. The routing term follows the excess-loss form of lenient regret (Merlis & Mannor, 2021), assigning zero loss to handlers within ϵ of the best and charging only the excess otherwise.

Theorem 5 (Organizational tolerance)**.** *Under Assumptions 1–2, for any $\epsilon > 0$ and $\beta_T > 0$, run* DELEGATE-UCB *with $\tau_\epsilon = \epsilon/(2\beta_T)$. Its expected ϵ-insensitive total regret satisfies*

$$\mathrm{Reg}^{(\epsilon)}_{\tau_\epsilon}(T) := \mathbb{E}[R^{(\epsilon)}_{\tau_\epsilon}(T)] \le \widetilde{O}\left(\frac{(H+1)d^2}{\epsilon} + \frac{c_R d^2}{\epsilon^2}\right) + \frac{M}{T}.$$

Unlike the environmental margin in Theorem 4, ϵ is an organizational choice: the acceptable utility shortfall on each task. A larger tolerance raises the audit threshold and lowers the audit-cost bound, $\widetilde{O}(c_R d^2/\epsilon^2)$. For fixed ϵ and other problem parameters, the expected cumulative loss beyond the tolerance and expected audit expenditure have polylogarithmic upper bounds in T. The corresponding ordinary-regret bound is $\mathrm{Reg}_{\tau_\epsilon}(T) \le \epsilon T + \mathrm{Reg}^{(\epsilon)}_{\tau_\epsilon}(T)$, which makes the allowance associated with the organization's chosen tolerance explicit. Appendix B.4 provides the proofs and shows how a suitable horizon-dependent choice of ϵ recovers the gap-free rate.

5 EMPIRICAL EVALUATION

In this section, we empirically examine how audit placement and precision affect routing regret and audit expenditure. Full protocols and paired-seed results are provided in Appendices C-D. The audit-rule comparisons share the optimistic router: MACHINE-ONLY audits machine-routed tasks at or above the uncertainty threshold, whereas UNCERTAINTY-ONLY audits either route at or above that threshold. We include SYMMETRIC-OVERLAP as an experimental control to examine the effect of requiring confidence overlap on machine-routed tasks. This control audits a task exactly when $\nu_t(x_t) \ge \tau$ and the confidence intervals of the machine and h_t overlap, regardless of which handler executes the task.

5.1 CONTROLLED MECHANISM STUDY

To compare audit placement at matched audit counts, we use a $T = 1000$ controlled pool with LLM-generated contexts and base scores, linear machine residuals, and two human utilities. Twenty paired seeds share arrivals and noises, and all policies use empirical exploration widths. For each audit cap $B \in \{100, 200, 400\}$ and seed, we construct two post-hoc controls that match DELEGATE-UCB's realized audit count N. MATCHED RANDOM audits N rounds chosen uniformly without replacement, whereas FRONT-LOADED audits the first N rounds. DELEGATE-UCB has lower total regret than MATCHED RANDOM at all three caps and FRONT-LOADED at $B = 100, 200$.

When random auditing uses the full budget B, its mean total regret is lower than DELEGATE-UCB's at $B = 100$ but higher at $B = 200, 400$ (Appendix C.9).

To examine audit precision, we use a separate two-context experiment with $T = 1200$, $c_R = 0.05$, orthogonal contexts, and no historical human feedback. All policies use theoretical confidence radii. Reducing τ from 0.3 to 0.0380 reduces DELEGATE-UCB's mean routing regret from 108.01 to 0.20. At this finer precision, DELEGATE-UCB matches UNCERTAINTY-ONLY's mean routing loss with fewer audits (583.86 versus 791), giving total regret 29.39 versus 39.75. MACHINE-ONLY incurs higher routing loss (8.22), but its smaller audit count (75.18) gives lower total regret (11.98; Table 12). Appendix C.10 further compares margin- and tolerance-based thresholds at common evaluation tolerances.

5.2 RECORDED-CONTEXT REPLAY

We evaluate the policies on ImageNet16H, which pairs human classifications with VGG-19 predictions on noisy images (Steyvers et al., 2022). The replay uses one pooled-human arm: human assignments reveal a correctness indicator sampled from recorded human responses, while audits reveal the machine's recorded correctness. We split the base images into 240 for calibration and 960 for evaluation, keeping all noise variants of each image together. Policies use empirical exploration widths and are tuned independently at each audit price on the calibration split, with evaluation

Table 1: ImageNet16H replay: mean total regret with pointwise 95% bootstrap confidence intervals across 50 seeds. Lower values are better. Bold marks the lowest mean in each column.

	$T = 960$			$T = 3840$		
Method	$c_R = 0.01$	$c_R = 0.05$	$c_R = 0.20$	$c_R = 0.01$	$c_R = 0.05$	$c_R = 0.20$
DELEGATE-UCB	**70.45** [69.41, 71.61]	**78.88** [77.15, 80.80]	95.83 [93.18, 98.52]	**252.53** [249.33, 255.84]	**265.77** [261.48, 270.19]	303.46 [294.16, 314.93]
MACHINE-ONLY	72.39 [71.41, 73.43]	80.13 [78.43, 81.92]	95.60 [93.09, 98.17]	260.30 [257.57, 263.09]	272.81 [270.07, 275.57]	**297.20** [290.53, 304.36]
UNCERTAINTY-ONLY	71.50 [70.59, 72.51]	79.43 [77.76, 81.23]	95.84 [93.20, 98.57]	260.85 [257.89, 263.80]	270.52 [266.87, 274.50]	302.22 [292.86, 313.92]
SYMMETRIC-OVERLAP	77.63 [75.65, 79.79]	84.76 [82.98, 86.70]	98.13 [95.70, 100.56]	262.25 [255.05, 270.23]	282.46 [275.25, 290.50]	310.52 [299.15, 323.02]
RANDOM	77.43 [76.31, 78.61]	88.70 [87.20, 90.17]	106.16 [104.20, 108.12]	271.05 [267.88, 274.22]	300.57 [295.76, 305.74]	351.07 [337.97, 364.94]
FRONT-N	71.31 [70.31, 72.34]	78.91 [77.03, 80.92]	**93.69** [91.18, 96.29]	261.98 [259.44, 264.46]	270.02 [266.85, 273.38]	299.71 [291.15, 311.12]
DMR-POOL	71.92 [70.73, 73.12]	80.94 [79.39, 82.59]	101.48 [99.27, 103.72]	256.80 [253.55, 260.02]	270.73 [267.12, 274.51]	329.32 [313.58, 346.79]

At $T = 960$, each seed averages four episodes with fresh learners. At $T = 3840$, each seed is one continuous run with prespecified horizon-scaled audit thresholds. Calibration-selected audit fractions and the DMR-POOL multiplier are retained.

conducted over 50 paired seeds. Evaluation covers 960-round episodes with fresh learners and continuous 3840-round episodes, for which the threshold-based policies use prespecified horizon scaling of their audit thresholds (Appendix D.4).

Table 1 reports the replay results. FRONT-N audits the first N tasks, with its audit fraction selected on calibration data, while DMR-POOL is a finite-pool adaptation of DMR (Tucker et al., 2023).

Across both horizons, DELEGATE-UCB has the lowest observed mean total regret among the compared policies at $c_R = 0.01$ and 0.05. At $c_R = 0.20$, FRONT-N has the lowest mean at $T = 960$, whereas MACHINE-ONLY has the lowest mean at $T = 3840$. The paired 95% confidence intervals for differences between DELEGATE-UCB and these two baselines include zero at the highest audit price for both horizons, as does the interval for its difference from FRONT-N at the moderate price in the 960-round setting.

Our theoretical analysis balances residual uncertainty against audit expenditure to control worst-case total regret. The replay examines this trade-off on a fixed task distribution, where the relative performance of audit rules depends on the routing improvement obtained for their cost. At $T = 960$ and $c_R = 0.20$, DELEGATE-UCB's small reduction in mean routing regret relative to MACHINE-ONLY is outweighed by its additional audit cost, whereas FRONT-N purchases more audits (48 versus 31.54) and achieves sufficiently lower mean routing regret to offset its higher audit expenditure. At $T = 3840$ and the same audit price, MACHINE-ONLY incurs both lower mean routing regret and lower audit costs, while FRONT-N's lower mean routing regret again outweighs its additional audit expenditure. The paired-bootstrap intervals summarize variability conditional on the fixed image groups and selected configurations.

6 CONCLUSION

Our results link the cost of learning for delegation to the audit precision needed to distinguish handlers. A known margin for suboptimal machine assignments determines a sufficient precision level, while an organizational tolerance lets the organization choose which utility differences warrant further assessment. Human-route audits can acquire information before later machine assignments, and the experiments show how audit placement and precision affect the balance between routing loss and audit cost.

AI USE STATEMENT

Generative AI was used to produce the controlled experiment's synthetic task text and initial base-machine scores, and assisted with theoretical derivation, proof checking, experiment implementation, and manuscript editing. Complete proofs, validation checks, and experimental protocols are provided in the appendix. The authors take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

This work conducts no new human-participant data collection. The recorded-context replay is a secondary offline analysis of the published ImageNet16H behavioral dataset and makes no identity-level inference. Selective audit could reduce unnecessary expert review, but a misspecified policy could divert attention from underrepresented or high-risk regions. High-stakes use requires domain constraints, subgroup checks, distribution-shift monitoring, and human oversight.

REPRODUCIBILITY STATEMENT

Appendix B contains complete proofs. Appendix C specifies the controlled task generation, algorithms, hyperparameters, seeds, baselines, and validation checks, including the precision-threshold study in Appendix C.10, while Appendix D documents the data join, split, replay protocol, and paired uncertainty analysis. Supplementary scripts record the mechanism checks, shared-seed controlled comparisons, and calibration-only replay selection. The original task-generation code and cached controlled inputs are available at `https://anonymous.4open.science/r/Delegate_UCB-242D`.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Junyu Cao, Ruijiang Gao, and Esmaeil Keyvanshokooh. HR-Bandit: Human-AI collaborated linear recourse bandit. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3016–3024. PMLR, 2025.

Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. Human-AI collaboration with bandit feedback. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pp. 1722–1728, 2021.

Ruijiang Gao, Maytal Saar-Tsechansky, and Maria De-Arteaga. Learning complementary policies for human-AI teams. *arXiv preprint arXiv:2302.02944v2*, 2023. Revised November 2025.

Thodoris Lykouris and Wentao Weng. Learning to defer in congested systems: the AI-human interplay. *arXiv preprint arXiv:2402.12237*, 2024.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3578–3606, 2023.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable H-consistency and strong guarantees for learning to defer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43035–43066. PMLR, 2025.

Nadav Merlis and Shie Mannor. Lenient regret for multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8950–8957, 2021. doi: 10.1609/aaai.v35i10.17082.

Yannis Montreuil, Hoang Duy Dang, Maxime Meyer, Lai Xing Ng, Axel Carlier, and Wei Tsang Ooi. Online learning-to-defer with varying experts. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2701–2709. PMLR, 2026a.

Yannis Montreuil, Leïna Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Learning-to-defer with expert-conditional advice. *arXiv preprint arXiv:2603.14324*, 2026b.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10520–10545. PMLR, 2023.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29292–29304, 2022.

Manhin Poon, Xiangxiang Dai, Xutong Liu, Fang Kong, John C. S. Lui, and Jinhang Zuo. Online multi-LLM selection via contextual bandits under unstructured context evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24855–24863, 2026.

Andrea Pugnana, Giovanni De Toni, Cesare Barbera, Roberto Pellungrini, Bruno Lepri, and Andrea Passerini. To ask or not to ask: Learning to require human feedback. *arXiv preprint arXiv:2510.08314*, 2025.

Mirabel Reid, Tom Sühr, Claire Vernade, and Samira Samadi. Online decision deferral under budget constraints. *arXiv preprint arXiv:2409.20489*, 2024.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11261–11295, 2023a.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Selective sampling and imitation learning via online regression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67213–67268, 2023b.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022. doi: 10.1073/pnas.2111547119.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3475–3483. PMLR, 2024.

Aaron D. Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2147–2156. PMLR, 2023.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11415–11434. PMLR, 2023.

Wang Wei, Tiankai Yang, Hongjie Chen, Yue Zhao, Franck Dernoncourt, Ryan A Rossi, and Hoda Eldardiry. Learning to route LLMs from bandit feedback: One policy, many trade-offs. *arXiv preprint arXiv:2510.07429*, 2025.

Zixi Wei, Yuzhou Cao, and Lei Feng. Exploiting human-AI dependence for learning to defer. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52484–52499. PMLR, 2024.

Appendix A develops the audit rules and their information mechanism; Appendix B collects the regret proofs. Appendices C and D document the controlled experiments and recorded-context replay, respectively. Appendix E covers scope, broader impacts, and reproducibility.

A Algorithmic Details and Audit Mechanisms

Algorithm 1 uses the estimates and updates below. The audit-rule analysis then establishes when audits can be skipped and how early human-route feedback can improve later decisions.

A.1 Estimation, routing, and updates

The policy maintains separate ridge estimates for human skills and for the machine residual because the feedback channels are distinct. For each human $j \geq 1$, the estimate $\widehat{\theta}_{j,t} = V_{j,t}^{-1} b_{j,t}$ uses

$$V_{j,t} = \lambda I + \sum_{s<t: j_s = j} x_s x_s^\top, \qquad b_{j,t} = \sum_{s<t: j_s = j} x_s Y_s(j).$$

The machine-residual estimate $\widehat{w}_t = V_t^{-1} b_t$ uses

$$V_t = \lambda I + \sum_{s<t} q_s x_s x_s^\top, \qquad b_t = \sum_{s<t} q_s x_s \big(Z_s - b_0(x_s)\big).$$

The estimates give the utility bounds in (2).

Stage 1 chooses $h_t \in \arg\max_{j \geq 1} U_{j,t}(x_t)$ and routes to the larger of $U_{0,t}(x_t)$ and $U_{h_t,t}(x_t)$. Stage 2 applies (4). A routed human observation updates only that human's ridge statistics. An audited target updates only (V_t, b_t) through $Z_t - b_0(x_t)$.

A.2 Admissible audit rules and certified skips

A task is informative for the machine if $\nu_t(x_t) \geq \tau$. For a human-routed task, audit additionally requires

$$[L_{0,t}(x_t), U_{0,t}(x_t)] \cap [L_{h_t,t}(x_t), U_{h_t,t}(x_t)] \neq \varnothing.$$

On a human route, the rule skips an audit when the selected human's confidence interval is disjoint from the machine's. On a machine route, it audits whenever machine uncertainty reaches the threshold.

Definition 1 (τ-admissible audit rule)**.** An audit rule is τ-admissible if, on every round, it audits only when $\nu_t(x_t) \geq \tau$ and audits every machine-routed task satisfying this inequality.

The rule in (4) is τ-admissible.

Lemma 1 (Certified skips and symmetric overlap)**.** *Let $U_{H,t}(x) = \max_{j \geq 1} U_{j,t}(x)$. Under optimistic routing, all audit-count, budgeted, known-margin, and tolerance bounds in this paper also hold for any pre-feedback audit rule satisfying*

$$q_t = 1 \Rightarrow \nu_t(x_t) \geq \tau,$$
$$\big\{j_t = 0,\ \nu_t(x_t) \geq \tau,\ L_{0,t}(x_t) < U_{H,t}(x_t)\big\} \Rightarrow q_t = 1.$$

In particular, this class contains Symmetric-overlap *auditing,*

$$q_t^{\mathrm{sym}} = \mathbf{1}\{\nu_t(x_t) \geq \tau\}\mathbf{1}\{[L_{0,t}, U_{0,t}] \cap [L_{h_t,t}, U_{h_t,t}] \neq \varnothing\}.$$

Proof. On $\mathcal{E}_T(\delta)$, a machine route with $L_{0,t} \geq U_{H,t}$ satisfies $\mu_0 \geq L_{0,t} \geq U_{H,t} \geq \mu_H^\star$ and has zero routing loss. Every other unaudited machine route has $\nu_t < \tau$. Thus the unaudited machine terms in the proofs of Theorems 3, 4, and 5 retain their respective bounds $2\beta_T\tau$, zero under the margin condition, and zero excess loss at the tolerance threshold. The human and audited-machine terms are unchanged. The implication $q_t = 1 \Rightarrow \nu_t \geq \tau$ gives the same deterministic audit-count bound, and the same failure-event envelope gives the expectation bounds. On a machine route $U_{0,t} \geq U_{h_t,t}$; if $L_{0,t} < U_{h_t,t}$, the two closed intervals overlap. Hence the symmetric rule satisfies the displayed conditions, including when upper bounds tie. □

A.3 EARLY HUMAN-ROUTE INFORMATION

An audit on a correctly routed task can still improve later decisions. If an audit is purchased at x_t, the Sherman–Morrison identity gives, for any future context x',

$$\nu_{t+1}^2(x') = \nu_t^2(x') - \frac{(x'^\top V_t^{-1} x_t)^2}{1 + x_t^\top V_t^{-1} x_t}. \tag{5}$$

An audit can therefore reduce future uncertainty even when the current intervals are separated. The original rule audits every machine route with $\nu_t(x_t) \geq \tau$, whereas the symmetric rule requires overlap on both routes. The resulting design matrices differ, so neither rule necessarily makes fewer audits over the full sequence.

Proposition 2 (Early human-route information)**.** *Fix $a \in (0, 1/2]$, $c_R > 0$, $d = H = 1$, $\lambda = 1$, $S = 1/2$, and $\sigma = c_0 = c_1 = 0$. Let $\theta_1^\star = 1/2$, $w^\star = -1/2$, with 100 initial tasks at $x = 1$, $b_0(1) = -1/10$, followed by tasks at $x = a$, $b_0(a) = 2a/5$. For $T \geq 100 + \lceil 2/a^2 \rceil$, use the common threshold $\tau = 3a/5$ and the prescribed confidence radii $\beta_{0,t} = \beta_{1,t} = 1/2$. Then* DELEGATE-UCB *and* SYMMETRIC-OVERLAP *each make exactly two human-route audits, incur zero routing regret, and have total regret $2c_R$.* MACHINE-ONLY *auditing satisfies*

$$R_{\mathrm{M}}(T) \geq \left(c_R + \frac{3a}{5}\right)\left(\left\lfloor \frac{3}{4a^2} \right\rfloor + 1\right).$$

All confidence intervals cover the true utilities throughout the sequence.

Proof. After m noise-free observations at context 1, the human ridge estimate is $m/[2(1+m)]$ and its radius per unit context is $1/[2\sqrt{1+m}]$. The human upper bound is at least $1/2$ for every $m \geq 0$, so all initial tasks are human-routed even without machine feedback, whose initial upper bound is $2/5$. For DELEGATE-UCB, the first intervals overlap. After one human observation and one audit, the machine upper bound is $-1/10 - 1/4 + 1/(2\sqrt{2}) > 0$, while the human lower bound is $1/4 - 1/(2\sqrt{2}) < 0$, so the second task is audited as well. Both audit uncertainties exceed $\tau \leq 3/10$. After the first two audits, $V_3 = 3$ and $\widehat{w}_3 = -1/3$, so

$$U_0(1) = -1/10 - 1/3 + 1/(2\sqrt{3}) < 0 < 1/3 - 1/(2\sqrt{3}) = L_1(1).$$

Subsequent human observations increase this lower bound, so no further initial task is audited. At the transition,

$$U_0(a)/a = 2/5 - 1/3 + 1/(2\sqrt{3}) < 0.356,$$
$$L_1(a)/a = 50/101 - 1/(2\sqrt{101}) > 0.445.$$

The intervals stay separated as the human design grows. Both overlap policies therefore choose the human without further audits.

MACHINE-ONLY acquires no initial audits. After n later machine assignments, each audited while the threshold permits, write $z = na^2$. Its normalized upper bound is

$$U_0(a)/a = 2/5 - \frac{z}{2(1+z)} + \frac{1}{2\sqrt{1+z}}.$$

This expression decreases in z. For $z \leq 3/4$ it is at least $2/5 - 3/14 + 1/\sqrt{7} > 0.563$, whereas the unchanged human upper bound is $50/101 + 1/(2\sqrt{101}) < 0.545$. The machine is selected and $\nu_t(a) = a/\sqrt{1+z} \geq 2a/\sqrt{7} > 3a/5$, so it is audited. Induction gives at least $\lfloor 3/(4a^2) \rfloor + 1$ such rounds. Each incurs routing loss $3a/5$ and audit cost c_R. The assumed horizon contains these rounds. Finally, with zero noise, the ridge parameter error for either scalar parameter of magnitude $1/2$ is $1/(2V)$, bounded by $1/(2\sqrt{V})$ since $V \geq 1$. This verifies confidence coverage at both contexts for every policy. □

The policies use the same ridge-UCB update, initialization, and audit threshold on this block-ordered sequence. Early contexts supply more information per audit than later ones, allowing shadow assessment to improve later routing while the human handles the current task. Appendix C.11 examines noise, arrival order, and context geometry separately.

Table 2: Early calibration with valid confidence intervals. Here $\lambda = 1$, $S = 0.5$, $\sigma = 0$, $\tau = 0.3$, $c_R = 0.05$, and execution costs are zero. Counts are exact.

Policy	Human-route audits	Machine-route audits	Routing regret	Total regret
DELEGATE-UCB	2	0	0	0.10
SYMMETRIC-OVERLAP	2	0	0	0.10
MACHINE-ONLY	0	4	1.20	1.40
UNCERTAINTY-ONLY	11	0	0	0.55

Lemma 5 relates overlap to the true utility difference between the machine and the best human at the current confidence resolution.

B PROOFS OF REGRET AND AUDIT GUARANTEES

The proofs are grouped by role: shared concentration tools (Appendix B.1), information-theoretic lower bounds (Appendix B.2), gap-free upper bounds (Appendix B.3), and precision-dependent guarantees (Appendix B.4).

B.1 CONFIDENCE AND POTENTIAL TOOLS

B.1.1 CONFIDENCE EVENT

For a confidence level $\delta \in (0,1)$, let $\mathcal{E}_T(\delta)$ be the event on which the human and machine prediction errors satisfy

$$\left|x^\top(\widehat{\theta}_{j,t} - \theta_j^\star)\right| \le \beta_{j,t}\,\|x\|_{V_{j,t}^{-1}} \quad (j = 1,\dots,H),$$
$$\left|x^\top(\widehat{w}_t - w^\star)\right| \le \beta_{0,t}\,\|x\|_{V_t^{-1}}$$

for all $t \le T$ and $x \in \mathcal{X}$. The norm bounds $\|\widehat{\theta}_{j,t} - \theta_j^\star\|_{V_{j,t}} \le \beta_{j,t}$ and $\|\widehat{w}_t - w^\star\|_{V_t} \le \beta_{0,t}$, uniformly in time, imply these pointwise inequalities by Cauchy–Schwarz.

Lemma 2 (Uniform self-normalized confidence)**.** *Under Assumptions 1 and 2, fix $\delta \in (0,1)$ and allocate failure probabilities $\delta_0, \delta_1, \dots, \delta_H$ with $\sum_{j=0}^H \delta_j \le \delta$. Choose the human and machine radii as*

$$\beta_{j,t}(\delta_j) = \sigma\sqrt{\log\frac{\det(V_{j,t})}{\lambda^d \delta_j^2}} + \sqrt{\lambda}S, \quad j = 1,\dots,H,$$

$$\beta_{0,t}(\delta_0) = \sigma\sqrt{\log\frac{\det(V_t)}{\lambda^d \delta_0^2}} + \sqrt{\lambda}S.$$

Then $\mathbb{P}(\mathcal{E}_T(\delta)) \ge 1 - \delta$.

Proof. For each human j, write the adaptive regression on the full time index using design vectors $\tilde{x}_{j,t} := \mathbf{1}\{j_t = j\}x_t$ and noises $\tilde{\eta}_{j,t} := \mathbf{1}\{j_t = j\}\eta_{j,t}$. The action j_t is $\mathcal{G}_t$-measurable and precedes observation of the current noises, so the design sequence is predictable and the masked noises are conditionally mean-zero and σ-sub-Gaussian. The self-normalized concentration inequality for adaptive linear regression (Abbasi-Yadkori et al., 2011) gives, with probability at least $1 - \delta_j$,

$$\|\widehat{\theta}_{j,t} - \theta_j^\star\|_{V_{j,t}} \le \sigma\sqrt{\log\frac{\det(V_{j,t})}{\lambda^d \delta_j^2}} + \sqrt{\lambda}S \quad \text{for all } t \le T.$$

The machine residual regression uses design vectors $\tilde{x}_{0,t} := q_t x_t$ and noises $\tilde{\xi}_t := q_t \xi_t$. Here q_t is also $\mathcal{G}_t$-measurable, and $Z_t - b_0(x_t) = x_t^\top w^\star + \xi_t$ whenever $q_t = 1$. The same concentration inequality gives, with probability at least $1 - \delta_0$,

$$\|\widehat{w}_t - w^\star\|_{V_t} \le \sigma\sqrt{\log\frac{\det(V_t)}{\lambda^d \delta_0^2}} + \sqrt{\lambda}S \quad \text{for all } t \le T.$$

A union bound over $j = 0, 1, \dots, H$ yields the norm bounds simultaneously. The pointwise inequalities in $\mathcal{E}_T(\delta)$ follow by Cauchy-Schwarz. □

B.1.2 SCALE-AWARE ELLIPTICAL POTENTIAL

Lemma 3 (Scale-aware elliptical potential). *Let $V_1 = \lambda I$ with $\lambda > 0$, and let $V_{t+1} = V_t + a_t x_t x_t^\top$ for predictable indicators $a_t \in \{0, 1\}$. If $\|x_t\|_2 \le 1$ for all t, then*

$$\sum_{t=1}^{T} a_t \|x_t\|^2_{V_t^{-1}} \le A_\lambda \log \frac{\det(V_{T+1})}{\det(\lambda I)}.$$

Moreover, if $n_T := \sum_{t=1}^{T} a_t$, then

$$\sum_{t=1}^{T} a_t \|x_t\|^2_{V_t^{-1}} \le A_\lambda d \log\Big(1 + \frac{n_T}{\lambda d}\Big).$$

Proof. Since $V_t \succeq \lambda I$ and $\|x_t\|_2 \le 1$, we have $0 \le a_t \|x_t\|^2_{V_t^{-1}} \le 1/\lambda$. For any $u \in [0, 1/\lambda]$, $u \le A_\lambda \log(1 + u)$, because $\log(1 + u) \ge u/(1 + u) \ge u/A_\lambda$. Applying this inequality and the matrix determinant lemma gives

$$\begin{aligned}\sum_{t=1}^{T} a_t \|x_t\|^2_{V_t^{-1}} &\le A_\lambda \sum_{t=1}^{T} \log\Big(1 + a_t \|x_t\|^2_{V_t^{-1}}\Big) \\ &= A_\lambda \log \frac{\det(V_{T+1})}{\det(\lambda I)}.\end{aligned}$$

The context bound also gives

$$\operatorname{tr}(V_{T+1}) = \lambda d + \sum_{t=1}^{T} a_t \|x_t\|_2^2 \le \lambda d + n_T.$$

By the trace-determinant inequality, $\det(V_{T+1}) \le \big(\lambda + \frac{n_T}{d}\big)^d$. Together with $\det(\lambda I) = \lambda^d$, this proves the second bound. □

B.2 IDENTIFICATION AND AUDIT-COST LOWER BOUNDS

We first establish indistinguishability without audits, then quantify the information supplied by a finite audit budget and balance it against audit cost.

B.2.1 PROOF OF THEOREM 1

Proof. Consider the pair $\mathcal{I}_+(\gamma)$ and $\mathcal{I}_-(\gamma)$. Under a no-audit policy, $q_t = 0$ for every t. The human outcome is identically zero in both instances, and a machine-routed unaudited task produces no modeled learning feedback. Hence, by induction over t, the distribution of the learner's pre-action history is identical under the two instances.

Let $p_t := \mathbb{P}_{\mathcal{I}_+(\gamma)}(j_t = 0) = \mathbb{P}_{\mathcal{I}_-(\gamma)}(j_t = 0)$ be the common probability of routing round t to the machine. The machine is optimal under $\mathcal{I}_+(\gamma)$ and the human under $\mathcal{I}_-(\gamma)$. Each wrong route costs γ, and the policy pays no audit cost, giving

$$\begin{aligned}\mathrm{Reg}_\pi^{\mathcal{I}_+(\gamma)}(T) &= \gamma \sum_{t=1}^{T} (1 - p_t), \\ \mathrm{Reg}_\pi^{\mathcal{I}_-(\gamma)}(T) &= \gamma \sum_{t=1}^{T} p_t.\end{aligned}$$

Adding these regrets yields $\mathrm{Reg}_\pi^{\mathcal{I}_+(\gamma)}(T) + \mathrm{Reg}_\pi^{\mathcal{I}_-(\gamma)}(T) = \gamma T$. At least one of the two regrets is thus at least $\gamma T/2$. □

B.2.2 Proof of Theorem 2

Proof. Fix $S > 0$ and $\sigma > 0$. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the laws of the complete modeled interaction history under $\mathcal{I}_+(\Delta_B)$ and $\mathcal{I}_-(\Delta_B)$, respectively.

Given the same revealed history, the policy uses the same stochastic action kernel under $\mathbb{P}_+$ and $\mathbb{P}_-$, so the action contributes zero conditional KL. Observation distributions also agree unless $q_t = 1$. Conditional on the past and current action, an audited residual has distribution $\mathcal{N}(\Delta_B, \sigma^2)$ under $\mathbb{P}_+$ and $\mathcal{N}(-\Delta_B, \sigma^2)$ under $\mathbb{P}_-$. The relative-entropy chain rule therefore gives

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_+ \| \mathbb{P}_-) &= \mathbb{E}_+ \left[\sum_{t=1}^{T} q_t D_{\mathrm{KL}} \left(\mathcal{N}(\Delta_B, \sigma^2) \middle\| \mathcal{N}(-\Delta_B, \sigma^2) \right) \right] \\ &= \frac{2\Delta_B^2}{\sigma^2} \mathbb{E}_+[N_{\mathrm{aud}}(T)] \\ &\leq \frac{2B\Delta_B^2}{\sigma^2} \leq \frac{1}{8}, \end{aligned}$$

where the final inequality follows from $\Delta_B \leq \sigma/(4\sqrt{B+1})$.

For each round, let $M_t = \{j_t = 0\}$ denote the event that the policy routes to the machine. This event is measurable with respect to the interaction history and the policy's internal randomization. Its probability difference is bounded by total variation, which Pinsker's inequality controls:

$$|\mathbb{P}_+(M_t) - \mathbb{P}_-(M_t)| \leq \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_+ \| \mathbb{P}_-)} \leq \frac{1}{4}.$$

The two expected routing regrets then satisfy

$$\begin{aligned} &\mathrm{Reg}_{\pi,\mathrm{route}}^{\mathcal{I}_+(\Delta_B)}(T) + \mathrm{Reg}_{\pi,\mathrm{route}}^{\mathcal{I}_-(\Delta_B)}(T) \\ &\quad = \Delta_B \sum_{t=1}^{T} [\mathbb{P}_+(M_t^c) + \mathbb{P}_-(M_t)] \\ &\quad = \Delta_B \sum_{t=1}^{T} [1 - \mathbb{P}_+(M_t) + \mathbb{P}_-(M_t)] \\ &\quad \geq \Delta_B T \left(1 - \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)\right) \\ &\quad \geq \frac{3}{4} \Delta_B T. \end{aligned}$$

At least one of the two routing regrets must consequently be at least $3\Delta_B T/8$. □

B.2.3 Proof of Corollary 1

Proof. For $S > 0$ and $\sigma > 0$, fix any policy π. Its worst-case expected audit usage over the hard family is $B_\pi := \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{cal}}} \mathbb{E}_{\mathcal{I}}[N_{\mathrm{aud}}(T)]$. The pathwise bound $0 \leq N_{\mathrm{aud}}(T) \leq T$ implies $B_\pi \in [0, T]$.

Apply Theorem 2 to $\mathcal{I}_+(\Delta_{B_\pi})$ and $\mathcal{I}_-(\Delta_{B_\pi})$, with $\Delta_{B_\pi} = \min\{S, \sigma/(4\sqrt{B_\pi + 1})\}$. Both instances belong to $\mathfrak{I}_{\mathrm{cal}}$ and have expected audit counts bounded by B_π. Their routing lower bound gives the first inequality below. Nonnegativity of routing regret gives the second:

$$\begin{aligned} \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{cal}}} \mathrm{Reg}_\pi^{\mathcal{I}}(T) &\geq \frac{3}{8} T \Delta_{B_\pi}, \\ \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{cal}}} \mathrm{Reg}_\pi^{\mathcal{I}}(T) &\geq c_R B_\pi. \end{aligned}$$

Combining these bounds using $\max\{a, b\} \geq (a+b)/2$ and the definition of Δ_{B_π} yields

$$\begin{aligned} \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{cal}}} \mathrm{Reg}_\pi^{\mathcal{I}}(T) &\geq \max \left\{ \frac{3}{8} T \Delta_{B_\pi}, c_R B_\pi \right\} \\ &\geq c \left[T \min \left\{ S, \frac{\sigma}{\sqrt{B_\pi + 1}} \right\} + c_R B_\pi \right] \end{aligned}$$

for a universal constant $c > 0$.

Writing $f(B) := T \min\{S, \sigma/\sqrt{B+1}\} + c_R B$ for $B \in [0, T]$, we can take the infimum over policies:

$$\inf_{\pi} \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{cal}}} \mathrm{Reg}_{\pi}^{\mathcal{I}}(T) \geq c \inf_{\pi} f(B_\pi) \geq c \inf_{B \in [0,T]} f(B).$$

The second inequality uses the fact that $\{B_\pi : \pi \text{ admissible}\}$ is a subset of $[0, T]$. This proves the first display of the corollary.

When S does not bind and the minimizer is interior, balancing the information term $\sigma T/\sqrt{B}$ against the audit cost $c_R B$ gives

$$B^\star \asymp \left(\frac{\sigma T}{c_R}\right)^{2/3},$$
$$\frac{\sigma T}{\sqrt{B^\star}} + c_R B^\star = \Theta\Big((c_R \sigma^2)^{1/3} T^{2/3}\Big).$$

□

B.3 Audit complexity and gap-free upper bounds

Audit counts are controlled by elliptical potential and routing loss by the chosen handler's confidence width. We use these bounds for fixed thresholds, external budgets, and cost-tuned budgets, and characterize human-route audits through confidence overlap.

B.3.1 Proof of Proposition 1

Proof. By τ-admissibility, every audited round satisfies $\nu_t(x_t) = \|x_t\|_{V_t^{-1}} \geq \tau$, so $q_t \leq q_t \nu_t(x_t)^2/\tau^2$. Summing and applying Lemma 3 to the machine design with $a_t = q_t$ gives

$$N_{\mathrm{aud}}(T) \leq \frac{1}{\tau^2} \sum_{t=1}^{T} q_t \nu_t(x_t)^2 \leq \frac{A_\lambda d}{\tau^2} \log\left(1 + \frac{T}{\lambda d}\right).$$

The last step uses $\sum_t q_t \leq T$; this also gives the other part of the stated minimum bound, $N_{\mathrm{aud}}(T) \leq T$. □

B.3.2 Optimistic routing decomposition

Lemma 4 (Optimistic routing decomposition)**.** *On the confidence event $\mathcal{E}_T(\delta)$, for every round t,*

$$\mu_{j_t^\star}(x_t) - \mu_{j_t}(x_t) \leq U_{j_t,t}(x_t) - \mu_{j_t}(x_t).$$

Proof. Let $j_t^\star \in \arg\max_{j \in \mathcal{J}} \mu_j(x_t)$ be the oracle's handler. The stationary utility $\mu_0(x) = b_0(x) + x^\top w^\star - c_0$ is the same for the learner and oracle, so the machine UCB bounds that same utility on $\mathcal{E}_T(\delta)$. On that event,

$$\mu_{j_t^\star}(x_t) - \mu_{j_t}(x_t) = \underbrace{\mu_{j_t^\star}(x_t) - U_{j_t^\star,t}(x_t)}_{\leq 0} + \underbrace{U_{j_t^\star,t}(x_t) - U_{j_t,t}(x_t)}_{\leq 0} + \underbrace{U_{j_t,t}(x_t) - \mu_{j_t}(x_t)}_{\text{chosen-arm estimation gap}}.$$

The first term is nonpositive by confidence coverage, and the second by the optimistic choice of j_t in Stage 1. □

B.3.3 FRONTIER INTERPRETATION OF HUMAN-ROUTED AUDITS

Lemma 5 (Frontier interpretation of human-routed audits)**.** *Recall $\mu_H^\star(x) = \max_{j\geq 1}\mu_j(x)$. On the confidence event $\mathcal{E}_T(\delta)$, suppose $j_t = h_t \geq 1$ and $q_t = 1$ under* DELEGATE-UCB*. Write $s_{0,t} := \beta_{0,t}\nu_t(x_t)$ and $s_{h_t,t} := \beta_{h_t,t}\|x_t\|_{V_{h_t,t}^{-1}}$. Then*

$$\mu_H^\star(x_t) - \mu_0(x_t) \leq 2s_{0,t} + 2s_{h_t,t},$$
$$\mu_0(x_t) - \mu_H^\star(x_t) \leq 2s_{h_t,t}.$$

Proof. By the choice of h_t and confidence coverage, $\mu_H^\star(x_t) \leq \max_{j\geq 1} U_{j,t}(x_t) = U_{h_t,t}(x_t)$. An audit through the human branch requires interval overlap, which gives $L_{h_t,t}(x_t) \leq U_{0,t}(x_t)$. Therefore,

$$\begin{aligned}\mu_H^\star(x_t) &\leq U_{h_t,t}(x_t) = L_{h_t,t}(x_t) + 2s_{h_t,t}\\ &\leq U_{0,t}(x_t) + 2s_{h_t,t} \leq \mu_0(x_t) + 2s_{0,t} + 2s_{h_t,t}.\end{aligned}$$

For the reverse direction, human routing implies $U_{0,t}(x_t) \leq U_{h_t,t}(x_t)$, and hence

$$\begin{aligned}\mu_0(x_t) - \mu_H^\star(x_t) &\leq U_{0,t}(x_t) - \mu_H^\star(x_t) \leq U_{h_t,t}(x_t) - \mu_H^\star(x_t)\\ &\leq U_{h_t,t}(x_t) - \mu_{h_t}(x_t) \leq 2s_{h_t,t},\end{aligned}$$

where the penultimate inequality uses $\mu_H^\star(x_t) \geq \mu_{h_t}(x_t)$. □

B.3.4 FULL PATHWISE REGRET THEOREM

Theorem 6 (Full pathwise regret under admissible audit)**.** *Consider the optimistic routing rule* (3) *combined with any τ-admissible audit rule. Let $T_j = \sum_{t=1}^T \mathbf{1}\{j_t = j\}$. On $\mathcal{E}_T(\delta)$,*

$$\begin{aligned}R_\tau(T) \leq\ & 2\sum_{j=1}^H \beta_{j,T}\sqrt{A_\lambda d T_j \log\left(1 + \frac{T_j}{\lambda d}\right)}\\ &+ 2\beta_{0,T}\sqrt{A_\lambda d N_{\mathrm{aud}}(T)\log\left(1+\frac{T}{\lambda d}\right)} + 2\beta_{0,T}T\tau + c_R N_{\mathrm{aud}}(T).\end{aligned}$$

If $L_T = \log(1 + T/(\lambda d))$ and

$$\bar\beta_T(\delta) := \sigma\sqrt{dL_T + 2\log\frac{H+1}{\delta}} + \sqrt{\lambda}S,$$

then, with $\delta_j = \delta/(H+1)$, the determinant bound gives $\max_{j\in\mathcal{J}, t\leq T}\beta_{j,t} \leq \bar\beta_T(\delta)$ and $R_\tau(T) \leq B_T^{\mathrm{det}}(\tau,\delta)$, where

$$\begin{aligned}B_T^{\mathrm{det}}(\tau,\delta) :=\ & 2\bar\beta_T(\delta)\sqrt{A_\lambda d H L_T T} + \frac{2A_\lambda\bar\beta_T(\delta)dL_T}{\tau}\\ &+ 2\bar\beta_T(\delta)T\tau + \frac{A_\lambda c_R d L_T}{\tau^2}.\end{aligned}$$

Let $M := 2(B_0 + S + C_{\max}) + c_R$. Then

$$\mathrm{Reg}_\tau(T) \leq B_T^{\mathrm{det}}(\tau,\delta) + \delta MT.$$

Proof. By Lemma 4,

$$R_\tau(T) \leq \sum_{t=1}^T \big(U_{j_t,t}(x_t) - \mu_{j_t}(x_t)\big) + c_R N_{\mathrm{aud}}(T).$$

On a human route $j_t = j \geq 1$, confidence coverage bounds the estimation gap by $2\beta_{j,t}\|x_t\|_{V_{j,t}^{-1}}$, which is at most $2\beta_{j,T}\|x_t\|_{V_{j,t}^{-1}}$. Recall that $T_j = \sum_{t=1}^T \mathbf{1}\{j_t = j\}$. Cauchy–Schwarz and Lemma 3, applied to arm j's design, give

$$\sum_{t:j_t=j}\|x_t\|_{V_{j,t}^{-1}} \leq \sqrt{T_j\sum_{t:j_t=j}\|x_t\|^2_{V_{j,t}^{-1}}} \leq \sqrt{A_\lambda d T_j\log\left(1+\frac{T_j}{\lambda d}\right)}.$$

Thus the total human-routed contribution is at most

$$2\sum_{j=1}^{H}\beta_{j,T}\sqrt{A_\lambda dT_j\log\left(1+\frac{T_j}{\lambda d}\right)}.$$

On a machine route, the estimation gap is at most $2\beta_{0,t}\nu_t(x_t) \le 2\beta_{0,T}\nu_t(x_t)$. For the audited machine routes, Cauchy–Schwarz gives

$$\sum_{t:j_t=0,q_t=1}\nu_t(x_t) \le \sum_{t:q_t=1}\nu_t(x_t) \le \sqrt{N_{\mathrm{aud}}(T)\sum_{t:q_t=1}\nu_t(x_t)^2}.$$

The machine-design potential bound, with $a_t = q_t$, gives $\sum_{t:q_t=1}\nu_t(x_t)^2 \le A_\lambda d\log(1+T/(\lambda d))$. Thus audited machine routes contribute at most

$$2\beta_{0,T}\sqrt{A_\lambda dN_{\mathrm{aud}}(T)\log\left(1+\frac{T}{\lambda d}\right)}.$$

If $j_t = 0$ and $q_t = 0$, then τ-admissibility implies $\nu_t(x_t) < \tau$. Each such round contributes at most $2\beta_{0,T}\tau$, giving a total contribution of at most $2\beta_{0,T}T\tau$.

Adding these routing bounds and the audit cost proves the pathwise inequality.

For the deterministic envelope, use $T_j \le T$ and $\sum_{j=1}^{H} T_j \le T$ to obtain

$$\sum_{j=1}^{H}\sqrt{T_j\log\left(1+\frac{T_j}{\lambda d}\right)} \le \sqrt{HTL_T}.$$

The audit-count bound $N_{\mathrm{aud}}(T) \le A_\lambda dL_T/\tau^2$ controls both the audited machine term and the audit cost:

$$2\bar\beta_T(\delta)\sqrt{A_\lambda dL_T\,N_{\mathrm{aud}}(T)} \le \frac{2A_\lambda\bar\beta_T(\delta)dL_T}{\tau},$$
$$c_RN_{\mathrm{aud}}(T) \le \frac{A_\lambda c_RdL_T}{\tau^2}.$$

These bounds give $R_\tau(T) \le B_T^{\mathrm{det}}(\tau,\delta)$ on $\mathcal{E}_T(\delta)$. Assumption 1 and the finite-cost convention give $|\mu_j(x)| \le B_0 + S + C_{\max}$ for every agent and context, so each one-period pseudo-regret, including audit cost, is at most $M = 2(B_0 + S + C_{\max}) + c_R$. The failure probability is at most δ; taking expectations therefore gives $\mathrm{Reg}_\tau(T) \le B_T^{\mathrm{det}}(\tau,\delta) + \delta MT$. □

B.3.5 Proof of Theorem 3

Proof. Set $\delta_T = (T+1)^{-2}$ and $\beta_T = \bar\beta_T(\delta_T)$. With $\tau_B = \sqrt{A_\lambda dL_T/B}$, Proposition 1 gives $N_{\mathrm{aud}}(T) \le B$. On $\mathcal{E}_T(\delta_T)$, Theorem 6 gives the refined pathwise bound

$$\begin{aligned}R_{\tau_B}(T) \le{}& 2\sum_{j=1}^{H}\beta_T\sqrt{A_\lambda dT_j\log\left(1+\frac{T_j}{\lambda d}\right)} + 2\beta_T\sqrt{A_\lambda dL_TB}\\ &+ 2\beta_TT\sqrt{\frac{A_\lambda dL_T}{B}} + c_RB.\end{aligned} \tag{6}$$

The bound

$$\sum_{j=1}^{H}\sqrt{T_j\log\left(1+\frac{T_j}{\lambda d}\right)} \le \sqrt{HTL_T}$$

makes the first term at most $2\beta_T\sqrt{A_\lambda dHL_TT}$.

For the expected-regret statement, each per-round pseudo-regret is at most $M = 2(B_0 + S + C_{\max}) + c_R$. Lemma 2 gives $\mathbb{P}(\mathcal{E}_T(\delta_T)^c) \le \delta_T$, so the failure contribution is at most $\delta_TMT = MT/(T+1)^2 \le M/T$. Taking expectations and using $\beta_T = \widetilde{O}(\sqrt{d})$ gives estimation terms of order $\widetilde{O}(d\sqrt{HT} + d\sqrt{B} + dT/\sqrt{B})$. Since $B \le T$ and $H \ge 1$, the $d\sqrt{B}$ term is absorbed by $d\sqrt{HT}$. This proves Theorem 3. □

B.3.6 Proof of Corollary 2

Proof. Let $D_T := A_\lambda d L_T$ and $\widetilde{B}_T := (\beta_T^2 D_T T^2 / c_R^2)^{1/3}$. At this audit level, the terms $\beta_T T\sqrt{D_T/B}$ and $c_R B$ are both $Q_T := \beta_T^{2/3} D_T^{1/3} c_R^{1/3} T^{2/3}$. We bound the three budget-dependent terms in (6) for each clipping regime.

If $\widetilde{B}_T < 1$, then $B_T^\star = 1$ and $c_R > \beta_T T\sqrt{D_T}$. Since $T \geq 1$, $\beta_T\sqrt{D_T} \leq \beta_T T\sqrt{D_T} < c_R$, giving

$$\beta_T\sqrt{D_T B_T^\star} + \beta_T T\sqrt{\frac{D_T}{B_T^\star}} + c_R B_T^\star \leq 3c_R.$$

When $1 \leq \widetilde{B}_T \leq T$, rounding gives $\widetilde{B}_T \leq B_T^\star \leq \widetilde{B}_T + 1 \leq 2\widetilde{B}_T$. The three terms satisfy

$$\begin{aligned}
\beta_T T\sqrt{\frac{D_T}{B_T^\star}} &\leq \beta_T T\sqrt{\frac{D_T}{\widetilde{B}_T}} = Q_T, \\
c_R B_T^\star &\leq 2c_R\widetilde{B}_T = 2Q_T, \\
\beta_T\sqrt{D_T B_T^\star} &\leq \sqrt{2}\,\beta_T\sqrt{D_T\widetilde{B}_T} = \sqrt{2}\,\frac{\widetilde{B}_T}{T}Q_T \leq \sqrt{2}\,Q_T,
\end{aligned}$$

where the last inequality uses $\widetilde{B}_T \leq T$.

Finally, if $\widetilde{B}_T > T$, then $B_T^\star = T$ and $c_R < \beta_T\sqrt{D_T/T}$, so $c_R T < \beta_T\sqrt{D_T T}$. The two estimation terms become

$$\beta_T\sqrt{D_T B_T^\star} = \beta_T T\sqrt{\frac{D_T}{B_T^\star}} = \beta_T\sqrt{D_T T}.$$

Since $H \geq 1$, all three budget-dependent terms are bounded by a constant multiple of $\beta_T\sqrt{D_T H T}$.

Combining the three cases with (6) and the expectation conversion above gives

$$\mathrm{Reg}_{\tau_{B_T^\star}}(T) \leq \widetilde{O}\left(\beta_T\sqrt{A_\lambda d H T} + Q_T + c_R\right) + \frac{M}{T}.$$

With $\beta_T = \widetilde{O}(\sqrt{d})$ and fixed λ,

$$\begin{aligned}
\beta_T\sqrt{A_\lambda d H T} &= \widetilde{O}(d\sqrt{HT}), \\
\beta_T^{2/3}(A_\lambda d)^{1/3} c_R^{1/3} T^{2/3} &= \widetilde{O}(d^{2/3} c_R^{1/3} T^{2/3}).
\end{aligned}$$

Substituting these rates proves Corollary 2. □

B.4 Known margins and organizational tolerance

For each threshold choice, Lemma 2 gives a corresponding event $\mathcal{E}_T(\delta_T)$ of probability at least $1 - \delta_T$. On this event, every confidence radius is bounded by β_T, and Lemma 4 gives

$$g_t \leq \begin{cases} 2\beta_T \|x_t\|_{V_{j_t,t}^{-1}}, & j_t \geq 1, \\ 2\beta_T \nu_t(x_t), & j_t = 0. \end{cases} \tag{7}$$

Each human selection updates that human's design, and each audit updates the residual design. Writing $\Lambda_T = A_\lambda d L_T$, Lemma 3 yields

$$\begin{aligned}
&\sum_{t:j_t=j} \|x_t\|^2_{V_{j,t}^{-1}} \leq \Lambda_T \quad (j = 1, \ldots, H), \\
&\sum_{t=1}^{T} q_t \nu_t(x_t)^2 \leq \Lambda_T.
\end{aligned} \tag{8}$$

Both total regret and ϵ-insensitive total regret are bounded by MT on every sample path, since $[g_t - \epsilon]_+ \leq g_t$. The failure event therefore contributes at most $\delta_T MT = MT/(T+1)^2 \leq M/T$ to each expected bound.

The formulas below use $\beta_T > 0$. If $\beta_T = 0$, then $S = \sigma = 0$, so all unknown human and residual parameters, and their observation noises, are zero. The known base score and execution costs determine every utility exactly. Choosing $\tau > 1/\sqrt{\lambda}$ produces no audits and optimistic routing is optimal, giving zero regret for both objectives.

B.4.1 Proof of Theorem 4

Proof. Proposition 1 with $\tau_\Delta = \underline{\Delta}/(2\beta_T)$ gives the deterministic bound

$$N_{\mathrm{aud}}(T) \le \min\left\{T, \frac{4\beta_T^2 \Lambda_T}{\underline{\Delta}^2}\right\}.$$

Consider an unaudited machine route. Admissibility implies $\nu_t(x_t) < \tau_\Delta$, so (7) gives $g_t < \underline{\Delta}$. Whenever the machine is suboptimal, $g_t = \mu_H^\star(x_t) - \mu_0(x_t) \ge \underline{\Delta}$ by the margin condition. Hence every unaudited machine route has $g_t = 0$.

Every machine routing error therefore occurs on an audited round, where $g_t \ge \underline{\Delta}$ gives $g_t \le g_t^2/\underline{\Delta}$. By (7), the latter is at most $4\beta_T^2 \nu_t(x_t)^2/\underline{\Delta}$. Summing and using (8) gives

$$\sum_{t:j_t=0} g_t \le \frac{4\beta_T^2}{\underline{\Delta}} \sum_{t=1}^{T} q_t \nu_t(x_t)^2 \le \frac{4\beta_T^2 \Lambda_T}{\underline{\Delta}}.$$

For human routes, write $T_j = \sum_{t=1}^{T} \mathbf{1}\{j_t = j\}$. Cauchy–Schwarz and (8) give

$$\begin{aligned}\sum_{t:j_t \ge 1} g_t &\le 2\beta_T \sum_{j=1}^{H} \sqrt{T_j \sum_{t:j_t=j} \|x_t\|^2_{V_{j,t}^{-1}}} \\ &\le 2\beta_T \sqrt{\Lambda_T} \sum_{j=1}^{H} \sqrt{T_j} \le 2\beta_T \sqrt{H \Lambda_T T}.\end{aligned}$$

Adding the human and machine routing losses and $c_R N_{\mathrm{aud}}(T)$ yields

$$R_{\tau_\Delta}(T) \le 2\beta_T \sqrt{H\Lambda_T T} + \frac{4\beta_T^2 \Lambda_T}{\underline{\Delta}} + \frac{4c_R \beta_T^2 \Lambda_T}{\underline{\Delta}^2}. \tag{9}$$

Taking expectations adds at most M/T. Since $\beta_T = \widetilde{O}(\sqrt{d})$ and $\Lambda_T = \widetilde{O}(d)$, the stated rate follows. □

The one-sided condition constrains machine-routing errors. Suboptimal human routes can have arbitrarily small gaps, even when $H = 1$ and the machine is slightly better, so their contribution retains its square-root form. If every suboptimal handler is separated from the optimum by at least $\underline{\Delta}$, the squared-width argument also bounds the human contribution by $4H\beta_T^2\Lambda_T/\underline{\Delta}$. For $H = 1$, the additional two-sided condition $|\mu_0(x) - \mu_1(x)| \ge \underline{\Delta}$ has this implication.

B.4.2 Proof of Theorem 5

Proof. For $\tau_\epsilon = \epsilon/(2\beta_T)$, Proposition 1 gives

$$N_{\mathrm{aud}}(T) \le \min\left\{T, \frac{4\beta_T^2 \Lambda_T}{\epsilon^2}\right\}.$$

On an unaudited machine route, admissibility and (7) imply $g_t \le 2\beta_T \nu_t(x_t) < \epsilon$, so its excess loss is zero. For every $s \ge 0$,

$$[2s - \epsilon]_+ \le \frac{s^2}{\epsilon}. \tag{10}$$

Indeed, the right-hand side is nonnegative and $s^2/\epsilon - (2s - \epsilon) = (s - \epsilon)^2/\epsilon \ge 0$. Applying (10) to (7), then summing over human routes and audited machine routes, gives

$$\begin{aligned}\sum_{t=1}^{T} [g_t - \epsilon]_+ &\le \frac{\beta_T^2}{\epsilon} \left(\sum_{j=1}^{H} \sum_{t:j_t=j} \|x_t\|^2_{V_{j,t}^{-1}} + \sum_{t:j_t=0, q_t=1} \nu_t(x_t)^2 \right) \\ &\le \frac{(H+1)\beta_T^2 \Lambda_T}{\epsilon},\end{aligned}$$

where the last step uses (8). Adding the full audit cost proves

$$R^{(\epsilon)}_{\tau_\epsilon}(T) \le \frac{(H+1)\beta_T^2\Lambda_T}{\epsilon} + c_R \min\left\{T, \frac{4\beta_T^2\Lambda_T}{\epsilon^2}\right\}. \tag{11}$$

Taking expectations adds at most M/T. Substituting $\beta_T^2\Lambda_T = \widetilde{O}(d^2)$ gives Theorem 5. □

B.4.3 Decision precision and ordinary regret

For the same run of the policy, the identity $g_t = [g_t - \epsilon]_+ + \min\{g_t, \epsilon\}$ gives the pathwise relation

$$R_{\tau_\epsilon}(T) \le \epsilon T + R^{(\epsilon)}_{\tau_\epsilon}(T). \tag{12}$$

Passing from tolerant to ordinary regret adds at most ϵ per task. To recover the gap-free rate with a horizon-dependent tolerance, write $\Gamma_T = \beta_T^2\Lambda_T$ and $K_T = (H+1)\Gamma_T$. Choose

$$\epsilon_T = \max\left\{\sqrt{\frac{K_T}{T}}, \left(\frac{8c_R\Gamma_T}{T}\right)^{1/3}\right\},$$

and run the policy with $\tau_{\epsilon_T} = \epsilon_T/(2\beta_T)$. Equations (11) and (12) give, on $\mathcal{E}_T(\delta_T)$,

$$R_{\tau_{\epsilon_T}}(T) \le \epsilon_T T + \frac{K_T}{\epsilon_T} + \frac{4c_R\Gamma_T}{\epsilon_T^2}.$$

The three terms satisfy

$$\begin{aligned}
\epsilon_T T &\le \sqrt{K_T T} + 2(c_R\Gamma_T)^{1/3}T^{2/3},\\
\frac{K_T}{\epsilon_T} &\le \sqrt{K_T T},\\
\frac{4c_R\Gamma_T}{\epsilon_T^2} &\le (c_R\Gamma_T)^{1/3}T^{2/3}.
\end{aligned}$$

Consequently,

$$R_{\tau_{\epsilon_T}}(T) \le 2\sqrt{K_T T} + 3(c_R\Gamma_T)^{1/3}T^{2/3}. \tag{13}$$

The square-root term in the definition of ϵ_T controls the inverse-tolerance contribution when the audit price is small. The same calculation holds at $c_R = 0$, with the audit-cost term equal to zero. The expected version of (13) adds at most M/T. For standard $\beta_T = \widetilde{O}(\sqrt{d})$ and fixed λ, its leading orders are $\widetilde{O}(d\sqrt{(H+1)T} + d^{2/3}c_R^{1/3}T^{2/3})$.

Remark 2 (Hard-threshold tolerance). For the alternative objective $\sum_t g_t\mathbf{1}\{g_t > \epsilon\} + c_R N_{\mathrm{aud}}(T)$, the same threshold gives the high-probability bound in (11) with the routing coefficient 1 replaced by 4. Indeed, unaudited machine routes have $g_t < \epsilon$, and on the remaining routes $g_t\mathbf{1}\{g_t > \epsilon\} \le g_t^2/\epsilon$. Equation (7) and the same potential sums therefore give routing loss at most $4(H+1)\beta_T^2\Lambda_T/\epsilon$.

Remark 3 (External audit budgets). For an external cap $B \in [1, T)$, Proposition 1 certifies $N_{\mathrm{aud}}(T) \le B$ whenever $\tau \ge \tau_B = \sqrt{\Lambda_T/B}$. Let ζ denote either $\underline{\Delta}$ or ϵ and write $\tau_\zeta = \zeta/(2\beta_T)$. The unaudited-route argument uses $\tau \le \tau_\zeta$. If $\tau_B \le \tau_\zeta$, choosing $\tau = \tau_\zeta$ supplies both the precision guarantee and the budget certificate. If $\tau_B > \tau_\zeta$, choosing $\tau = \tau_B$ supplies the guarantee in Theorem 3. Thus $\tau = \max\{\tau_B, \tau_\zeta\}$ implements these two choices while preserving admissibility throughout the run. These budget conditions are sufficient certificates; the cap $B = T$ is satisfied by every policy and imposes no additional threshold constraint.

C Additional Experimental Details

The controlled task-pool experiment uses LLM-generated task text and base-machine scores. Utilities and audited targets follow the linear-residual model. Task generation and base scoring use `qwen3.6-plus` through an OpenAI-compatible API, with temperature zero for base scoring.

C.1 TASK GENERATION AND NORMALIZED FEATURES

The cached pool contains 1000 tasks, 250 each from billing, refund, technical support, and compliance. Each JSON record contains `task_id`, `domain`, `difficulty`, `ambiguity`, `risk`, `task_text`, and `gold_resolution`; the three continuous metadata fields lie in $[0, 1]$. The learner uses the metadata, not the raw task text. With compliance as the omitted category, define

$$\tilde{x}_t = (1, \text{difficulty}_t, \text{ambiguity}_t, \text{risk}_t, \mathbf{1}\{\text{billing}\}, \mathbf{1}\{\text{refund}\}, \mathbf{1}\{\text{technical}\}).$$

We use $x_t = \tilde{x}_t/\sqrt{5}$; since $\|\tilde{x}_t\|_2^2 \le 5$, the contexts satisfy $\|x_t\|_2 \le 1$. Raw-basis coefficients are multiplied by $\sqrt{5}$ in the normalized basis, preserving all latent utilities and oracle decisions.

The scale-equivalent ridge parameter is $\lambda = \lambda_{\text{raw}}/5 = 0.2$, where $\lambda_{\text{raw}} = 1$. For any common observation indicators, the corresponding designs satisfy $V_t = \widetilde{V}_t/5$, and hence

$$x_t^\top \widehat{\theta}_t = \tilde{x}_t^\top \widehat{\theta}_t^{\text{raw}}, \qquad x_t^\top V_t^{-1} x_t = \tilde{x}_t^\top \widetilde{V}_t^{-1} \tilde{x}_t.$$

Thus normalization preserves the fitted scores and widths as well as the environment; the scale factor in the theory is $A_\lambda = 6$.

C.2 BASE-MACHINE SCORING PROMPT

For each task, the base model receives the metadata and customer ticket text, without the hidden resolution. It generates a first-draft answer and a score for the probability that the answer would pass a strict human audit without revision. We obtain the base prior $b_0(x_t)$ from this score using the following prompt:

```
SYSTEM:
You are a strict, risk-averse quality auditor for customer-service and
compliance responses. You are not scoring whether the answer sounds fluent.
You are estimating the probability that the machine answer would be accepted
by a strict human auditor without any revision.

Important scoring rule:
The score is NOT the model's confidence.
The score is NOT whether the answer is plausible.
The score is the probability that the answer would pass strict audit as-is.

Be conservative:
- Start from a default score around 0.55 for nontrivial tickets.
- Scores above 0.80 should be rare.
- Scores above 0.85 should only be used for trivial, low-risk, unambiguous
 tickets.
- High-risk tickets should usually be below 0.70.
- Ambiguous tickets should usually be below 0.70.
- Compliance-related tickets should usually be below 0.65 unless the answer
 is extremely complete and safe.
- If the answer requires policy details, account verification, legal or
 compliance judgment, or missing customer information, the score should
 usually be below 0.65.
- If the answer may need human audit, escalation, or additional information,
 the score should be between 0.40 and 0.65.
- If the answer is unsafe, overconfident, or makes unsupported assumptions,
 the score should be below 0.50.

Score bands:
- 0.90-1.00: Almost certainly accepted without revision. Use only for
 trivial, low-risk cases.
- 0.75-0.89: Likely accepted, but only for clear, low-risk tickets with a
 complete answer.
- 0.60-0.74: Possibly accepted, but some details or caveats may be missing.
- 0.40-0.59: Uncertain; likely needs human audit or revision.
- 0.20-0.39: Likely rejected or unsafe.
- 0.00-0.19: Clearly wrong, dangerous, or unusable.

Return strict JSON only:
{
 "machine_answer": "...",
 "predicted_quality": 0.0
}

USER:
Task metadata:
```

```
- domain: {domain}
- difficulty: {difficulty}
- ambiguity: {ambiguity}
- risk: {risk}

Customer ticket:
{task_text}

Generate a first-draft machine answer, then assign predicted_quality
according to the strict audit acceptance probability rubric above.
```

Let $\widetilde{b}_0(x_t)$ denote the raw predicted quality score returned by the base model. We set $b_0(x_t) = \mathrm{clip}\{\widetilde{b}_0(x_t), 0.2, 0.9\}$. The base prior is kept separate from the normalized context x_t. In the simulation, $b_0(x_t)$ and the corrected target $\bar{Z}_t = b_0(x_t) + x_t^\top w^\star$ serve as latent routing-quality scores, not calibrated probabilities. The residual model determines utilities and audited targets; the LLM supplies task text and the base score.

C.3 Controlled utilities and feedback

In the raw metadata basis defined above, the fixed coefficients are

$$\begin{aligned} w^\star_{\mathrm{raw}} &= (0.04, -0.06, -0.06, -0.10, 0.08, 0.05, -0.04), \\ \theta^\star_{1,\mathrm{raw}} &= (0.62, -0.05, -0.04, -0.10, 0.10, 0.08, 0), \\ \theta^\star_{2,\mathrm{raw}} &= (0.71, -0.04, -0.03, 0.10, -0.16, -0.16, -0.16). \end{aligned}$$

Set $w^\star = \sqrt{5} w^\star_{\mathrm{raw}}$ and $\theta^\star_j = \sqrt{5}\theta^\star_{j,\mathrm{raw}}$. Human 1 specializes in billing and refunds. Human 2 benefits from risk and compliance; its raw intercept incorporates the omitted compliance indicator. All three normalized coefficient norms are below 2.

The mean machine score is $\bar{Z}_t = b_0(x_t) + x_t^\top w^\star$, and the human scores are $s_j(x_t) = x_t^\top \theta^\star_j$. Utilities subtract execution costs: $\mu_0 = \bar{Z}_t - c_0$, $\mu_j = s_j - c_j$, with $(c_0, c_1, c_2) = (0.02, 0.10, 0.18)$. The oracle maximizes these utilities over the three handlers. An audit reveals $Z_t = \bar{Z}_t + \xi_t$, and a human assignment reveals $Y_t(j) = s_j(x_t) + \eta_{j,t}$, with independent Gaussian noises of standard deviation 0.03. Only a paid audit reveals the machine residual $Z_t - b_0(x_t)$; all updates affect future decisions.

C.4 Algorithms

All policies use the optimistic router and ridge updates in Algorithm 1 and differ only in their audit decisions. They share potential human and audit noises for each (t, seed). The fixed exploration multipliers are $\beta_0^{\mathrm{emp}} = 0.25$ and $\beta_h^{\mathrm{emp}} = 0.35$. These widths are exploration scores, not calibrated confidence bands. For $B > 0$, the audit rule in (4) uses the threshold

$$\tau_B^{\mathrm{emp}} = \sqrt{\frac{2d \log(1 + T/(\lambda d))}{B}}.$$

Writing q_t^{uncap} for the uncapped decision, we enforce the budget with $q_t = \mathbf{1}\{q_t^{\mathrm{uncap}} = 1,\ N_{\mathrm{aud}}(t-1) < B\}$. No reported selective trajectory reaches the cap. Auditing is disabled at $B = 0$.

Audit-rule ablations. MACHINE-ONLY audit uses the same threshold only when $j_t = 0$. UNCERTAINTY-ONLY audits every task above the threshold. The κ-frontier ablation replaces interval overlap by $U_{h_t,t} - U_{0,t} \le \kappa \beta_0^{\mathrm{emp}} \nu_t(x_t)$ and sweeps $\kappa \in \{0.5, 1, 2, 4\}$. NO AUDIT sets $q_t = 0$ throughout.

Random and equal-count baselines. For each seed, the complete task table is randomly permuted once, and the same seed-specific permutation is shared across all policies and budget caps. FULL-BUDGET RANDOM chooses exactly B uniformly random rounds. For each (B, seed), MATCHED RANDOM and FRONT-LOADED first record the realized audit count N of the main policy. MATCHED RANDOM chooses exactly N rounds uniformly without replacement, while FRONT-LOADED audits the first N rounds of the shared random permutation. Matching N isolates audit placement from quantity. These are post-hoc evaluation controls: N is obtained from the completed DELEGATE-UCB trajectory, not supplied to an online policy.

C.5 BUDGETS, METRICS, AND IMPLEMENTATION

We use caps $B \in \{0, 25, 50, 100, 200, 400\}$ and the same 20 seeds $s = 0, \ldots, 19$ for the controlled comparisons and robustness baseline. NumPy SeedSequence streams $[s, 3101]$, $[s, 1103]$, and $[s, 2207]$ generate the shared task permutations, human potential noises, and audit noises. FULL-BUDGET RANDOM uses $[s, B, 4103]$ and MATCHED RANDOM $[s, B, 9049]$; FRONT-LOADED uses the initial positions of the shared order. An independent implementation reproduces the original matrix-update DELEGATE-UCB trajectories.

With audit price $c_R = 0.05$ and audit count $N_{\mathrm{aud}}(T) = \sum_{t=1}^{T} q_t$, we report routing and total regret:

$$R_{\mathrm{route}}(T) = \sum_{t=1}^{T} [\max_{j \in \{0,1,2\}} \mu_j(x_t) - \mu_{j_t}(x_t)],$$
$$R(T) = R_{\mathrm{route}}(T) + c_R N_{\mathrm{aud}}(T).$$

Net utility is $\sum_t \mu_{j_t}(x_t) - c_R N_{\mathrm{aud}}(T)$. All methods satisfy their audit caps without truncating selective trajectories.

Paired-bootstrap intervals. For each reported comparison, the resampling unit is the paired seed-level regret difference. We draw 20,000 bootstrap samples by resampling the 20 paired differences with replacement and report the 2.5th and 97.5th percentiles of the resampled means, with bootstrap seed 2026090799. The MATCHED RANDOM schedule randomness is part of each seed-level observation through its deterministic seed-specific random-number stream.

C.6 EMPIRICAL-WIDTH STRESS TEST

The stress family is exactly linearly realizable and uses empirical exploration widths. We set $T = 1200$, $d = 7$, use contexts $x_t \in \{e_1, \ldots, e_7\}$, and retain the same costs, ridge parameter, noise scale, and empirical widths as in the controlled task-pool experiment. The human score vectors and machine residual are

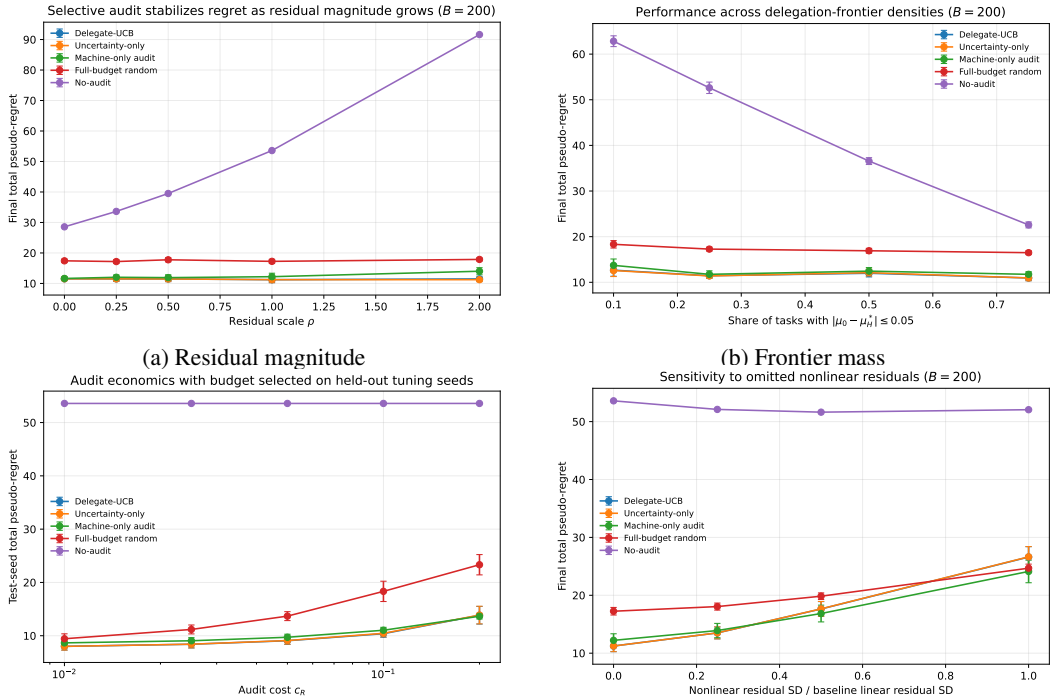
$$\theta_1 = (0.75, 0.75, 0.75, 0.75, 0.75, 0.35, 0.35),$$
$$\theta_2 = (0.25, 0.25, 0.25, 0.25, 0.25, 0.20, 0.20),$$
$$w^\star = (0, 0, 0, 0, 0, 0.75, -0.25).$$

The first five directions are human-dominant decoys. Their base score is set so that the true human-machine gap is in $\{0.6, 0.8, 1.0\}$; because their residual is zero, audit can reduce statistical uncertainty but cannot change the optimal handler. On e_6 the base score is -0.35 and the positive residual is initially hidden behind a human route. On e_7 the base score is 0.48 and the residual is negative. We vary the decoy mass over $\{0, 0.25, 0.5, 0.75\}$, use eight thresholds $\tau \in \{1.2, 1.0, 0.8, 0.6, 0.5, 0.4, 0.3, 0.25\}$, and average 100 paired seeds. No hard cap binds. A diagonal implementation is algebraically identical to the ridge updates because all contexts are basis vectors; it was validated exactly against the generic matrix implementation.

On e_6, the initial machine upper score is $-0.35 - 0.02 + 0.25/\sqrt{0.2} \simeq 0.189$, whereas the best human has utility 0.25 and the machine has utility 0.38. Thus this family measures behavior under initial machine undercoverage. Across its 72 nonzero-decoy cells, DELEGATE-UCB matches UNCERTAINTY-ONLY routing regret with 21.9 fewer audits on average; MACHINE-ONLY incurs 38.8 more routing regret. These differences measure performance under the empirical initialization. Appendix C.11 compares the same policies using theoretical confidence radii to separate early information from undercoverage.

C.7 ONE-FACTOR ROBUSTNESS MATRIX

The robustness matrix changes one factor at a time from the controlled task pool. We evaluate DELEGATE-UCB, UNCERTAINTY-ONLY, MACHINE-ONLY audit, NO AUDIT, and FULL-BUDGET RANDOM audit using 20 paired seeds. Unless otherwise stated, we use $B \in \{50, 100, 200, 400\}$ and report $B = 200$ below. Every seed induces one random task order shared by all policies and budgets, and all policies share the same potential human and audit noises.



(a) Residual magnitude (b) Frontier mass

(c) Audit cost (d) Nonlinear misspecification

Figure 2: One-factor robustness analyses. Panels (a), (b), and (d) use $B = 200$ and report mean total pseudo-regret with normal-approximation 95% error bars over 20 paired seeds. Panel (c) selects the budget cap on ten tuning seeds and evaluates it on ten disjoint test seeds.

Residual magnitude. Let $r_t = \bar{Z}_t - b_0(x_t)$ denote the baseline controlled residual. We replace it by $r_t(\rho) = \rho r_t$ for $\rho \in \{0, 0.25, 0.5, 1, 2\}$, then recompute the machine utility and oracle. Human utilities and base scores are unchanged. At $\rho = 0$, the learner must still estimate the unknown correction; audits can reduce routing regret by establishing that it is close to zero.

Frontier mass. We vary frontier mass by reweighting arrivals while keeping each task's utilities fixed, so its loss under an incorrect route is unchanged. We define a near-frontier stratum by $|\mu_0(x) - \mu_H^\star(x)| \le 0.05$ and sample a fixed population of 5000 rows with replacement so that its near-frontier share is $p_F \in \{0.10, 0.25, 0.50, 0.75\}$. For each seed, we randomly permute this population and use the first $T = 1000$ records, sharing the sequence across policies and budgets. The resulting routing-flip share rises from 0.088 to 0.338 as p_F increases, while the original utility attached to each sampled task is unchanged.

Audit economics. We compare policies at five audit prices, $c_R \in \{0.01, 0.025, 0.05, 0.1, 0.2\}$, with audit caps $\{25, 50, 100, 200, 400\}$. For each method and cost, we select the cap that minimizes mean total pseudo-regret on seeds 0-9 and evaluate it on the disjoint seeds 10-19. This comparison measures finite-grid tuning performance rather than estimating a $c_R^{-2/3}$ scaling law.

Nonlinear misspecification. Let $d_t, a_t, r_t^{\text{risk}}$ denote the difficulty, ambiguity, and risk metadata. We construct the omitted component from

$$v_t = (d_t - \bar{d})(r_t^{\text{risk}} - \bar{r}^{\text{risk}}) + \frac{1}{2}(a_t - \bar{a})^2,$$

then center and normalize it as $\psi_t = (v_t - \bar{v})/\operatorname{sd}(v)$. Means and standard deviations are computed over the controlled task pool. If σ_r is the standard deviation of the baseline linear residual, the true target becomes $\bar{Z}_t(\alpha) = b_0(x_t) + r_t + \alpha\sigma_r\psi_t$. We vary $\alpha \in \{0, 0.25, 0.5, 1\}$. The learner continues to fit the seven-dimensional linear model, which omits this nonlinear component when $\alpha > 0$.

Table 3: Robustness at $B = 200$. Values are mean total pseudo-regret over 20 paired seeds; "Audits" is the DELEGATE-UCB realized count. The nonlinear strength is measured relative to the standard deviation of the baseline linear residual.

Axis	Level	DELEGATE-UCB	Audits	MACHINE-ONLY	FULL RANDOM	NO AUDIT
Residual scale ρ	0.00	11.51	12.35	11.64	17.42	28.57
	0.25	11.49	12.35	12.03	17.18	33.63
	0.50	11.49	12.30	11.91	17.76	39.51
	1.00	11.24	12.30	12.22	17.25	53.59
	2.00	11.49	12.20	14.00	17.88	91.62
Frontier mass p_F	0.10	12.66	11.55	13.71	18.32	62.83
	0.25	11.44	11.85	11.75	17.27	52.62
	0.50	11.98	11.80	12.45	16.91	36.56
	0.75	10.95	12.00	11.74	16.49	22.58
Nonlinear strength α	0.00	11.24	12.30	12.22	17.25	53.59
	0.25	13.50	12.35	13.91	18.05	52.11
	0.50	17.64	12.35	16.85	19.84	51.63
	1.00	26.61	12.40	24.12	24.68	52.06

Table 4: Held-out audit-cost tuning. Caps are selected on seeds 0-9 and reported metrics are evaluated on seeds 10-19.

	DELEGATE-UCB			FULL-BUDGET RANDOM		MACHINE-ONLY
c_R	Selected cap	Audits	Total regret	Selected cap	Total regret	Total regret
0.010	400	26.4	8.01	200	9.43	8.66
0.025	400	26.4	8.40	100	11.18	9.05
0.050	400	26.4	9.06	100	13.68	9.71
0.100	400	26.4	10.38	50	18.32	11.03
0.200	200	12.1	13.84	50	23.32	13.66

DELEGATE-UCB's mean total regret stays nearly flat across the tested residual scales. Across the frontier-mass levels, DELEGATE-UCB has lower mean total regret than MACHINE-ONLY audit; reweighting changes both the frequency and size of routing mistakes, so regret need not vary monotonically with frontier mass. Under nonlinear misspecification, DELEGATE-UCB has higher mean total regret than MACHINE-ONLY audit at $\alpha = 0.5$ and 1, and also exceeds FULL-BUDGET RANDOM audit at $\alpha = 1$. Decision-focused filtering therefore depends on the residual geometry assumed by the learner.

C.8 PARAMETERS AND VALIDATION

Table 5 collects the configuration and task-pool checks. The oracle uses all three handlers, and the base score varies across tasks. NO AUDIT routing and total regret coincide because its audit count is zero. Observed budget feasibility is distinct from the uncapped τ-admissibility condition used in the theory.

C.9 MATCHED-COUNT AND AUDIT-RULE DIAGNOSTICS

The main policy never reaches the hard cap at a reported positive budget. FULL-BUDGET RANDOM always uses the nominal cap, which can lower routing regret while increasing audit cost linearly in B. Its total regret is lower at $B = 100$, but the ranking reverses at $B = 200$ and $B = 400$.

MATCHED RANDOM-minus-DELEGATE-UCB paired total-regret differences are 20.36 $[15.87, 24.77]$, 18.30 $[14.61, 22.05]$, and 10.07 $[8.49, 11.68]$ at $B = 100, 200, 400$. The corresponding FRONT-LOADED differences are 8.65 $[4.63, 12.93]$, 3.60 $[1.86, 5.64]$, and 0.48 $[-0.04, 1.07]$. The paired intervals for the differences between the $\kappa = 2$ ablation and DELEGATE-UCB include zero. The κ sweep measures sensitivity to the frontier width.

Table 5: Controlled-pool configuration and validation. Noise entries are standard deviations; oracle counts refer to the fixed 1000-task pool.

Quantity	Value
Tasks / domains / humans / dimension	1000 / 4 (250 tasks each) / 2 / 7
Paired seeds / budget caps	20 / $\{0, 25, 50, 100, 200, 400\}$
Normalization / maximum context norm	$x = \tilde{x}/\sqrt{5}$ / 0.961
Ridge: normalized / raw-equivalent	0.2 / 1 ($A_\lambda = 6$)
Exploration: machine / human	0.25 / 0.35
Feedback noise: machine / human	0.03 / 0.03
Execution costs: machine / humans	0.02 / $(0.10, 0.18)$
Audit price / κ sweep	0.05 / $\{0.5, 1, 2, 4\}$
Ground truth / LLM judge	Controlled linear residual / none
Base score: mean / standard deviation	0.615 / 0.154
Base score: minimum / maximum	0.20 / 0.90
Clipping: lower / upper / any	0.008 / 0.038 / 0.046
Oracle counts: machine / humans	609 / $(139, 252)$
NO AUDIT: count / routing and total regret	0 / 53.5875
Audit-budget violations	0

Table 6: Controlled-pool audit placement (20 paired seeds). MATCHED RANDOM and FRONT-LOADED use DELEGATE-UCB's exact seed-specific audit count N. Routing regret excludes audit cost; total regret adds $0.05N$ for the three matched-count policies. FULL RANDOM buys B audits.

		Routing regret			Total regret		
B	N	DELEGATE-UCB	MATCHED RANDOM	FRONT-LOADED	DELEGATE-UCB	MATCHED RANDOM	FULL RANDOM
100	5.15	19.80	40.16	28.45	20.06	40.42	13.65
200	12.30	10.62	28.92	14.22	11.24	29.54	17.25
400	26.95	7.37	17.44	7.85	8.72	18.78	26.45

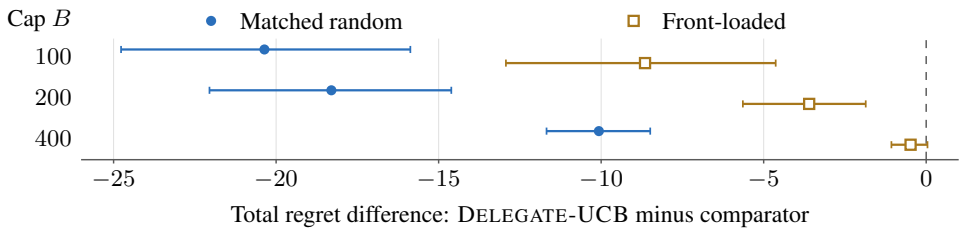


Figure 3: Audit placement at matched realized counts ($T = 1000$, 20 seeds, $c_R = 0.05$). Points and bars show mean differences and pointwise 95% paired-bootstrap intervals. Negative differences favor DELEGATE-UCB. Its mean realized audit counts are 5.15, 12.30, and 26.95 for caps $B = 100, 200, 400$, respectively.

Table 7: Audit-rule comparisons under the common controlled-pool protocol. Each entry is mean total regret / mean audit count over 20 paired seeds.

Policy	$B = 100$	$B = 200$	$B = 400$
DELEGATE-UCB	20.06 / 5.15	11.24 / 12.30	8.72 / 26.95
SYMMETRIC-OVERLAP	19.99 / 4.85	11.15 / 11.95	8.66 / 26.75
MACHINE-ONLY	21.40 / 4.90	12.22 / 11.60	9.27 / 26.45
UNCERTAINTY-ONLY	20.05 / 5.15	11.23 / 12.35	8.71 / 27.65
$\kappa = 2$ frontier	20.05 / 5.15	11.28 / 12.45	8.68 / 27.10
NO AUDIT	53.59 / 0	53.59 / 0	53.59 / 0

C.10 DECISION PRECISION IN A TWO-CONTEXT MODEL

We isolate the audit-resolution choice with two humans and contexts drawn independently and uniformly from $\{e_1, e_2\}$, so $d = H = 2$. With baseline $a = 0.5$ and human advantage $\delta_h = 0.02$, the first human has parameter $\theta_1^\star = (a, a - \delta_h)$ and the second has $\theta_2^\star = (a - \delta_h, a)$. The machine has base score $b_0(x) = a$ and residual $w^\star = (\Delta_1, -\Delta_2)$. Different employees are best among the humans in the two contexts. The machine's utility minus the best human's is exactly Δ_1 on e_1 and $-\Delta_2$ on e_2. Execution costs are zero and the audit price is $c_R = 0.2$. Each feedback channel has independent Gaussian noise with standard deviation 0.1.

All policies use the stated determinant-based confidence radii with $\lambda = S = 1$ and failure probability $\delta_T = (T + 1)^{-2}$. They share optimistic routing and the confidence-overlap audit rule, with no stopping cap. Only the threshold changes: the cost-tuned configuration uses Corollary 2, the margin-tuned configuration uses $\tau_\Delta = \underline{\Delta}/(2\beta_T)$, and the tolerance-based configurations use $\tau_\epsilon = \epsilon_{\text{policy}}/(2\beta_T)$. We fix $T = 50{,}000$ before observing outcomes. This is the first horizon in $\{10{,}000, 20{,}000, 50{,}000, 100{,}000\}$ exceeding eight times the per-direction update scale $\lceil \tau_{0.05}^{-2} - \lambda \rceil$, giving $\beta_T = 1.8107$. Results average 50 paired seeds with common contexts and potential feedback noises.

The homogeneous instances use $(\Delta_1, \Delta_2) = (0.05, 0.05)$ and $(0.15, 0.15)$; the heterogeneous instance fixes $(0.05, 0.15)$. The recorded artifact additionally reports audit cost, human- and machine-routing losses, both excess-loss and hard-threshold losses at common evaluation tolerances 0.05, 0.10, and 0.20, and paired differences from the cost-tuned configuration. These finite-horizon comparisons examine resolution and acquisition cost; no convergence exponent is estimated.

At margin 0.05, the margin-tuned rule reduces routing loss but its extra audits increase total regret (Table 8). At margin 0.15, the margin-tuned rule has 0.08 higher mean routing regret than the cost-tuned rule, with a 95% paired interval of $[-1.32, 1.48]$ for the difference. For heterogeneous gaps, $\epsilon_{\text{policy}} = 0.20$ improves total regret over cost-tuning (paired difference -74.68 $[-76.35, -73.02]$) and lowers loss at the common evaluation tolerance 0.10. The finer 0.05 and 0.10 thresholds spend more on audits without improving total regret at this horizon.

Table 8: Audit resolution and utility loss in the two-context model. Parentheses are 95% t-interval half-widths over 50 seeds. Total regret includes 0.2 per audit. The final column uses the same evaluation tolerance $\epsilon_{\text{eval}} = 0.10$ for every policy; it is distinct from the policy's threshold input.

Threshold	Audits	Routing	Total regret	$R^{(0.10)}$
Margin 0.05				
Cost-tuned	998.00 (0.00)	1049.42 (20.04)	1249.02 (20.04)	199.60 (0.00)
Margin-tuned	10492.00 (0.00)	109.56 (2.39)	2207.96 (2.39)	2098.40 (0.00)
Margin 0.15				
Cost-tuned	998.00 (0.00)	75.82 (1.77)	275.42 (1.77)	205.76 (0.18)
Margin-tuned	1164.00 (0.00)	75.90 (1.98)	308.70 (1.98)	239.23 (0.18)
Heterogeneous				
Cost-tuned	998.00 (0.00)	78.56 (2.15)	278.16 (2.15)	199.65 (0.00)
$\epsilon_{\text{policy}} = 0.05$	6277.18 (17.64)	111.87 (2.38)	1367.30 (3.59)	1255.49 (3.53)
$\epsilon_{\text{policy}} = 0.10$	2341.56 (17.97)	89.08 (2.33)	557.40 (3.77)	468.37 (3.59)
$\epsilon_{\text{policy}} = 0.20$	654.00 (0.00)	72.68 (2.03)	203.48 (2.03)	130.85 (0.00)

C.11 Early information, coverage, and geometry

We use four policies with the same optimistic router: Delegate-UCB, Symmetric-overlap, Machine-only, and Uncertainty-only. The symmetric rule is defined in Lemma 1. Each stochastic condition uses 50 paired seeds, 2026090700–2026090749, sharing arrivals and pre-sampled Gaussian potential outcomes. Before feedback, we record the route, audit decision, current machine interval, and whether it contains the true utility. We separate audits and routing losses by execution route and record machine uncertainty at the later context. Truth enters only these diagnostics, regret evaluation, and the feedback revealed after the decisions. Theoretical-radius runs use the determinant-based radii in Appendix B with $\delta = (T+1)^{-2}$; empirical-radius runs use $\beta_0 = 0.25$ and $\beta_h = 0.35$.

For the early-information study, we use Proposition 2 with $a \in \{0.5, 0.25, 0.125\}$, $c_R = 0.05$, and Gaussian noise scales $\sigma \in \{0, 0.03\}$. The horizon is 200 for $a = 0.5$ and $100 + \lceil 2/a^2 \rceil$ otherwise. Each sequence contains 100 initial-type contexts and the remaining later-type contexts. We compare their block order with a uniformly random permutation of the same multiset. All other parameters, including $\tau = 0.6a$, stay fixed. In every noiseless block condition, the two overlap policies make exactly two audits with zero routing regret, and Machine-only satisfies the proposition's audit lower bound. Table 9 reports the $a = 0.5$ comparisons. All recorded current-context confidence intervals cover their means throughout these runs. This observed coverage supplements the deterministic proof for $\sigma = 0$.

Table 9: Noise and ordering in the early-information construction ($a = 0.5$). Entries are mean total regret / audit count across 50 paired seeds. Symmetric-overlap and Delegate-UCB coincide in these conditions.

Noise	Order	Delegate-UCB	Machine-only	Uncertainty-only
0	Block	0.100 / 2.00	1.400 / 4.00	0.550 / 11.00
0	Permuted	0.496 / 3.92	1.414 / 4.04	0.906 / 12.12
0.03	Block	0.110 / 2.20	1.883 / 5.38	0.550 / 11.00
0.03	Permuted	0.484 / 3.92	1.876 / 5.36	0.894 / 12.12

The coverage-controlled stress slice uses the vectors and costs in Appendix C.6, with decoy gap 0.8, uniformly sampled basis contexts (decoy mass $5/7$), $T = 1200$, and $\tau \in \{0.3, 0.6\}$. We hold each seed's arrivals and noises fixed while replacing the empirical multipliers by the theoretical radii with $S = 2$. At $\tau = 0.3$, empirical Machine-only has a mean of 170.76 rounds on which a human is selected, the machine is optimal, and its upper score is below its true utility. Delegate-UCB has one such round on average. With theoretical radii, both counts are zero. Table 10 separates these coverage effects from differences in audit placement. The theoretical-radius runs retain lower

DELEGATE-UCB routing regret even though both policies' intervals cover the observed contexts throughout.

Table 10: Stress slice with fixed task and noise streams. Each entry is mean routing regret / audit count over 50 seeds. SYMMETRIC-OVERLAP coincides with DELEGATE-UCB. Audit cost is 0.05 per observation.

Radii	τ	DELEGATE-UCB	MACHINE-ONLY	UNCERTAINTY-ONLY
Empirical	0.3	4.87 / 27.00	26.44 / 10.88	4.87 / 77.00
Theoretical	0.3	17.90 / 76.36	25.87 / 32.00	17.90 / 77.00
Empirical	0.6	7.47 / 11.00	29.47 / 3.00	7.47 / 21.00
Theoretical	0.6	16.98 / 21.00	24.96 / 16.00	16.98 / 21.00

To vary information transfer while fixing utility gaps, we next take $x_A = (1, 0)$ and $x_B = (\rho, \sqrt{1-\rho^2})$ for $\rho \in \{0, 0.5, 0.9\}$. Set $\theta^\star = (1/2, (1-\rho)/(2\sqrt{1-\rho^2}))$ and $w^\star = 0$. The base scores $b_0(x_A) = -0.3$ and $b_0(x_B) = 0.4$ keep human utility at 0.5 and human–machine gaps at 0.8 and 0.1, respectively. We use 100 A tasks followed by 1100 B tasks, $\lambda = S = 1$, $\sigma = 0.03$, $\tau = 0.3$, zero execution costs, and $c_R = 0.05$. Historical-feedback conditions initialize every policy's human ridge model with 20 noiseless observations alternating between A and B; machine initialization is unchanged. Theoretical radii are used throughout.

Table 11: Geometry and shared human history. Entries are mean total regret / audit count over 50 paired seeds. SYMMETRIC-OVERLAP coincides with DELEGATE-UCB.

ρ	Historical samples	DELEGATE-UCB	MACHINE-ONLY	UNCERTAINTY-ONLY
0.0	0	109.021 / 20.22	109.376 / 12.00	109.110 / 22.00
0.5	0	109.047 / 19.22	109.476 / 11.00	109.138 / 21.00
0.9	0	109.165 / 17.22	109.634 / 11.00	109.210 / 19.00
0.0	20	110.653 / 16.90	110.384 / 12.00	110.910 / 22.00
0.5	20	110.648 / 15.40	110.474 / 11.00	110.928 / 21.00
0.9	20	110.683 / 13.98	110.648 / 11.00	111.020 / 19.00

Increasing alignment reduces the audit count of the overlap policies. At the chosen threshold, the 0.1 gap on later tasks remains small relative to residual uncertainty, so routing regret dominates the total and MACHINE-ONLY has a lower mean in the historical-feedback conditions. The threshold is held fixed across geometries rather than calibrated to the margin. All pre-feedback current-context intervals in these runs cover their true means.

To separate geometric information sharing from audit resolution, we repeat the experiment at the finer threshold $\tau_{\text{fine}} = \Delta/(2.2\beta_T) = 0.0380155$. Here $\Delta = 0.1$ and the machine confidence bound is

$$\beta_T = 0.03\sqrt{2\log(1 + 1200/2) + 2\log(2 \cdot 1201^2)} + 1 = 1.1956845.$$

Thus $2\beta_T\tau_{\text{fine}} < \Delta$. The gap is available to the synthetic experiment designer and is not used for replay parameter selection. All other settings, including history, the 50 paired seeds, and potential outcomes, are those of Table 11.

Table 12: Fine-precision geometry experiment, $\tau = 0.0380155$, $T = 1200$, $c_R = 0.05$. Entries are mean routing regret / audit count / total regret over 50 paired seeds. SYMMETRIC-OVERLAP coincides with DELEGATE-UCB in every run. n_H counts historical human observations.

n_H	ρ	DELEGATE-UCB	MACHINE-ONLY	UNCERTAINTY-ONLY
0	0.0	0.20 / 583.86 / 29.39	8.22 / 75.18 / 11.98	0.20 / 791.00 / 39.75
0	0.5	0.10 / 579.20 / 29.06	8.12 / 74.22 / 11.83	0.10 / 791.00 / 39.65
0	0.9	0.10 / 571.88 / 28.69	8.09 / 73.90 / 11.78	0.00 / 787.00 / 39.35
20	0.0	1.56 / 578.30 / 30.48	8.19 / 74.86 / 11.93	1.57 / 791.00 / 41.12
20	0.5	1.50 / 573.14 / 30.16	8.12 / 74.18 / 11.83	1.50 / 791.00 / 41.05
20	0.9	1.46 / 565.78 / 29.75	8.12 / 74.20 / 11.83	1.28 / 787.00 / 40.63

At $\rho = 0$ without history, DELEGATE-UCB's routing regret falls from 108.01 under $\tau = 0.3$ to 0.20 under τ_{fine}. At the finer precision, UNCERTAINTY-ONLY has the same mean routing regret but buys 791 audits rather than 583.86. DELEGATE-UCB's paired total-regret difference is -10.36 $[-10.68, -10.05]$. Across the six correlation/history settings, its mean total-regret reduction relative to UNCERTAINTY-ONLY ranges from 10.36 to 10.90. SYMMETRIC-OVERLAP coincides with DELEGATE-UCB in every run of this experiment.

Relative to MACHINE-ONLY in the orthogonal cold-start setting, DELEGATE-UCB reduces routing regret by 8.02 but incurs 25.43 more audit cost. Its total regret is therefore 17.42 $[17.09, 17.73]$ higher. MACHINE-ONLY has lower mean total regret in all six fine-precision settings, whereas DELEGATE-UCB has lower routing regret. These differences quantify the cost of continuing to resolve overlapping intervals after human routing already has a higher optimistic value.

Without history, the mean uncertainty of the later context just before the second phase is 1.000, 0.880, and 0.519 for $\rho = 0$, 0.5, and 0.9, respectively. These measurements track information transfer from the initial contexts. They are recorded before the current audit update; checkpoints occur at rounds $1, 50, 100, 101, 200, 400, 800, 1200$. All 2400 coarse- and fine-precision trajectories have zero observed interval violations. The paired intervals use 10,000 seed-level bootstrap resamples with seed 2026090719.

Holding the orthogonal cold-start policies, thresholds, and potential feedback fixed, repricing their trajectories gives a mean total-regret difference between DELEGATE-UCB and MACHINE-ONLY of $-8.018 + 508.68c_R$, with a break-even price of approximately 0.01576. This is a fixed-policy cost comparison; selecting new thresholds at each price would define a different comparison.

D IMAGENET16H REAL-CONTEXT OFFLINE REPLAY

The main replay comparison includes matched-horizon episodes (Appendix D.3) and longer episodes with horizon-scaled thresholds (Appendix D.4). Additional analyses examine fixed-threshold transfer and common historical human feedback.

D.1 DATA AND FEEDBACK

ImageNet16H records human classifications and predictions from five CNN families on 1,200 images, 16 classes, and four phase-noise levels $(80, 95, 110, 125)$ (Steyvers et al., 2022). VGG-19 is selected before policy comparison. The behavioral file has 28,997 rows from 145 participants; the machine file has 36,000 rows, 7,200 per CNN. The VGG-19/human inner join covers 4,800 unique image-condition keys without label disagreement or row multiplication. Machine probabilities are in $[0, 1]$, sum to one within 1.45×10^{-15}, and agree with recorded predictions and correctness; human responses also agree with recorded correctness. There are six human responses for 4,603 contexts and seven for 197. Six repeated participant-context keys (12 rows, 0.0207%) represent nonidentical recorded trials and are retained.

A class-stratified split assigns 240 base images (15 per class) to calibration and 960 (60 per class) to evaluation, keeping each image's noise variants together. These sets contain 960 and 3840 contexts, respectively. The main replay uses one pooled-human arm. For the $H = 2$ sensitivity, worker accuracy on calibration images defines 73 high-pool and 72 low-pool workers at the calibration median. Requiring responses from both pools leaves 3741 evaluation contexts (97.42%). In this sensitivity, each human arm represents a calibration-defined pool of workers.

The 21-dimensional context comprises an intercept, VGG-19 top-1 confidence, top-2 margin, normalized entropy, scaled noise, and 16 predicted-class indicators, normalized to norm at most one. The base score is top-1 confidence; audit reveals machine correctness $Z_t \in \{0, 1\}$ and updates the residual with $Z_t - b_0(x_t)$. A human assignment reveals one correctness indicator sampled from that context's recorded pool before drawing the replay order. The learner receives this indicator, not the image label; machine correctness is revealed only by audit. Execution costs are zero. The oracle compares deterministic machine correctness with each pool's recorded mean correctness. The pool mean estimates human utility from the available responses, without extrapolating to the worker population.

All policies share task order and potential human feedback within each of 50 paired seeds. Random schedules have separate streams. We use ridge parameter $\lambda = 1$ and empirical exploration multipliers $\beta_0 = 0.25$, $\beta_h = 0.35$, not calibrated confidence radii. Total regret is $R_{\text{route}} + c_R N_{\text{aud}}$, and mean net utility is $T^{-1}(\sum_t \mu_{j_t}(x_t) - c_R N_{\text{aud}})$. Reported 95% intervals use 10,000 paired seed-bootstrap resamples. The fixed-transfer and historical-feedback comparisons use bootstrap seed 2026090799; the matched-horizon, scaled-threshold, and audit-fraction range comparisons use 2026090719. These pointwise intervals quantify replay stochasticity conditional on the images and selected configurations, without multiple-comparison adjustment or population resampling.

D.2 Calibration and audit rules

Each method selects parameters independently at $c_R \in \{0.01, 0.05, 0.20\}$ using the 960 calibration contexts. Calibration selects configurations, and both calibration and evaluation start from zero ridge coefficients. The historical-feedback comparison instead initializes the human model from calibration observations.

For each price, configurations are evaluated on the same 20 validation seeds, 1000–1019. DELEGATE-UCB, SYMMETRIC-OVERLAP, MACHINE-ONLY, and UNCERTAINTY-ONLY each use $\tau \in \{0.15, 0.25, 0.4, 0.6, 0.9\}$. RANDOM and FRONT-$N$ independently choose an audit fraction $f \in \{0.005, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1.0\}$. They purchase $\text{round}(fT)$ audits, either at uniformly sampled positions or in the first rounds. Each method selects its configuration by minimizing mean validation total regret, breaking ties in favor of the smaller parameter. RANDOM and FRONT-N therefore select their own audit quantities. Configurations are fixed before evaluation on the 50 test seeds, 0–49.

The DMR-POOL baseline adapts the Delta Max Regret criterion of Tucker et al. (2023) to the present feedback protocol. We form a fixed reference set $\mathcal{X}_{\text{cal}}$ of 64 feature centroids by predicted class and noise level, using calibration contexts only. For a residual design matrix V, let $\Phi(V) = \max_{u \in \mathcal{X}_{\text{cal}}} \beta_0 \sqrt{u^\top V^{-1} u}$ be the largest machine exploration width on this set. The audit rule is

$$q_t^{\text{DMR}} = \mathbf{1}\{(T-t)[\Phi(V_t) - \Phi(V_t + x_t x_t^\top)] > \alpha c_R\}.$$

The policy uses the common optimistic router and selects $\alpha \in \{0.1, 0.3, 1, 3, 10\}$ on validation images. Selected-human feedback updates the human model at no additional audit charge; a paid audit updates the machine residual under either route. This finite-pool adaptation uses the original criterion's projected reduction in uncertainty, with a reference geometry available before test-time decisions. All replay methods use the same empirical multipliers and ridge parameter. Selective and DMR-POOL rules run without an additional hard cap; we report their realized counts.

To assess the audit-fraction range, we compare a grid capped at 0.2 with the extended grid above, which includes full auditing. Both grids use the same calibration data, seeds, and objective. RANDOM selects the 0.2 boundary at the two lower prices on the capped grid. At $c_R = 0.01$, the extended grid changes the selected fractions from 0.2 to 0.4 for RANDOM and from 0.1 to 0.3 for FRONT-N; the other selections remain unchanged (Table 13). The method comparisons use the extended-grid selections. Relative to the original grid, their matched-horizon total regret decreases by 3.59 $[1.86, 5.24]$ and 3.75 $[2.36, 5.22]$, respectively. On the continuous 3840-round test, the corresponding extended-minus-original differences are 1.20 $[-4.29, 6.37]$ and 7.32 $[4.63, 9.82]$. A calibration-selected audit fraction can therefore also change its cost balance across horizons.

Table 13: Audit-fraction range check. Fractions minimize mean total regret on the same 20 calibration seeds. The extended grid includes full auditing.

Policy	Grid	$c_R = 0.01$	$c_R = 0.05$	$c_R = 0.20$
RANDOM	Original	0.2	0.2	0.02
RANDOM	Extended	0.4	0.2	0.02
FRONT-N	Original	0.1	0.1	0.05
FRONT-N	Extended	0.3	0.1	0.05

This range comparison was specified after the initial analysis and reuses the same image split. Fractions are selected from calibration outcomes before the additional test runs. The resulting comparisons use the paired-bootstrap protocol above.

The selected thresholds for DELEGATE-UCB, MACHINE-ONLY, and UNCERTAINTY-ONLY are $(0.25, 0.4, 0.6)$ across increasing prices; SYMMETRIC-OVERLAP selects $(0.25, 0.25, 0.6)$. RANDOM selects fractions $(0.4, 0.2, 0.02)$, FRONT-N selects $(0.3, 0.1, 0.05)$, and DMR-POOL selects multipliers $(1, 1, 10)$.

D.3 MATCHED-HORIZON EVALUATION

The matched-horizon evaluation uses the same 240 calibration images, 20 validation seeds, candidate grids, and per-price selections as the calibration protocol above. We partition the 960 test images once into four disjoint groups using seed 2026090717. Each group contains 15 images from each of the 16 classes, together with all four noise variants, giving 960 contexts. For each test seed, we restrict the original full-test permutation and potential human feedback to that group. All policies share these outcomes and the relative arrival order, and each starts from zero ridge coefficients with $V = I$. RANDOM auditing uses a panel-specific schedule stream. Group assignment uses image identity and class; parameter selection uses calibration outcomes.

Let $R_{s,k}$ be the regret in test group k under seed s. We analyze $\bar{R}_s = \frac{1}{4}\sum_{k=1}^{4} R_{s,k}$ and report its mean over 50 seeds. Audit counts and net utility are aggregated in the same order. The estimand is mean regret per 960-round episode with four learner restarts. For policy comparisons, 10,000 paired bootstrap resamples select seed indices from the 50 differences in $\bar{R}_s$, using seed 2026090719. This preserves the dependence among the four groups within a seed.

Table 14: Matched-horizon replay decomposition: mean total regret / audit count per 960-round episode, averaging four test episodes within each of 50 paired seeds. Configurations are selected on calibration data.

Policy	$c_R = 0.01$	$c_R = 0.05$	$c_R = 0.20$
DELEGATE-UCB	70.45 / 302.83	78.88 / 107.68	95.83 / 31.54
SYMMETRIC-OVERLAP	77.63 / 178.25	84.76 / 178.25	98.13 / 26.86
MACHINE-ONLY	72.39 / 278.71	80.13 / 98.39	95.60 / 29.00
UNCERTAINTY-ONLY	71.50 / 321.05	79.43 / 111.92	95.84 / 32.02
RANDOM	77.43 / 384.00	88.70 / 192.00	106.16 / 19.00
FRONT-N	71.31 / 288.00	78.91 / 96.00	93.69 / 48.00
DMR-POOL	71.92 / 232.97	80.94 / 126.42	101.48 / 13.50

At $c_R = 0.01$, the DELEGATE-UCB-minus-MACHINE-ONLY difference is -1.94 $[-2.56, -1.32]$, the difference from FRONT-N is -0.86 $[-1.47, -0.22]$, and the difference from DMR-POOL is -1.46 $[-2.23, -0.70]$. At $c_R = 0.05$, these differences are -1.25 $[-2.17, -0.35]$, -0.03 $[-1.48, 1.37]$, and -2.06 $[-3.87, -0.27]$. At $c_R = 0.20$, the differences from MACHINE-ONLY and FRONT-N are 0.23 $[-0.95, 1.42]$ and 2.14 $[-0.15, 4.51]$, respectively; both intervals include zero. Supplementary records provide all paired comparisons.

At $c_R = 0.01$, DELEGATE-UCB's total-regret difference from RANDOM is -6.98 $[-8.11, -5.82]$. Relative to UNCERTAINTY-ONLY, the differences at $c_R = 0.01$ and 0.05 are -1.05 $[-1.49, -0.62]$ and -0.55 $[-1.13, -0.04]$. DELEGATE-UCB and UNCERTAINTY-ONLY share the routing and calibration settings, so this comparison measures the effect of overlap filtering.

D.4 TRANSFERRING PRECISION ACROSS HORIZONS

Fixed-threshold transfer applies the calibration-selected numerical thresholds to the continuous 3840-round test.

Table 15: Fixed-threshold transfer to $T = 3840$: mean total regret / audit count over 50 paired seeds. Parameters are selected independently for each method and price using calibration images only.

Policy	$c_R = 0.01$	$c_R = 0.05$	$c_R = 0.20$
DELEGATE-UCB	253.64 / 344.44	276.54 / 114.78	335.88 / 33.00
SYMMETRIC-OVERLAP	266.55 / 288.04	278.07 / 288.04	341.71 / 31.46
MACHINE-ONLY	260.71 / 327.74	279.06 / 106.14	328.79 / 30.48
UNCERTAINTY-ONLY	256.97 / 373.12	278.67 / 121.96	335.49 / 34.08
RANDOM	271.05 / 1536	300.57 / 768	351.07 / 77
FRONT-N	261.98 / 1152	270.02 / 384	299.71 / 192
DMR-POOL	256.80 / 659.62	270.73 / 315.02	329.32 / 44.64

Table 15 reports the resulting test outcomes. At the lowest price, the paired DELEGATE-UCB-minus- MACHINE-ONLY difference is -7.07 $[-11.02, -3.07]$, and the difference from SYMMETRIC-OVERLAP is -12.90 $[-21.34, -5.13]$. At $c_R = 0.05$, FRONT-N and DMR-POOL have lower mean regret than DELEGATE-UCB, with respective DELEGATE-UCB-minus-baseline intervals 6.53 $[-0.19, 13.50]$ and 5.82 $[-0.06, 12.28]$. At $c_R = 0.20$, FRONT-N improves on DELEGATE-UCB by 36.17 $[19.51, 53.19]$. The selective methods transfer fixed numerical thresholds from 960 calibration rounds to 3840 test rounds; FRONT-N and RANDOM transfer audit fractions, while DMR-POOL uses the remaining horizon. For the selective methods, we also evaluate horizon-scaled thresholds without retuning configurations.

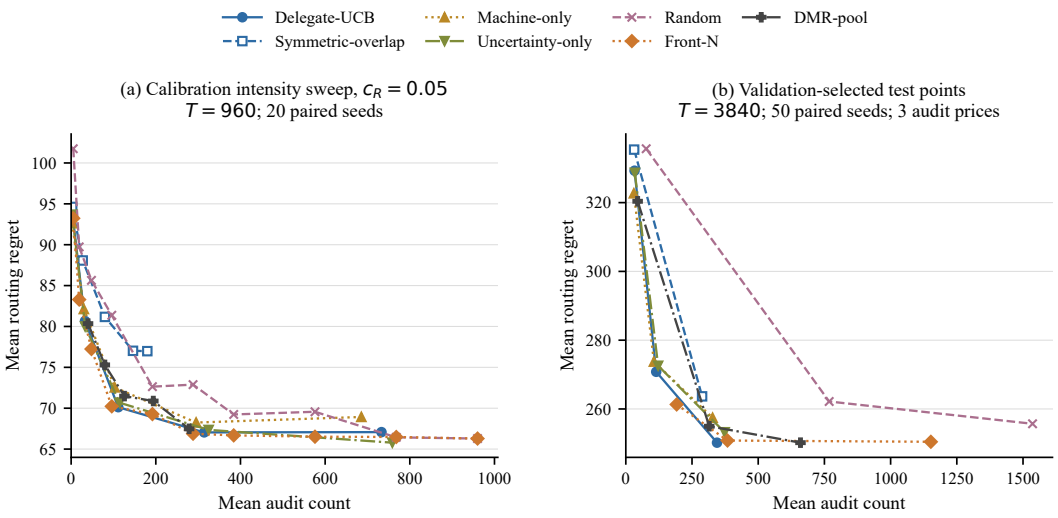


Figure 4: Audit quantity and routing loss under fixed-threshold transfer on ImageNet16H. Panel (a) shows candidate configurations on calibration images at $c_R = 0.05$: ten audit fractions for Random and Front-N, and five configurations for each other method. Panel (b) shows only test configurations selected on calibration images for $c_R \in \{0.01, 0.05, 0.20\}$. Points are seed means, and lines connect observed configurations; repeated selected configurations coincide. The panels have different horizons (960 and 3840 contexts). Paired total-regret intervals and the price-specific operating points are reported in the text and Table 15.

In the unclipped budget balance of Corollary 2, τ_T scales as $(c_R \Lambda_T / (\beta_T T))^{1/3}$. We use this dependence to specify a threshold transfer rule before the scaled test runs. With the replay's fixed exploration multipliers and $d = 21$, $\lambda = 1$, the common price and constants cancel, giving

$$\tau_{\text{test}} = \tau_{\text{val}} \left[\frac{\log(1 + T_{\text{test}}/21)/T_{\text{test}}}{\log(1 + T_{\text{val}}/21)/T_{\text{val}}} \right]^{1/3}.$$

For $T_{\text{val}} = 960$ and $T_{\text{test}} = 3840$, the factor is 0.6973408142. We apply it to all four selective methods. The extended-grid calibration selections for RANDOM and FRONT-N are transferred as audit fractions; the DMR-POOL multiplier remains unchanged. The fixed- and scaled-threshold runs

share test order, potential outcomes, costs, and cold starts. The scaling rule applies the theoretical horizon dependence using the replay's fixed exploration multipliers and approximate linear model.

Table 16: Long-horizon replay with prespecified threshold scaling: mean total regret / audit count over 50 paired seeds, $T = 3840$. Each selective threshold is multiplied by 0.69734; the selected fractions and DMR-POOL multiplier are unchanged.

Policy	$c_R = 0.01$	$c_R = 0.05$	$c_R = 0.20$
DELEGATE-UCB	252.53 / 718.70	265.77 / 271.56	303.46 / 102.52
SYMMETRIC-OVERLAP	262.25 / 505.40	282.46 / 505.40	310.52 / 94.60
MACHINE-ONLY	260.30 / 688.76	272.81 / 257.10	297.20 / 94.08
UNCERTAINTY-ONLY	260.85 / 779.96	270.52 / 293.28	302.22 / 108.18
RANDOM	271.05 / 1536.00	300.57 / 768.00	351.07 / 77.00
FRONT-N	261.98 / 1152.00	270.02 / 384.00	299.71 / 192.00
DMR-POOL	256.80 / 659.62	270.73 / 315.02	329.32 / 44.64

At $c_R = 0.20$, scaling changes DELEGATE-UCB's audit count from 33.00 to 102.52 and its routing regret from 329.28 to 282.95. After the extra audit cost, total regret falls from 335.88 to 303.46; the paired scaled-minus-fixed difference is -32.43 $[-46.14, -19.49]$. The scaled DELEGATE-UCB-minus-FRONT-N difference is 3.75 $[-8.43, 16.62]$, whose interval includes zero. At $c_R = 0.05$, scaled DELEGATE-UCB has total regret 265.77, with differences -7.04 $[-10.69, -3.35]$ from MACHINE-ONLY and -4.25 $[-7.68, -0.68]$ from FRONT-N.

The matched-horizon and scaled-threshold protocols were specified after the initial fixed-transfer analysis and retain its evaluation images. All configuration choices remain restricted to calibration outcomes; the paired intervals also condition on the specified transfer rule.

Resampling the original 20 validation seeds selects DELEGATE-UCB's $\tau = 0.6$ at $c_R = 0.20$ in 96.48% of 10,000 bootstrap replicates. At $c_R = 0.05$, $\tau = 0.4$ is selected in 99.27%. The bootstrap measures selection stability within the fixed calibration pool; it neither resamples images nor changes the configurations used in evaluation.

D.5 REFERENCE REPLAY AND ORGANIZATIONAL HISTORY

The reference replay fixes $c_R = 0.05$ and $B = \mathrm{round}(0.2T)$, and uses the empirical audit rule with $\tau_B = \sqrt{2d \log(1 + T/d)/B}$ (0.533994 for $T = 3840$). FULL-BUDGET RANDOM selects exactly B positions; MATCHED RANDOM uses the seed-specific realized DELEGATE-UCB count and is a post-hoc placement control. No selective trajectory reaches the cap. Tables 17 and 18 report the one- and two-pool results; accuracy is tie-aware.

For $H = 1$, DELEGATE-UCB's paired total-regret difference from NO AUDIT is -144.32 $[-158.37, -128.85]$. The paired intervals for comparisons with the other selective rules include zero, as do those for MATCHED RANDOM (-18.72 $[-37.90, 1.68]$) and FULL-BUDGET RANDOM (14.22 $[-0.07, 29.78]$). FULL-BUDGET RANDOM has the lowest observed mean. For $H = 2$, DELEGATE-UCB improves over NO AUDIT by 180.10 $[162.93, 195.49]$ and over MATCHED RANDOM by 51.99 $[28.35, 74.58]$; the paired intervals for comparisons with UNCERTAINTY-ONLY, MACHINE-ONLY, and FULL-BUDGET RANDOM include zero. This sensitivity excludes 99 contexts lacking both pools.

Method	R_{route}	R_{total}	N_{aud}	Routing acc. (%)
DELEGATE-UCB	315.41 [301.65, 330.80]	317.87 [303.69, 333.25]	49.2 [48.4, 49.9]	85.8 [85.4, 86.3]
UNCERTAINTY-ONLY	313.15 [299.07, 328.55]	315.73 [301.80, 331.16]	51.6 [51.2, 52.0]	85.9 [85.4, 86.4]
MACHINE-ONLY	308.94 [296.99, 321.89]	311.18 [299.39, 324.15]	44.8 [43.7, 45.9]	86.1 [85.7, 86.5]
NO AUDIT	462.19 [462.19, 462.19]	462.19 [462.19, 462.19]	0.0 [0.0, 0.0]	81.7 [81.7, 81.7]
FULL-BUDGET RANDOM	265.25 [260.26, 270.77]	**303.65** **[298.73, 309.28]**	768.0 [768.0, 768.0]	86.6 [86.4, 86.8]
MATCHED RANDOM	334.14 [322.78, 346.07]	336.59 [325.18, 348.67]	49.2 [48.4, 49.9]	85.2 [84.8, 85.5]

Table 17: ImageNet16H real-context offline replay with one pooled-human arm and audit price $c_R = 0.05$. Entries are means with 95% intervals from 10,000 paired seed-level bootstrap resamples. The intervals quantify replay stochasticity, not population sampling uncertainty. Bold marks the lowest observed mean total regret, not a resolved ranking.

Method	R_{route}	R_{total}	N_{aud}	Routing acc. (%)
DELEGATE-UCB	352.78 [337.18, 370.01]	355.15 [339.80, 372.23]	47.4 [46.6, 48.2]	85.8 [85.3, 86.2]
UNCERTAINTY-ONLY	351.96 [336.20, 369.05]	354.46 [338.81, 371.14]	50.1 [49.7, 50.4]	85.8 [85.3, 86.2]
MACHINE-ONLY	351.13 [336.08, 367.41]	353.21 [338.17, 369.03]	41.7 [40.3, 43.0]	85.8 [85.4, 86.2]
NO AUDIT	535.25 [535.25, 535.25]	535.25 [535.25, 535.25]	0.0 [0.0, 0.0]	81.8 [81.8, 81.8]
FULL-BUDGET RANDOM	305.30 [294.75, 316.61]	**342.70** **[332.35, 353.75]**	748.0 [748.0, 748.0]	86.5 [86.1, 86.9]
MATCHED RANDOM	404.76 [387.73, 422.17]	407.13 [390.04, 424.41]	47.4 [46.6, 48.2]	84.3 [83.9, 84.8]

Table 18: Sensitivity analysis with two human pools defined only on calibration images and audit price $c_R = 0.05$; contexts missing either pool are excluded. Entries are means with 95% intervals from 10,000 paired seed-level bootstrap resamples. The intervals quantify replay stochasticity, not population sampling uncertainty. Bold marks the lowest observed mean total regret, not a resolved ranking.

With zero coefficients and $V_{j,1} = I$, NO AUDIT selects the machine on all tasks in all 50 seeds and receives no human feedback. ALWAYS-HUMAN uses the pooled-human arm throughout; the oracle's mean utility is 0.91594. The history sensitivity initializes each policy's human ridge model from the same 960 calibration contexts and recorded mean human correctness, leaving the machine residual at zero. Threshold, costs, evaluation images, and potential outcomes are unchanged. With this common human history, DELEGATE-UCB has lower total regret than NO AUDIT: the paired difference is -110.27 $[-117.45, -102.77]$ (Table 19).

Table 19: Frozen-threshold replay with and without common historical human feedback. Means over 50 paired seeds; $T = 3840$, $c_R = 0.05$. Machine assignments also determine the number of human observations, T minus that count.

Human history	Policy	Total regret	Audits	Machine routes	Net utility
None	DELEGATE-UCB	317.87	49.16	3082.16	0.8332
	SYMMETRIC-OVERLAP	323.11	46.68	3268.02	0.8318
	MACHINE-ONLY	311.18	44.82	3077.20	0.8349
	UNCERTAINTY-ONLY	315.73	51.62	3075.92	0.8337
	NO AUDIT / ALWAYS-MACHINE	462.19	0	3840	0.7956
	ALWAYS-HUMAN	488.12	0	0	0.7888
Calibration	DELEGATE-UCB	302.99	43.64	2909.98	0.8370
	SYMMETRIC-OVERLAP	304.78	41.96	2904.36	0.8366
	MACHINE-ONLY	307.82	39.84	2915.70	0.8358
	UNCERTAINTY-ONLY	296.54	51.62	2924.06	0.8387
	NO AUDIT	413.25	0	3563.22	0.8083

An independent implementation reproduces all 50 original frozen DELEGATE-UCB total-regret values to numerical precision. Supplementary records retain candidate scores, selected parameters, test outcomes, route-specific audits, and paired comparisons.

The replay combines recorded images, noise conditions, labels, CNN outputs, and human responses with simulated arrival order, response sampling, and paid selective revelation. It evaluates selective calibration on these recorded contexts using approximate low-dimensional ridge models, without establishing exact realizability or confidence coverage. The results concern this replay protocol, not causal audit effects, online deployment performance, or population generalization.

E SCOPE, BROADER IMPACTS, AND REPRODUCIBILITY

E.1 MODEL SCOPE AND EXTENSIONS

The theoretical guarantees assume linear realizability, conditionally sub-Gaussian noise, bounded contexts, and a stationary residual-correction model. Audits update estimates of fixed machine quality, while organizational learning improves allocation. Review that changes the current outcome would require its benefit to enter both the action value and the oracle's decision; feedback that changes future capabilities would make utilities history-dependent. Nonstationary humans, changing machine models, and strategic users likewise extend the present setting. Unknown horizons can be handled by a standard doubling trick, recomputing the confidence level, budget target, and threshold at the start of each epoch.

The nonlinear sensitivity experiment in Appendix C.7 examines a departure from linear realizability: when the omitted component has the same standard deviation as the modeled linear residual, DELEGATE-UCB has lower mean total regret than NO AUDIT but higher mean total regret than

MACHINE-ONLY and FULL-BUDGET RANDOM audit. This comparison shows how model error affects audit selection; the theoretical guarantees apply under the stated realizability conditions.

E.2 BROADER IMPACTS

Targeting calibration-relevant cases can reduce unnecessary expert review. The same targeting can be harmful if a misspecified model directs review away from underrepresented or high-risk regions. High-stakes use should add domain-specific safety constraints, subgroup and fairness checks, monitoring for distribution shift, and human oversight.

E.3 ARTIFACTS AND COMPUTATIONAL RESOURCES

The anonymized controlled-simulation artifact uses synthetic or LLM-assisted tasks, with no private user data. Its code is available at `https://anonymous.4open.science/r/Delegate_UCB-242D`. The controlled simulation is reproduced from the cached task table without GPU use or external API calls. On a 64-bit Windows machine with an Intel 11th Gen Core i7-class processor, 4 cores/8 threads, 15.7 GiB RAM, and Python 3.11.4, the run covering 3 policies, 6 budgets, 20 seeds, and 1000 rounds took 92.1 seconds and peaked at 316 MiB working set. Regenerating the figures took 8.8 seconds and peaked at 443 MiB.

The cached task table was generated with `qwen3.6-plus`. The CPU-only sensitivity experiments evaluate five methods across four one-factor axes and produce 7700 seed-level records. Runtime depends on the host and multiprocessing configuration. The validation script checks cell coverage, regret decomposition, audit-cost accounting, exact full-random counts, cap binding, and constructed frontier-mass targets. Exploratory development compute outside the reproducible runs was not tracked.

The horizon and precision supplement contains 7650 trajectory records and 19,200 geometry checkpoints, together with the fixed protocols, calibration selections, paired comparisons, and validation checks. It includes both fixed- and scaled-threshold results. The audit-fraction range check adds 400 calibration trajectories and 1500 test trajectories, preserving the original candidate scores and the extended-grid selections. Fixed-fraction calibration trajectories are repriced at the three audit costs, giving 1200 price-indexed records.