CERTIFYING LLM SERVING UPDATES WITH COST-OPTIMAL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When an LLM serving pool changes, an operator must decide whether to retain an incumbent, replace it with a candidate, or combine them in a cascade. Evaluating all models on common queries can itself be costly. We formulate this pre-deployment decision as fixed-confidence identification with costly evaluation actions, separating the serving configuration to be selected from the observations used to select it. The characteristic certification cost $C^\star(\theta)$ quantifies the minimum cost of distinguishing the true environment from alternatives that change the optimal serving decision. It yields an instance-dependent lower bound, which our TRACK-AND-CERTIFY procedure asymptotically attains as $\delta \to 0$ under finite-class, stationary, identifiable observation models with known costs and a unique true optimum. For LLM serving, replacement depends on marginal performance, whereas a cascade triggered by verified incumbent failure also depends on same-query rescue behavior. Controlled replay across 11 model pools shows that richer multi-candidate feedback reduces certification cost only on selected pools, and that the characteristic-cost ratio strongly predicts this realized benefit. Decision-aware singleton allocation reduces median certification cost by $8.53\times$ relative to uniform singleton evaluation. These results motivate designing model evaluations around the serving decision they must resolve.

1 INTRODUCTION

Large language model (LLM) serving increasingly relies on *model selection* to trade response quality against inference cost. Routing methods select a model for each query, while cascading methods may invoke additional models when the current response is deemed insufficient (Gupta et al., 2024; Ong et al., 2025; Dekoninck et al., 2025). These methods largely study how a given model pool should serve queries. In practice, however, the pool itself changes as new models become available or serving costs change. Before adopting such a change, an operator must determine whether the incumbent should be retained, replaced by a candidate, or augmented by a verifier-triggered cascade. Full-feedback evaluation could resolve this decision by running all relevant models on common queries, but obtaining such feedback can itself require substantial inference compute (Tsiourvas et al., 2025). *This paper asks: which model evaluations should be purchased to determine the optimal serving configuration?*

The key distinction is between the object ultimately deployed and the measurements used to select it. A *serving configuration* specifies how models will serve future queries, whereas an *evaluation action* specifies which model outcomes are observed before deployment. For an incumbent B and candidate j, replacement depends on their marginal performance and relative serving cost, whereas a verifier-triggered cascade additionally depends on whether j succeeds on queries for which B fails. Singleton and paired same-query evaluations can therefore have different information value for the deployment decision. This places pre-deployment serving evaluation within controlled sensing and fixed-confidence pure exploration, where experiments are chosen adaptively to resolve a terminal decision (Nitinawarat et al., 2013; Deshmukh et al., 2021; Fiez et al., 2019; Poiani et al., 2024). The LLM setting adds a specific observation structure: evaluation actions reveal different projections of a shared same-query model-outcome law, and those projections need not have the same value for keep, replace, and cascade decisions.

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.

We formulate this problem as fixed-confidence identification with costly evaluation actions. A certifier adaptively chooses which outcomes to observe and must return an optimal serving configuration with probability at least $1 - \delta$, while minimizing expected evaluation cost. It only needs to rule out environments under which the optimal serving decision changes. This leads to the *characteristic certification cost* $C^\star(\theta)$, the cost of an oracle information allocation that separates the true environment from all decision-changing alternatives. Every δ-correct procedure satisfies $\mathbb{E}_\theta[C_\tau] \geq \mathrm{kl}(1-\delta, \delta) C^\star(\theta)$. Our TRACK-AND-CERTIFY procedure tracks the characteristic allocation and attains $\mathbb{E}_\theta[C_{\tau_\delta}] \sim C^\star(\theta) \log(1/\delta)$ as $\delta \to 0$ under the assumptions of Theorem 2. The analysis connects the max–min information allocation of controlled sensing with fixed-confidence allocation tracking (Nitinawarat et al., 2013; Deshmukh et al., 2021; Kaufmann et al., 2016; Garivier and Kaufmann, 2016; Jedra and Proutiere, 2020; Wang et al., 2021).

The LLM specialization predicts selective value from richer feedback: singleton evaluation already reveals each incumbent–candidate same-query outcome, while multi-candidate observations can still reduce the cost of ruling out decision-changing alternatives. Controlled replay on 11 frozen model pools supports this distinction and shows a broader gain from decision-aware allocation. A characteristic statistic also predicts the useful evaluation family on held-out-task replay.

Our contributions are:

- **Formulation and characteristic cost.** We cast LLM serving updates as fixed-confidence identification with distinct serving configurations and costly evaluation actions. A cost-weighted information design characterizes first-order certification cost and yields identifiability, duality, and pruning results.
- **Adaptive certification.** TRACK-AND-CERTIFY tracks the characteristic allocation, is δ-correct, and asymptotically attains this cost under finite-class assumptions; independent auxiliary data can select a valid evaluation family before certification.
- **LLM evidence.** Controlled replay on 11 frozen model pools shows broad gains from decision-aware allocation, selective value from richer same-query feedback, and held-out prediction of the useful evaluation family by the characteristic-cost ratio.

2 RELATED WORK

LLM model selection and costly evaluation. Cost-aware LLM inference has been studied through routing, which selects a model for each query, and cascading, which may invoke additional models when an initial response is insufficient (Chen et al., 2023; Gupta et al., 2024; Yue et al., 2024; Ong et al., 2025; Dekoninck et al., 2025). These methods primarily optimize how a given model pool should serve queries. Recent routing work has also highlighted the cost of *full-feedback* data, in which multiple candidate models are evaluated on common queries (Tsiourvas et al., 2025). Active testing more broadly treats evaluation feedback as a costly acquisition resource (Kossen et al., 2021). These works motivate selective evaluation, but our objective is different: we use evaluation outcomes to certify which serving configuration should be deployed, rather than to estimate test risk or optimize query-time routing.

Fixed-confidence identification and adaptive experimental design. Fixed-confidence pure exploration characterizes instance-dependent information requirements for identifying a correct decision and motivates procedures that learn and track informative sampling proportions (Kaufmann et al., 2016; Garivier and Kaufmann, 2016; Jedra and Proutiere, 2020; Wang et al., 2021). Sequential controlled sensing studies the closely related problem of adaptively choosing observation controls for simple and composite hypothesis discrimination, including first-order asymptotically optimal sequential policies (Nitinawarat et al., 2013; Deshmukh et al., 2021). Transductive experimental design separates the available measurements from the objects to be identified (Fiez et al., 2019), while multi-fidelity identification incorporates observations with heterogeneous cost and informativeness (Poiani et al., 2024). Our serving problem brings these ideas to a setting in which deployment configurations and evaluation actions are distinct and each action reveals a different projection of the same model-outcome law. The resulting information geometry determines whether singleton, paired, or full-feedback evaluation is worth purchasing for a particular serving decision.

3 LLM SERVING UPDATES AND COSTLY EVALUATION

3.1 SERVING CONFIGURATIONS AND EVALUATION ACTIONS

We consider a serving-update problem indexed by an unknown environment $\theta \in \Theta$, where Θ is a finite environment class. Let Π be a finite set of *serving configurations* available after the update. Each configuration $\pi \in \Pi$ has expected utility $J_\pi(\theta) \in \mathbb{R}$. The certification task is to identify an element of $\Pi^\star(\theta) = \arg\max_{\pi\in\Pi} J_\pi(\theta)$. When the optimum is non-unique, we use a fixed deterministic selector $\pi^\circ(\theta) \in \Pi^\star(\theta)$. Our asymptotic analysis assumes that the true optimum is unique, denoted $\pi^\star(\theta)$.

To identify the optimal serving configuration, the certifier acquires evidence through a finite set of *evaluation actions* $\mathcal{Z}$. Each action $z \in \mathcal{Z}$ incurs evaluation cost $h_z > 0$ and produces an observation $O \sim P_\theta^z$ under environment θ. Serving configurations and evaluation actions play different roles: $J_\pi(\theta)$ may already account for the *serving* cost incurred after deployment, whereas h_z is the *evaluation* cost paid to acquire information before deployment. Evaluation actions may therefore differ both in the information they reveal and in the cost required to obtain it. In the LLM setting, an evaluation action may run a single model (*singleton evaluation*), two models on the same query (*paired same-query evaluation*), or the full model pool (*full-feedback evaluation*). Figure 1 previews how these evaluation actions support a certified serving update.

3.2 FIXED-CONFIDENCE CERTIFICATION

A certification procedure adaptively selects actions, stops at a time τ, and outputs a serving configuration $\widehat{\pi}$. At round t, its choice Z_t is measurable with respect to $\mathcal{F}_{t-1} = \sigma(Z_1, O_1, \ldots, Z_{t-1}, O_{t-1})$. Given this history and $Z_t = z$, the next observation follows P_θ^z; observations are fresh and conditionally stationary (Appendix A.1). Let $N_z(t) = \sum_{s\le t} \mathbf{1}\{Z_s = z\}$ count action selections. The cumulative evaluation cost is $C_t = \sum_z h_z N_z(t)$, and $\widehat{\pi}$ is $\mathcal{F}_\tau$-measurable.

A certification procedure is δ-correct if, for every $\theta \in \Theta$,

$$\mathbb{P}_\theta(\widehat{\pi} \notin \Pi^\star(\theta)) \le \delta. \tag{1}$$

Among δ-correct certification procedures, our objective is to minimize the expected evaluation cost $\mathbb{E}_\theta[C_\tau]$ for each environment θ. Crucially, certification need not recover the full environment: it only needs enough evidence to rule out environments under which the optimal serving configuration would change.

3.3 LLM SPECIALIZATION: SINGLE MODELS AND CASCADES

We now instantiate the abstract certification problem for an LLM serving update in which an incumbent model may be retained, replaced by a candidate, or combined with a candidate in a verifier-triggered cascade. Let B denote the incumbent model and let $j \in [K]$ index the candidate models. For a common query, let $Y = (Y_B, Y_1, \ldots, Y_K) \in \{0,1\}^{K+1}$ record the correctness outcomes of the incumbent and candidate models. We allow arbitrary dependence among these same-query outcomes. Let κ_B and κ_j denote their serving costs, and let $\eta > 0$ control the quality–cost tradeoff.

We restrict attention to three deployment choices. Their per-query utilities are

$$u_{\pi_B}(Y) = Y_B - \eta\kappa_B, \tag{2}$$

$$u_{\pi_j}(Y) = Y_j - \eta\kappa_j, \tag{3}$$

$$u_{\pi_{B\to j}}(Y) = Y_B + (1 - Y_B)Y_j - \eta[\kappa_B + (1 - Y_B)\kappa_j]\,. \tag{4}$$

Here π_B retains the incumbent, π_j replaces it with candidate j, and $\pi_{B\to j}$ is a verifier-triggered cascade that first serves with B and invokes j when an observable verifier identifies an incumbent failure. Accordingly, the cascade pays κ_j only on such queries. This abstraction models automatically verifiable settings and our controlled replay.

These utilities expose the distinct statistical quantities that determine the two possible updates. Writing $J_\pi(\theta) = \mathbb{E}_\theta[u_\pi(Y)]$, their gains relative to incumbent-only serving are

$$\Delta_j^{\mathrm{cand}} = \mathbb{E}[Y_j - Y_B] - \eta(\kappa_j - \kappa_B), \tag{5}$$

$$\Delta_j^{\mathrm{cas}} = \mathbb{P}(Y_B = 0)\,[\mathbb{P}(Y_j = 1 \mid Y_B = 0) - \eta\kappa_j]\,. \tag{6}$$

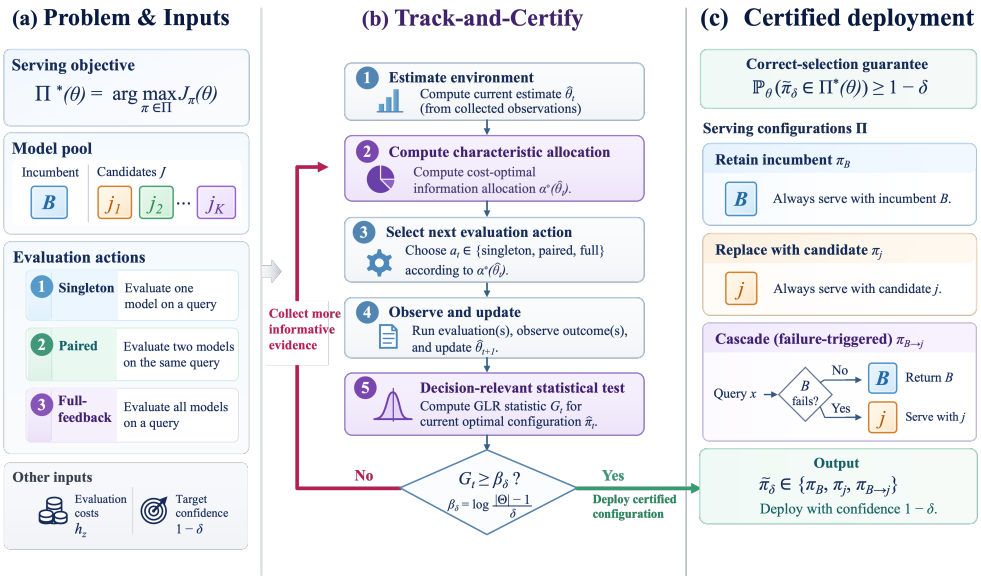


Figure 1: **Decision-aware certification for an LLM serving update.** TRACK-AND-CERTIFY buys informative evaluations and stops when a serving configuration is certified.

Replacement depends on marginal performance and relative serving cost, while cascade value additionally depends on the candidate's rescue probability conditional on incumbent failure. Singleton observations may therefore suffice for replacement, whereas same-query observations can provide additional information for cascade certification.

4 CHARACTERISTIC CERTIFICATION COST

This section characterizes the minimum evaluation cost required for fixed-confidence certification. The key object is an instance-dependent information design that allocates evaluation effort to distinguish the true environment from alternatives that would change the optimal serving configuration. It is a cost-weighted version of the max–min information designs that underlie controlled sensing and fixed-confidence pure exploration.

4.1 CONFIGURATION-CHANGING ALTERNATIVES

Certification does not need to distinguish the true environment θ from every other environment. Only alternatives under which its optimal serving configuration is no longer optimal can lead to a different deployment decision. Accordingly, for a true environment θ with unique optimum, define

$$\mathrm{Alt}(\theta) = \{\lambda \in \Theta : \pi^\star(\theta) \notin \Pi^\star(\lambda)\}\,. \tag{7}$$

One use of action z supplies expected log-likelihood evidence $d_z(\theta, \lambda) = D_{\mathrm{KL}}(P_\theta^z \| P_\lambda^z)$ against λ. Here D_{KL} denotes KL divergence, and kl its Bernoulli specialization. An oracle experiment chooses how much evaluation effort to allocate to each action. To rule out every decision-changing alternative, the accumulated information must reach at least one unit against each $\lambda \in \mathrm{Alt}(\theta)$. The cheapest such experiment defines the *characteristic certification cost*

$$C^\star(\theta) = \min_{n \geq 0} \left\{ \sum_z h_z n_z : \sum_z n_z d_z(\theta, \lambda) \geq 1,\ \forall \lambda \in \mathrm{Alt}(\theta) \right\}. \tag{8}$$

The optimizer $n^\star \in \mathbb{R}_+^{|\mathcal{Z}|}$ is a continuous allocation of evaluation effort per unit of required evidence, not an integer sample count. Environments sharing the true optimal configuration impose no constraint.

4.2 A FUNDAMENTAL LOWER BOUND

The same quantity also lower-bounds the evaluation cost of every fixed-confidence certifier.

Theorem 1 (Fundamental certification lower bound). *For any $\delta \in (0, 1/2)$ and any δ-correct certifier,*

$$\mathbb{E}_\theta[C_\tau] \geq \mathrm{kl}(1-\delta, \delta)\, C^\star(\theta). \tag{9}$$

Consequently, the expected cost is at least $C^\star(\theta)\log(1/\delta)$ to first order as $\delta \downarrow 0$.

Thus $C^\star(\theta)$ is the first-order instance complexity of certification: as the target error probability vanishes, no procedure can achieve expected evaluation cost smaller than $C^\star(\theta)\log(1/\delta)$ to leading order. The proof is given in Appendix B.1.

4.3 STRUCTURAL CONSEQUENCES AND A CASCADE SPECIALIZATION

The characteristic design also reveals when certification is possible and which evaluation actions can be useful. A first consequence is a necessary *decision-identifiability* condition. For every $\lambda \in \mathrm{Alt}(\theta)$, some $z \in \mathcal{Z}$ must satisfy $P^z_\theta \neq P^z_\lambda$. Otherwise, the two environments induce identical observation laws under every adaptive evaluation policy despite requiring different serving decisions, making δ-correct certification impossible for $\delta < 1/2$.

For a finite alternative set, Eq. (8) is a linear program with dual

$$C^\star(\theta) = \max_{w \geq 0} \sum_{\lambda \in \mathrm{Alt}(\theta)} w_\lambda \quad \text{s.t.} \quad \sum_{\lambda} w_\lambda d_z(\theta, \lambda) \leq h_z,\ \forall z. \tag{10}$$

The dual assigns weight to decision-changing alternatives subject to an information budget for each evaluation action. It therefore identifies the weighted collection of alternatives that is hardest to rule out per unit evaluation cost. The same representation yields a simple pruning rule.

Proposition 1 (Cost–Blackwell pruning). *If evaluation action z_2 Blackwell-dominates z_1 for every environment and $h_{z_2} \leq h_{z_1}$, then there exists an optimal characteristic design with $n^\star_{z_1} = 0$.*

Decision identifiability, LP duality, and the pruning result are proved in Appendix B.2.

Identifiability versus information efficiency. All retain/replace/cascade utilities are determined by the singleton laws $\{P_\theta(Y_B, Y_j)\}_{j=1}^K$, so every decision-changing alternative differs under some singleton action. Richer feedback can still add conditional information:

$$d_{jk}(\theta, \lambda) = d_j(\theta, \lambda) + \mathbb{E}_{P_\theta(Y_B, Y_j)} D_{\mathrm{KL}}(P_\theta(Y_k \mid Y_B, Y_j) \| P_\lambda(Y_k \mid Y_B, Y_j))\,. \tag{11}$$

Thus multi-candidate feedback can reduce certification cost without changing what is identifiable, a distinction isolated experimentally in Section 6.

The characteristic cost also makes explicit which aspects of an LLM serving instance make certification difficult. Consider one candidate and a restricted verifier-triggered cascade family. Let $\rho = \mathbb{P}(Y_B = 0)$ and $q = \mathbb{P}(Y_j = 1 \mid Y_B = 0)$, suppose the cascade is preferred iff $q > c$, and let a paired same-query evaluation cost h. If alternatives change only q, the closest configuration-changing environment lies at the decision boundary $q' = c$, giving

$$C^\star(q) = \frac{h}{\rho\ \mathrm{kl}(q, c)} = \Theta\!\left(\frac{h}{\rho(q-c)^2}\right) \qquad (q \to c). \tag{12}$$

Hence certification becomes expensive for two distinct reasons: the informative event $Y_B = 0$ is rare ($\rho \downarrow 0$), or the cascade is close to the deployment boundary ($q \to c$). A derivation is given in Appendix B.3.

5 COST-OPTIMAL CERTIFICATION

Section 4 characterizes the cost-optimal experiment when the environment is known. Following the tracking-and-stopping paradigm of fixed-confidence pure exploration, we turn this oracle design into an adaptive certification procedure. Since the true environment is unknown, TRACK-AND-CERTIFY repeatedly estimates it, computes the corresponding characteristic allocation, tracks that allocation online, and stops once the current serving decision has sufficient likelihood evidence against every decision-changing alternative.

5.1 TRACK-AND-CERTIFY

Write p_ϑ^z for the observation mass function and define the log-likelihood

$$\ell_t(\vartheta) = \sum_{s=1}^{t} \log p_\vartheta^{Z_s}(O_s). \tag{13}$$

Algorithm 1 combines forced exploration, allocation tracking, and generalized likelihood-ratio (GLR) stopping. When the characteristic program has multiple optima, $n^\circ(\vartheta)$ denotes its unique minimum-ℓ_2-norm optimizer; other finite argmax ties use a fixed deterministic order.

Forced exploration guarantees persistent coverage of the evaluation channels, while the remaining rounds track the characteristic allocation predicted by the current environment estimate. The GLR stopping rule certifies the current serving configuration once every decision-changing alternative has accumulated sufficient evidence against it.

5.2 CORRECTNESS AND ASYMPTOTIC INSTANCE OPTIMALITY

We now show that the sampling and stopping rules above are both valid and first-order cost-optimal. Assume finite Θ and $\mathcal{Z}$, finite observation alphabets with uniformly positive common support, known fixed evaluation costs $h_z > 0$, fresh conditionally stationary observations, global identifiability, and a unique optimal serving configuration under the true environment.

Theorem 2 (Correctness and instance optimality)**.** *Under the conditions above,* TRACK-AND-CERTIFY *is δ-correct for every $\delta \in (0,1)$ and satisfies*

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}_\theta[C_{\tau_\delta}]}{\log(1/\delta)} = C^\star(\theta). \tag{14}$$

Together with Theorem 1, Theorem 2 establishes $C^\star(\theta)$ as the exact first-order certification cost: no δ-correct procedure can asymptotically improve on this constant, and TRACK-AND-CERTIFY attains it. The proof is given in Appendix C.1–Appendix C.3.

5.3 HISTORY-GUIDED EVALUATION-FAMILY DESIGN

An independent auxiliary dataset D_{hist} can select a restricted evaluation family $\widehat{\mathcal{Z}} = g(D_{\mathrm{hist}})$, after which fresh certification uses $\widehat{\mathcal{Z}}$. This allows prior information about the useful observation structure to reduce the set of actions explored during certification.

Corollary 1 (History-guided validity)**.** *Suppose D_{hist} is independent of the fresh certification stream conditional on θ, and every evaluation family in the range of g supports a valid δ-correct certifier. Then the resulting history-guided procedure remains δ-correct.*

The proof is given in Appendix C.4. Section 6 tests whether a characteristic statistic computed from independent DEV data can predict which evaluation family is useful on held-out TEST replay.

6 EXPERIMENTS

We organize the experiments around three questions. First, how much does adaptive allocation reduce certification cost relative to uniform allocation, and how broadly do these gains hold across model pools? Second, when is richer multi-candidate feedback worth purchasing, and does the characteristic cost predict that regime? Finally, do these conclusions persist when we search the full 64-state simplex for harder alternatives, reconstruct the evaluation geometry from limited historical calibration data, and select an evaluation family on a held-out task?

6.1 EXPERIMENTAL SETUP

We study certification using same-query binary correctness outcomes and serving costs from LLMRouterBench/Kairos (Li et al., 2026). The main benchmark contains 11 frozen model pools, each with one incumbent and five candidates. We label these pools M1–M11; each label refers to a

Algorithm 1 TRACK-AND-CERTIFY

Require: Θ, $\mathcal{Z}$, confidence level δ

$N_z(0) \leftarrow 0$ for all $z \in \mathcal{Z}$; $\beta_\delta \leftarrow \log((|\Theta| - 1)/\delta)$
for $t = 1, 2, \ldots$ **do**
  $\widehat{\theta}_{t-1} \in \arg\max_{\vartheta \in \Theta} \ell_{t-1}(\vartheta)$
  if $\mathrm{Alt}(\widehat{\theta}_{t-1}) = \varnothing$ **then**
    return $\pi^\circ(\widehat{\theta}_{t-1})$
  else if $\exists z : \ N_z(t-1) < \sqrt{t}/|\mathcal{Z}|$ **then**
    Choose any such action Z_t
  else
    Compute $n^\circ(\widehat{\theta}_{t-1})$ and set $\alpha^\circ_z = n^\circ_z / \sum_k n^\circ_k$
    $Z_t \in \arg\max_z \{\alpha^\circ_z t - N_z(t-1)\}$
  end if
  Observe O_t and update counts and likelihoods
  $\widehat{\theta}_t \in \arg\max_{\vartheta \in \Theta} \ell_t(\vartheta)$; $\widehat{\pi}_t \leftarrow \pi^\circ(\widehat{\theta}_t)$
  $G_t \leftarrow \inf_{\lambda : \widehat{\pi}_t \notin \Pi^\star(\lambda)} [\ell_t(\widehat{\theta}_t) - \ell_t(\lambda)]$
  if $G_t \geq \beta_\delta$ **then**
    return $\widehat{\pi}_t$
  end if
end for

complete incumbent–candidate set, not an individual model. Each pool yields 11 serving configurations and 16 evaluation actions. For a candidate subset S, an action observes $(Y_B, (Y_j)_{j \in S})$; the incumbent outcome is shared and is not charged again, so evaluation cost counts only additional candidate evaluations. Hence a singleton j observes (Y_B, Y_j), a pair j, k observes (Y_B, Y_j, Y_k), and full feedback observes $(Y_B, Y_1, \ldots, Y_K)$. For each pool, we construct a finite decision-relevant class Θ from the smoothed empirical 64-state joint distribution P_0 and the nearest decision-changing alternative for each feasible route. Characteristic costs in the main replay are computed as $C^\star_\Theta$. We compare four policies:

- UNIFORM-S: cycles evenly through the five singleton actions.
- TRACK-S: adaptively tracks the characteristic allocation over those five actions.
- UNIFORM-ALL: cycles evenly through all 16 actions.
- TRACK-ALL: adaptively tracks the characteristic allocation over all 16 actions.

For the singleton-only and all-action families, let $C^\star_{\mathrm{S}}$ and $C^\star_{\mathrm{A}}$ denote their characteristic costs and define $R = C^\star_{\mathrm{S}}/C^\star_{\mathrm{A}}$. We call pools with $R \geq 1.25$ higher-R and those with $R \leq 1.05$ near-singleton. Uniformity is over action counts rather than evaluation cost. Within each action family, methods share the likelihood model, stopping rule, confidence target, and cost accounting. Appendix D.1 gives the complete construction, seed counts, horizons, and reliability accounting.

6.2 ADAPTIVE ALLOCATION SUBSTANTIALLY REDUCES CERTIFICATION COST

We first ask how much certification cost is lost to uniform evaluation. Across the 11 frozen pools, the median pool-level mean cost ratio is 8.533 for UNIFORM-S/TRACK-S and 7.370 for UNIFORM-ALL/TRACK-ALL within the same action family (Table 1). In Figure 2, the circular UNIFORM-S/TRACK-S ratios exceed one for every pool, often by a wide margin: adaptive allocation broadly shifts effort toward actions informative for the serving decision. The triangular TRACK-S/TRACK-ALL ratios instead straddle one, so richer feedback reduces finite-sample cost for some pools but increases it for others. Light-blue shading marks the four higher-R pools ($R \geq 1.25$); the seven unshaded pools are near-singleton ($R \leq 1.05$), and pools are ordered by R. Full error, censoring, and pool-level accounting for the 9,900 baseline runs appear in Appendix D.3.

Table 1: **Certification-cost comparisons across model pools.** Median ratios of pool-level mean costs; values above one favor the denominator.

Comparison	Pools	Median cost ratio
UNIFORM-S/TRACK-S	11	$\mathbf{8.533}\times$
UNIFORM-ALL/TRACK-ALL	11	$\mathbf{7.370}\times$
TRACK-S/TRACK-ALL, higher-R group	4	$1.826\ [1.376, 2.128]$
TRACK-S/TRACK-ALL, near-singleton group	7	$0.394\ [0.283, 0.878]$

6.3 WHEN DOES RICHER FEEDBACK PAY OFF?

Does the characteristic-cost ratio R distinguish pools that benefit from richer multi-candidate feedback at finite confidence? For each pool, let $\overline{C}_m$ denote the mean certification cost of method m across replay seeds and define the realized mean-cost ratio $G = \overline{C}_{\text{Track-S}}/\overline{C}_{\text{Track-All}}$. Across the 11 frozen pools, R tracks the ordering of realized gains: Spearman $\rho = 0.8823$ with 95% bootstrap interval $[0.551, 0.965]$ (Figure 3(a)). The four higher-R pools have median TRACK-S/TRACK-ALL ratio 1.826, compared with 0.394 for the seven near-singleton pools (Table 1). Thus R is informative about *which* pools benefit from richer feedback, while finite-confidence gains vary in magnitude.

Dependence and cost sensitivity. To probe one source of this heterogeneity, we use two fixed controlled environments that preserve singleton identifiability while varying the candidate-dependence contrast $d \in [0, 1]$ and the normalized cost $h_{12} \in \{1, 2, 4\}$ of evaluating candidates 1 and 2 jointly on a query. At $d = 0$, candidate outcomes are conditionally independent given the incumbent outcome; $d = 1$ is the strongest contrast in the sweep. The three costs are controlled sensitivity settings, not measured benchmark prices. For each (d, h_{12}), Figure 3(b) plots $R_{\text{ctrl}} = \sqrt{R_0 R_1}$ and $G_{\text{ctrl}} = \sqrt{G_0 G_1}$, where R_e and G_e are the characteristic- and mean-cost ratios for each environment $e \in \{0, 1\}$. Greater dependence can make joint feedback more informative per unit cost about decision-changing alternatives. At $d = 1$, $h_{12} = 1, 2, 4$ yield $R_{\text{ctrl}} = 30.815, 15.407, 7.704$, respectively, and corresponding G_{ctrl} values of 19.989 (95% CI $[18.444, 21.616]$), 11.688 ($[10.818, 12.608]$), and 6.126 ($[5.696, 6.590]$). Conversely, when dependence is weak and paired feedback is expensive, the richer action family can incur finite-sample overhead. Richer feedback can therefore substantially reduce certification cost even when it is unnecessary for identifiability, but its value is instance- and cost-dependent.

6.4 ROBUSTNESS, CALIBRATION, AND TRANSFER

We first test whether the selective value of richer feedback depends on the route-wise finite alternative class. Solving continuous hardest-alternative problems over the full 64-state simplex, all 22 pool–family problems converge (maximum cutting-plane gap 9.84×10^{-6}), and singleton evaluation remains identifiable on all 11 pools. The harder alternatives increase median $C^\star$ by $1.159\times$ for singleton and $1.181\times$ for all-action evaluation, yet the finite and continuous characteristic ratios have Spearman $\rho = 0.8724$ and the frozen $R \geq 1.25$ family choice agrees on all 11 pools (Appendix D.5). Adding the discovered adversarial alternatives to each finite class yields 2,200 replay runs without errors or censoring. The augmented characteristic ratio remains associated with realized TRACK-S/TRACK-ALL cost (Spearman $\rho = 0.8818$), and TRACK-ALL is cheaper on all four richer-feedback pools M1, M5, M7, and M10. Smoothing, route-margin, and probability-floor perturbations also preserve the family choice (Appendix D.6). Thus the finite construction understates absolute difficulty but preserves the selective feedback conclusion.

Calibration and held-out transfer. To assess how much historical data is needed to recover the evaluation-family geometry, we rebuild P_0 and the complete finite class from independent historical full-feedback subsamples. Relative to the full-data reference, the frozen family rule agrees on 94.7% of pool–replicates with 128 prompts, 98.5% with 256, and 100% with 512; the median rank correlation of characteristic ratios reaches 0.989 by 256 prompts. Exact serving decisions stabilize more slowly, reaching 96.4% agreement at 512 prompts (Appendix D.7). Finally, after removing `mmlupro` from DEV, the frozen family rule matches the lower mean-cost family on all 11 held-out MMLU-Pro pools, with overlapping uncertainty for M9. Full transfer and finite-confidence diagnostics appear in Appendices D.8 and D.9.

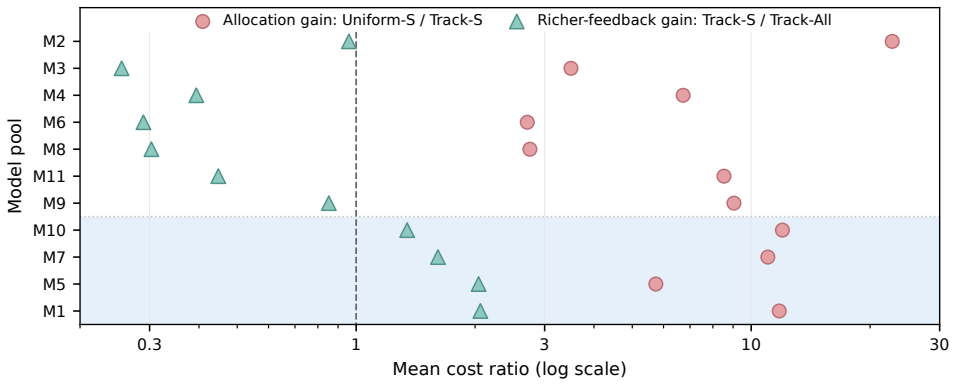
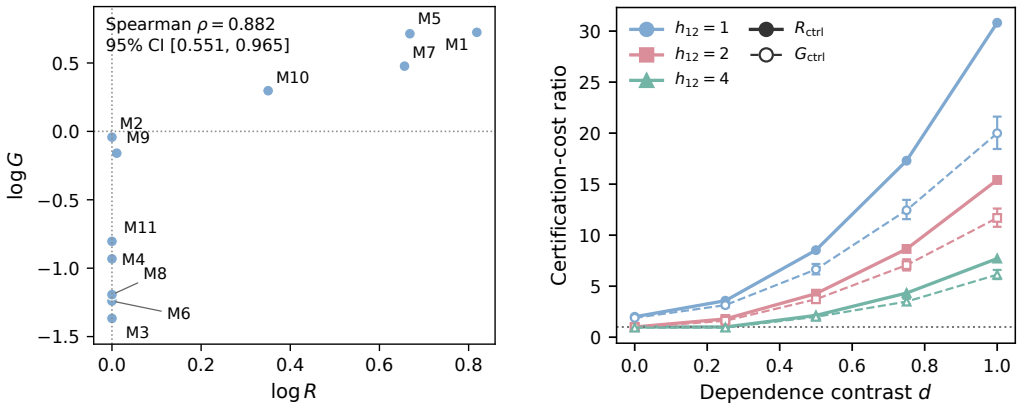


Figure 2: **Pool-level certification gains.** Pools are ordered by increasing characteristic-cost ratio R from top to bottom; light blue marks $R \geq 1.25$; ratios above one favor the denominator.



(a) Frozen model pools **(b)** Controlled information-efficiency study

Figure 3: **When richer feedback pays off.** (a) Pool-level prediction; (b) a controlled information-efficiency study with singleton identifiability.

7 Conclusion and Limitations

We studied how to allocate costly model evaluations before an LLM serving update. The characteristic certification cost $C^\star(\theta)$ captures the minimum first-order evaluation cost required to rule out every environment that would change the optimal serving configuration. We proved that this cost is unavoidable for any δ-correct certifier and that Track-and-Certify asymptotically attains it. In the LLM specialization, the same framework exposes a structural distinction between marginal and joint information: candidate replacement depends on marginal model performance, whereas verifier-triggered cascades can additionally depend on same-query rescue behavior. The replay experiments reflect this structure—decision-aware allocation provides broad savings, while paired same-query evaluation is useful only on model pools where joint information reduces the characteristic cost. The current achievability result is stated for finite, stationary hypothesis classes with known evaluation costs, and the empirical characteristic costs are defined with respect to the frozen finite classes used in replay. The experiments further use binary correctness outcomes and verifier-triggered one-stage cascades. Extending the framework to noisy or open-ended quality evaluation, learned deferral signals, and richer serving policies is a natural direction for future work.

AI USE STATEMENT

Generative AI tools were used to polish and reorganize manuscript text, and assist in debugging code. The authors reviewed the AI-assisted work and take responsibility for the final content.

REFERENCES

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of PMLR, pages 12987–13010, 2025.

Aditya Deshmukh, Venugopal V. Veeravalli, and Srikrishna Bhashyam. Sequential controlled sensing for composite multihypothesis testing. *Sequential Analysis*, 40(2):259–289, 2021.

Tanner Fiez, Lalit Jain, Kevin G. Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of PMLR, pages 998–1027, 2016.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *International Conference on Learning Representations*, 2024.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of PMLR, pages 5753–5763, 2021.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, Zhen Wang, and Shuyue Hu. LLMRouterBench: A massive benchmark and unified framework for LLM routing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37733–37754, 2026.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Asterios Tsiourvas, Wei Sun, and Georgia Perakis. Causal LLM routing: End-to-end regret minimization from observational data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via Frank-Wolfe. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *International Conference on Learning Representations*, 2024.

A Supplementary Material for Section 3

A.1 Formal conventions and LLM specialization

We use the filtration $\mathcal{F}_t = \sigma(Z_1, O_1, \ldots, Z_t, O_t)$. Each evaluation choice Z_t is $\mathcal{F}_{t-1}$-measurable, and observations are conditionally fresh:

$$P_\theta(O_t \in A \mid \mathcal{F}_{t-1}, Z_t = z) = P_\theta^z(A).$$

This predictability convention is used throughout the likelihood-ratio and concentration arguments below. When an estimated environment has multiple optimal serving configurations, the fixed selector π° from Section 3 is used. The lower-bound and asymptotic optimality results are stated at true environments with a unique optimal serving configuration, denoted $\pi^\star(\theta)$.

For the LLM specialization, the verifier-triggered cascade in Eq. (4) uses observable, verified incumbent failure as its escalation signal. This matches the automatically verifiable controlled replay analyzed here. The general certification theory is formulated in terms of the induced utilities $J_\pi(\theta)$ and observation channels P_θ^z.

B Supplementary Material for Section 4

B.1 Proof of the fundamental lower bound

We first record the stopped-transcript change-of-measure identity used by Theorem 1. Let $H_t = (Z_1, O_1, \ldots, Z_t, O_t)$ and $\tau_m = \tau \wedge m$. For any $\theta, \lambda \in \Theta$, the chain rule for relative entropy, predictability of Z_t, and conditional freshness give

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{H_{\tau_m}} \middle\| \mathbb{P}_\lambda^{H_{\tau_m}}\right) = \mathbb{E}_\theta\left[\sum_{s=1}^{\tau_m} \log \frac{p_\theta^{Z_s}(O_s)}{p_\lambda^{Z_s}(O_s)}\right] \tag{15}$$

$$= \sum_{z \in \mathcal{Z}} \mathbb{E}_\theta[N_z(\tau_m)]\, d_z(\theta, \lambda). \tag{16}$$

Passing to $m \to \infty$ gives the corresponding identity whenever the expected cost is finite; if it is infinite, the lower bound is immediate. Equivalently, the inequality needed below follows by Fatou's lemma.

Proof of Theorem 1. Fix $\lambda \in \mathrm{Alt}(\theta)$ and set $E = \{\widehat{\pi} = \pi^\star(\theta)\}$. By δ-correctness and uniqueness at θ, $\mathbb{P}_\theta(E) \geq 1 - \delta$. Since $\lambda \in \mathrm{Alt}(\theta)$, $\pi^\star(\theta)$ is not optimal under λ, hence $\mathbb{P}_\lambda(E) \leq \delta$. Binary data processing gives

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{H_\tau} \middle\| \mathbb{P}_\lambda^{H_\tau}\right) \geq \mathrm{kl}(\mathbb{P}_\theta(E), \mathbb{P}_\lambda(E)) \geq \mathrm{kl}(1-\delta, \delta), \tag{17}$$

where the last inequality is Bernoulli-KL monotonicity on $p \geq 1 - \delta$, $q \leq \delta$ for $\delta < 1/2$. Combining Eqs. (16)–(17), for every $\lambda \in \mathrm{Alt}(\theta)$,

$$\sum_z \mathbb{E}_\theta[N_z(\tau)] d_z(\theta, \lambda) \geq \mathrm{kl}(1-\delta, \delta).$$

Thus $n_z = \mathbb{E}_\theta[N_z(\tau)] / \mathrm{kl}(1-\delta, \delta)$ is feasible for Eq. (8), so

$$C^\star(\theta) \leq \sum_z h_z n_z = \frac{\mathbb{E}_\theta[C_\tau]}{\mathrm{kl}(1-\delta, \delta)}.$$

This proves Eq. (9); using $\mathrm{kl}(1-\delta, \delta) = \log(1/\delta) + O(1)$ gives the asymptotic statement. □

B.2 IDENTIFIABILITY, DUALITY, AND COST–BLACKWELL PRUNING

Decision identifiability. Suppose some $\lambda \in \mathrm{Alt}(\theta)$ satisfies $P_\theta^z = P_\lambda^z$ for every z. By induction over rounds, every predictable adaptive policy then induces the same transcript and output laws under θ and λ. With $E = \{\widehat{\pi} = \pi^\star(\theta)\}$, simultaneous δ-correctness would require $\mathbb{P}_\theta(E) \geq 1 - \delta$ and $\mathbb{P}_\lambda(E) \leq \delta$, impossible for $\delta < 1/2$. Hence decision identifiability is necessary. The stronger global identifiability condition used below is sufficient for our current MLE-based achievability proof; it is stronger than the decision-identifiability condition required merely to distinguish environments that induce different serving decisions.

LP duality. For a finite alternative class, write Eq. (8) as

$$\min_{n \geq 0} h^\top n \quad \text{s.t.} \quad Dn \geq \mathbf{1}, \qquad D_{\lambda z} = d_z(\theta, \lambda).$$

Finite-dimensional strong duality gives

$$C^\star(\theta) = \max_{w \geq 0} \mathbf{1}^\top w \quad \text{s.t.} \quad D^\top w \leq h,$$

which is Eq. (10).

Proof of Cost–Blackwell pruning. If z_2 Blackwell-dominates z_1, there is a Markov kernel K with $P_\eta^{z_1} = K P_\eta^{z_2}$ for every η. Data processing gives

$$d_{z_1}(\theta, \lambda) \leq d_{z_2}(\theta, \lambda) \qquad \forall \lambda.$$

Given any feasible design with $n_{z_1} > 0$, move this mass from z_1 to z_2. Every information constraint weakly increases and the objective weakly decreases because $h_{z_2} \leq h_{z_1}$. Applying this transformation to an optimum yields an optimum with $n^\star_{z_1} = 0$. □

B.3 VERIFIER-TRIGGERED CASCADE SPECIALIZATION

Let $\rho = \mathbb{P}(Y_B = 0)$ and $q = \mathbb{P}(Y_j = 1 \mid Y_B = 0)$. Consider the restricted family in which the serving choice changes at $q = c$, a paired evaluation costs h, and alternatives vary only q' while leaving the remaining observation law fixed. Then

$$D_{\mathrm{KL}}(P_q \| P_{q'}) = \rho \, \mathrm{kl}(q, q').$$

Among configuration-changing alternatives, Bernoulli KL is minimized at the boundary $q' = c$. With one evaluation action, Eq. (8) therefore reduces to $n\rho \, \mathrm{kl}(q, c) \geq 1$, so

$$C^\star(q) = \frac{h}{\rho \, \mathrm{kl}(q, c)}.$$

Using $\mathrm{kl}(q, c) = \frac{(q-c)^2}{2c(1-c)} + o((q-c)^2)$ as $q \to c$ gives the scaling in Eq. (12). □

C SUPPLEMENTARY MATERIAL FOR SECTION 5

C.1 FIXED-CONFIDENCE CORRECTNESS OF TRACK-AND-CERTIFY

For a false environment $\eta \neq \theta$, define

$$M_t^{\eta,\theta} = \exp\{\ell_t(\eta) - \ell_t(\theta)\}.$$

Predictability and conditional freshness imply

$$\mathbb{E}_\theta[M_t^{\eta,\theta} \mid \mathcal{F}_{t-1}] = M_{t-1}^{\eta,\theta},$$

so this is a nonnegative mean-one martingale. If the algorithm stops incorrectly at time t, then $\widehat{\pi}_t \notin \Pi^\star(\theta)$ and the true environment belongs to the infimum defining G_t. Hence

$$\ell_t(\widehat{\theta}_t) - \ell_t(\theta) \geq \beta_\delta.$$

An incorrect stop has $\widehat{\theta}_t \neq \theta$, so some false η satisfies $\ell_t(\eta) - \ell_t(\theta) \geq \beta_\delta$. Ville's inequality and a union bound yield

$$\begin{aligned}\mathbb{P}_\theta(\text{error}) &\leq \sum_{\eta \neq \theta} \mathbb{P}_\theta\left(\sup_t M_t^{\eta,\theta} \geq e^{\beta_\delta}\right) \\ &\leq (|\Theta| - 1)e^{-\beta_\delta} = \delta.\end{aligned}$$

This proves the fixed-confidence statement in Theorem 2.

C.2 MLE STABILIZATION AND ALLOCATION TRACKING

Let $K = |\mathcal{Z}|$. The forced-exploration rule gives deterministic $c_K, C_K > 0$ such that

$$N_z(t) \geq c_K\sqrt{t} - C_K, \qquad z \in \mathcal{Z}, \tag{18}$$

and the number $F(t)$ of forced rounds through t is $O(\sqrt{t})$. Each action is therefore sampled infinitely often while the forced fraction vanishes.

MLE stabilization. For each action z, expose an independent fresh stack $O_{z,1}, O_{z,2}, \ldots \overset{i.i.d.}{\sim} P_\theta^z$ and reveal its next element whenever z is selected. For $\lambda \neq \theta$,

$$\ell_t(\theta) - \ell_t(\lambda) = \sum_z \sum_{i=1}^{N_z(t)} \log \frac{p_\theta^z(O_{z,i})}{p_\lambda^z(O_{z,i})}.$$

The strong law makes the average contribution of action z converge to $d_z(\theta, \lambda) \geq 0$. Global identifiability provides at least one action with strictly positive divergence, and Eq. (18) sends its count to infinity. Since Θ is finite, there is an almost surely finite time T_θ after which $\widehat{\theta}_t = \theta$ forever.

For the quantitative tail, define the Hellinger affinity

$$r_z(\theta, \lambda) = \sum_o \sqrt{p_\theta^z(o)p_\lambda^z(o)}.$$

For every false λ, choose a distinguishing action with $r_z(\theta, \lambda) < 1$. Predictability and conditional freshness yield the adaptive Hellinger bound

$$\mathbb{E}_\theta \exp\left(\frac{\ell_t(\lambda) - \ell_t(\theta)}{2}\right) \leq r_z(\theta, \lambda)^{N_z(t)}.$$

Together with Eq. (18), Markov's inequality, and a finite union bound over false environments and future times, there are constants $C, c > 0$ such that

$$\mathbb{P}_\theta(T_\theta > t) \leq Ce^{-c\sqrt{t}}. \tag{19}$$

Quantitative tracking. Fix $\alpha \in \Delta(\mathcal{Z})$ and set

$$q_z(t) = \alpha_z t - N_z(t), \qquad D_\alpha(t) = \sum_z [q_z(t)]_+.$$

Suppose that after time T, every non-forced round chooses an action of maximum deficit, and let $F_T(t)$ be the number of forced rounds after T through t. Induction over rounds gives

$$D_\alpha(t) \leq \max\{D_\alpha(T), K + 1\} + F_T(t). \tag{20}$$

Since $\sum_z q_z(t) = 0$, with $B_T(t) = \max\{D_\alpha(T), K + 1\} + F_T(t)$,

$$\alpha_z t - B_T(t) \leq N_z(t) \leq \alpha_z t + (K - 1)B_T(t). \tag{21}$$

Also $D_\alpha(T) \leq T$ and $F_T(t) = O(\sqrt{t})$. Let $n^\circ(\theta)$ be the canonical minimum-ℓ_2 characteristic optimizer defined in Section 5, and normalize it to α°. After T_θ, the target equals this fixed allocation, so

$$\max_z |N_z(t) - \alpha_z^\circ t| \leq C\big(T_\theta + \sqrt{t} + 1\big). \tag{22}$$

Thus $N_z(t)/t \to \alpha_z^\circ$ almost surely. If $K = 1$, exact tracking is immediate.

C.3 Proof of asymptotic instance optimality

Let $n^\circ(\theta)$ be the canonical characteristic optimizer and normalize

$$\alpha_z^\circ = \frac{n_z^\circ}{\sum_k n_k^\circ}.$$

Assume $\mathrm{Alt}(\theta) \neq \varnothing$; the empty case is treated below. Define

$$I^\star(\theta) = \min_{\lambda \in \mathrm{Alt}(\theta)} \sum_z \alpha_z^\circ d_z(\theta, \lambda), \qquad H^\star(\theta) = \sum_z h_z \alpha_z^\circ.$$

At an optimum of Eq. (8), at least one information constraint is tight; otherwise all n_z° could be scaled down. If $S = \sum_z n_z^\circ$, then $I^\star = 1/S$ and $H^\star = C^\star/S$, hence

$$\frac{H^\star}{I^\star} = C^\star(\theta). \tag{23}$$

Pathwise optimality. By Eq. (22), $N_z(t)/t \to \alpha_z^\star$. The strong law on the action-specific fresh stacks and finiteness of $\mathrm{Alt}(\theta)$ give

$$\frac{G_t}{t} \to I^\star(\theta) \qquad \text{a.s.} \tag{24}$$

after MLE stabilization. Since $\beta_\delta \to \infty$,

$$\frac{\tau_\delta}{\beta_\delta} \to \frac{1}{I^\star(\theta)} \qquad \text{a.s.} \tag{25}$$

and $C_t/t \to H^\star$ implies

$$\frac{C_{\tau_\delta}}{\beta_\delta} \to \frac{H^\star}{I^\star} = C^\star(\theta) \qquad \text{a.s.} \tag{26}$$

Because $\beta_\delta = \log(1/\delta) + O(1)$, this proves pathwise first-order optimality.

Expected-cost optimality. For each $\lambda \in \mathrm{Alt}(\theta)$, write

$$L_t^\lambda = \ell_t(\theta) - \ell_t(\lambda), \quad A_t(\lambda) = \sum_{s=1}^t d_{Z_s}(\theta, \lambda), \quad M_t^\lambda = L_t^\lambda - A_t(\lambda).$$

The process M_t^λ is a martingale. Uniform positive common support bounds its one-step centered increments by a deterministic B_M, so Azuma–Hoeffding gives

$$\mathbb{P}_\theta(M_t^\lambda \leq -x) \leq \exp\left(-\frac{x^2}{2B_M^2 t}\right). \tag{27}$$

Split

$$\mathbb{P}_\theta(\tau_\delta > t) \leq \mathbb{P}_\theta(T_\theta > t^{2/3}) + \mathbb{P}_\theta(\tau_\delta > t,\ T_\theta \leq t^{2/3}). \tag{28}$$

The first term is at most $C_1 e^{-c_1 t^{1/3}}$ by Eq. (19). On $\{T_\theta \leq t^{2/3}\}$, Eq. (22) yields uniformly in $\lambda \in \mathrm{Alt}(\theta)$

$$A_t(\lambda) \geq t I^\star - C_D t^{2/3}. \tag{29}$$

Fix $K_0 > 1/I^\star$ and let $\Delta = I^\star - 1/K_0 > 0$. If $t \geq K_0 \beta_\delta$, then $\beta_\delta \leq t/K_0$; for all sufficiently large t, Eq. (29) gives

$$A_t(\lambda) - \beta_\delta \geq \frac{\Delta}{2} t \qquad \forall \lambda \in \mathrm{Alt}(\theta).$$

If additionally $\tau_\delta > t$ and $T_\theta \leq t^{2/3}$, then the MLE has stabilized, $G_t < \beta_\delta$, and for some $\lambda \in \mathrm{Alt}(\theta)$,

$$M_t^\lambda \leq -\frac{\Delta}{2} t.$$

Eq. (27) and a finite union bound therefore give

$$\mathbb{P}_\theta(\tau_\delta > t,\ T_\theta \leq t^{2/3}) \leq C_2 e^{-c_2 t}.$$

Hence, for $t \geq t_0 \vee K_0\beta_\delta$,

$$\mathbb{P}_\theta(\tau_\delta > t) \leq C_1 e^{-c_1 t^{1/3}} + C_2 e^{-c_2 t}, \tag{30}$$

with constants independent of δ.

Set $X_\delta = \tau_\delta/\beta_\delta$ and fix a sufficiently small δ_0. For $0 < \delta \leq \delta_0$ and $u \geq K_0$, Eq. (30) is bounded by the integrable envelope

$$g(u) = C_1 e^{-c_1 (u\beta_{\delta_0})^{1/3}} + C_2 e^{-c_2 u \beta_{\delta_0}}.$$

For nonnegative X,

$$\mathbb{E}[X\mathbf{1}\{X > M\}] = M\mathbb{P}(X > M) + \int_M^\infty \mathbb{P}(X > u)\, du.$$

Thus $\{X_\delta : 0 < \delta \leq \delta_0\}$ is uniformly integrable. Since $0 \leq C_{\tau_\delta} \leq h_{\max}\tau_\delta$, the scaled costs $C_{\tau_\delta}/\beta_\delta$ are uniformly integrable as well. Combining this with Eq. (26),

$$\frac{\mathbb{E}_\theta[C_{\tau_\delta}]}{\beta_\delta} \to C^\star(\theta).$$

Finally $\beta_\delta = \log(1/\delta) + O(1)$, proving the expectation-level statement in Theorem 2. □

If $\mathrm{Alt}(\theta) = \varnothing$, the selected configuration is optimal throughout the finite environment class, $C^\star(\theta) = 0$, and the procedure may stop without purchasing information; the theorem is immediate.

C.4 History-guided evaluation design

Condition on the independent historical dataset D_{hist}, which fixes $\widehat{\mathcal{Z}} = g(D_{\mathrm{hist}})$. By assumption, the fresh certifier within every family in the range of g is δ-correct. Thus for every realization of D_{hist},

$$\mathbb{P}_\theta(\widehat{\pi} \notin \Pi^\star(\theta) \mid D_{\mathrm{hist}}) \leq \delta.$$

Taking expectations over D_{hist} gives

$$\mathbb{P}_\theta(\widehat{\pi} \notin \Pi^\star(\theta)) = \mathbb{E}[\mathbb{P}_\theta(\widehat{\pi} \notin \Pi^\star(\theta) \mid D_{\mathrm{hist}})] \leq \delta.$$

This proves the history-guided validity corollary. □

D Additional Experimental Details

D.1 Replay protocol and finite hypothesis classes

Frozen model pools and serving configurations. The main certification replay uses 11 frozen five-candidate model pools. Each pool contains one incumbent B and five candidates, giving 11 serving configurations: incumbent-only, five candidate-only configurations, and five verifier-triggered cascades. The separate selection-neutral serving-geometry diagnostic uses 11 frozen incumbents and a canonical 33-model pool. Among 352 possible non-self incumbent–candidate pairs, 172 satisfy the frozen DEV support and identity rules; all 180 exclusions fail the common-prompt support threshold. The serving-cost scales $s \in \{0.25, 0.50, 0.75\}$ are fixed in advance, with $s = 0.50$ used for the multi-candidate certification study.

Observation semantics and costs. For a candidate subset S, an evaluation action observes

$$O^S = (Y_B, (Y_j)_{j \in S}).$$

The incumbent outcome is shared and incurs no additional evaluation charge; costs count only newly evaluated candidates. Consequently, a singleton action $\{j\}$ already identifies the full incumbent–candidate joint law (Y_B, Y_j), a pair action $\{j, k\}$ observes (Y_B, Y_j, Y_k), and full feedback observes all incumbent and candidate outcomes. The 16-action family therefore consists of five singleton, ten paired, and one full-feedback action. This distinction is important: richer actions need not expand identifiability of the serving utilities studied here, but they can increase information per unit evaluation cost about decision-changing alternatives.

Finite hypothesis class. For each frozen model pool, the replay class is constructed from a smoothed empirical joint distribution P_0 over the $2^6 = 64$ binary incumbent/candidate outcome states. For every feasible route that changes the optimal serving decision, we add its nearest forward-KL I-projection subject to the corresponding decision-changing constraint. Positive-support smoothing and a probability floor keep all likelihood ratios and KL divergences finite. The route margin, smoothing level, and probability floor are frozen before the main replay. Thus the replay characteristic quantity is $C^\star_\Theta$ for this compact, decision-relevant finite class.

Certification protocol. The main replay uses a common likelihood model, GLR stopping rule, target confidence, and cost accounting within each compared action family. UNIFORM-S and TRACK-S are restricted to the five singleton actions; UNIFORM-ALL and TRACK-ALL use all 16 actions. The main text reports pool-level ratios of mean certification cost and then aggregates those ratios across the 11 pools. Runs that reach the frozen horizon without certification are recorded as censored.

D.2 ADDITIONAL SERVING-GEOMETRY DIAGNOSTICS

The serving-update problem is nontrivial across the benchmark. Table D.1 reports the optimal serving configuration over 172 eligible incumbent–candidate pairs at three pre-specified serving-cost scales, and Figure D.1 shows the corresponding decision geometry. None of incumbent-only, candidate-only, or cascade serving is confined to a negligible region at the central operating point.

Among nonoptimal serving targets, 68 admit decision-changing alternatives while 42 are structurally infeasible. A separate maximum-achievable-margin LP reproduces the same 68/42 split. Every infeasible target is globally weakly dominated by at least one competing serving configuration and strictly dominated somewhere.

D.3 MAIN REPLAY ACCOUNTING AND POOL-LEVEL DIAGNOSTICS

Table D.2 summarizes the scale and reliability accounting of all seed-level trajectories in the completed experimental suites. The main baseline study contains 9,900 seed-level runs, including all nine incorrect certifications and two censored runs. The controlled mechanism study contains 15,600 runs with no censoring; in its Part B sweep, the overall observed error rates are 1.13% for TRACK-S and 0.63% for TRACK-ALL, and the largest single-condition observed error rate is 2.5%. These frequencies summarize finite-sample reliability across the controlled sweep.

The pool-level main result has two distinct components. First, adaptive allocation is broadly useful: the median UNIFORM-S/TRACK-S and UNIFORM-ALL/TRACK-ALL mean-cost ratios are 8.533 and 7.370. Second, expanding the feedback family is selective. Under the frozen characteristic partition, the four higher-R pools have median TRACK-S/TRACK-ALL ratio 1.826 with 95% hierarchical-bootstrap interval $[1.376, 2.128]$, whereas the seven near-singleton pools have median 0.394 with interval $[0.283, 0.878]$. The latter negative finite-sample effect shows that a richer action family can incur additional acquisition and exploration cost when its asymptotic information advantage is negligible.

Table D.2: **Experimental accounting.** Seed-level runs and reliability by suite; dashes denote non-certification studies.

Study	Runs	Censored	Reliability note
Main baseline replay	9,900	2	9 incorrect certifications
Controlled mechanism	15,600	0	Part B: 1.13% (TRACK-S), 0.63% (TRACK-ALL)
Finite-Θ sensitivity	15,400	0	no censoring
Calibration geometry	6,600	–	historical calibration only
Held-out MMLU-Pro transfer	2,200	0	no censoring

D.4 CONTROLLED RICHER-FEEDBACK MECHANISM STUDY

We use a controlled construction to separate the role of same-query dependence in the serving decision from the efficiency value of multi-candidate feedback. In Part A we fix $p_B = 0.70$ and $p_j = 0.65$ and vary only the feasible dependence parameter λ. Marginal accuracies therefore remain constant, but

Table D.1: Selection-neutral serving choices for 172 eligible pairs at three cost scales.

Cost scale s	Incumbent-only	Candidate-only	Cascade
0.25	51 (29.7%)	51 (29.7%)	70 (40.7%)
0.50	58 (33.7%)	57 (33.1%)	57 (33.1%)
0.75	64 (37.2%)	64 (37.2%)	44 (25.6%)

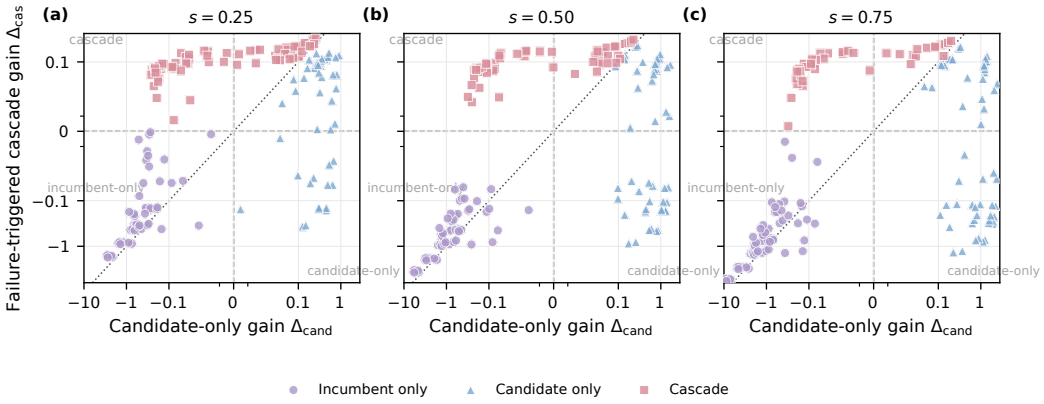


Figure D.1: **Serving-configuration geometry.** Axes show candidate-only and cascade gains over incumbent-only serving.

the cascade utility changes through the incumbent–candidate joint law. The optimal serving decision switches from Cascade to Keep at $\lambda^\star = 0.75$. Because singleton evaluation observes (Y_B, Y_j), singleton feedback remains sufficient to identify this joint law. Part A isolates how same-query dependence changes serving utility while singleton identifiability is preserved.

Part B introduces multiple candidates while preserving singleton identifiability. Its dependence contrast $d \in [0, 1]$ moves from conditional independence of candidate outcomes given the incumbent at $d = 0$ to the strongest contrast in the sweep at $d = 1$. Each singleton action costs 1; evaluating candidates 1 and 2 jointly costs $h_{12} \in \{1, 2, 4\}$, the other pairs cost 2, and full feedback costs $h_{12} + 1$. The shared incumbent outcome adds no evaluation charge. These normalized values are controlled cost-sensitivity settings, not benchmark-measured prices: $h_{12} = 2$ equals the sum of two singleton costs, while 1 and 4 represent discounted and surcharged pair actions. For each pair cost h_{12}, the plotted theoretical and empirical ratios are geometric means over two fixed environments, as defined in Section 6.3. At $d = 1$, pair costs $h_{12} = 1, 2, 4$ give $R_{\mathrm{ctrl}} = 30.815, 15.407, 7.704$, respectively. The corresponding G_{ctrl} values are 19.989 (95% CI $[18.444, 21.616]$), 11.688 ($[10.818, 12.608]$), and 6.126 ($[5.696, 6.590]$). In weak-dependence, high-pair-cost regimes the richer action family can instead incur a small finite-sample overhead. Figure D.2 shows the Part A serving-utility switch; Figure 3(b) shows the Part B cost-ratio sweep, and Figure D.3 shows the associated Track-All evaluation-cost composition.

D.5 Continuous hardest-alternative diagnostic and adversarial augmentation

The main replay uses a finite class built from the empirical 64-state joint law and route-wise forward-KL projections. To examine the structural effect of that finite construction, we also optimize over the full probability simplex. Let $p \in \Delta_{64}$ denote the smoothed empirical law of $(Y_B, Y_1, \ldots, Y_5)$ for a frozen pool, and let A_z be the marginalization operator for evaluation action z. For an action-family cost allocation $w \in \Delta(\mathcal{F})$, the information rate against an arbitrary alternative $q \in \Delta_{64}$ is

$$I_{\mathcal{F}}(w, q) = \sum_{z \in \mathcal{F}} \frac{w_z}{h_z} D_{\mathrm{KL}}(A_z p \| A_z q).$$

For each competing serving configuration, the condition that it weakly beats the true optimum is linear in q. We therefore minimize $I_{\mathcal{F}}(w, q)$ over each such decision-changing region with a

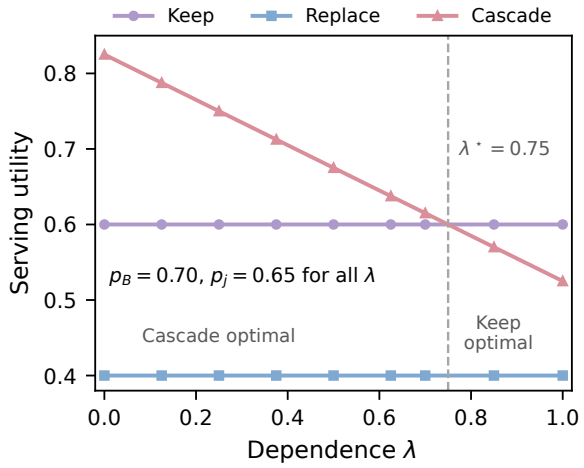


Figure D.2: **Dependence changes serving utility.** With fixed marginal accuracies, the optimal route switches from Cascade to Keep at $\lambda^\star = 0.75$.

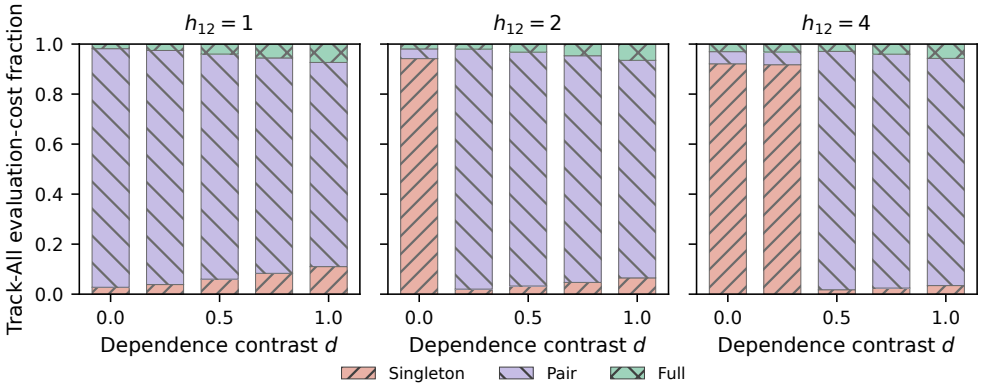


Figure D.3: **Evaluation-cost allocation.** Track-All spending across singleton, pair, and full-feedback actions in the controlled study.

convex separation oracle, and optimize the outer allocation by cutting planes. We solve this problem separately for the singleton and all-action families. This calculation supplies continuous characteristic quantities for comparison with the finite replay classes; the achievability theorem in Section 5 applies to the finite classes used by TRACK-AND-CERTIFY.

All 22 pool–family problems converge, with maximum final cutting-plane gap 9.84×10^{-6}, and singleton evaluation remains identifiable on all 11 pools. The continuous search finds harder alternatives than the original route-wise finite class: the median increase in characteristic cost is $1.159\times$ for singleton evaluation and $1.181\times$ for the all-action family. Despite this increase in absolute difficulty, the finite and continuous characteristic ratios $C^\star_{\mathrm{S}}/C^\star_{\mathrm{A}}$ have Spearman correlation 0.8724, and the frozen $R \geq 1.25$ family rule agrees on all 11 pools.

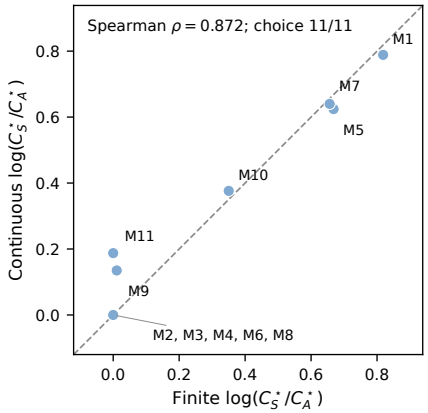


Figure D.4: **Finite versus continuous characteristic ratios.** One point per model pool; the dashed line is equality.

We next form an adversarially augmented finite class by adding every distinct hardest alternative returned by the continuous search to the original class and rerun TRACK-S and TRACK-ALL with the same replay protocol. The augmented study contains 2,200 seed-level runs with no errors or censoring. The augmented characteristic ratio has Spearman correlation 0.8818 with the realized TRACK-S/TRACK-ALL cost ratio, and TRACK-ALL is cheaper on all four richer-feedback pools M1, M5, M7, and M10.

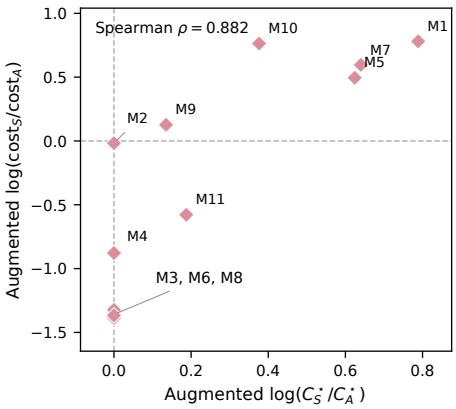


Figure D.5: **Predicted versus realized richer-feedback value** after adding continuous hardest alternatives to the finite classes.

D.6 ROBUSTNESS TO THE FINITE-Θ CONSTRUCTION

The finite class is generated from the smoothed empirical 64-state law and route-specific forward-KL I-projections described in Appendix D.1. We test six one-at-a-time perturbations around the original construction: lower and higher smoothing, lower and higher route margin, and lower and higher probability floor. Together with the original construction, this gives seven finite-Θ specifications.

Across every specification, all 11 singleton families remain identifiable and the optimal serving decision changes on 0/11 pools. Relative to the original construction, every alternative characteristic-ratio ranking has Spearman $\rho = 1.000$, and the frozen $R = 1.25$ action-family classification agrees on all 11 pools. The empirical TRACK-S/TRACK-ALL mean-cost ratio stays on the same side of one within each pool across all seven specifications (Table D.3). The table reports the largest absolute percentage change in this ratio relative to the original construction within a pool, across 22 pool–perturbation comparisons for each parameter. None of the 66 perturbed comparisons reverses

which method has the lower mean cost. Each of the 77 pool–specification cells uses 100 common seeds, and all 77 pointwise paired-seed bootstrap 95% intervals exclude one.

Table D.3: **Cheaper-method stability.** For TRACK-S/TRACK-ALL mean-cost ratios, we report the largest absolute percentage change from each pool's original across 22 cases.

Parameter varied	Max. relative change	Cheaper-method switches
Smoothing	3.5%	0/22
Route margin	12.4%	0/22
Probability floor	7.4%	0/22

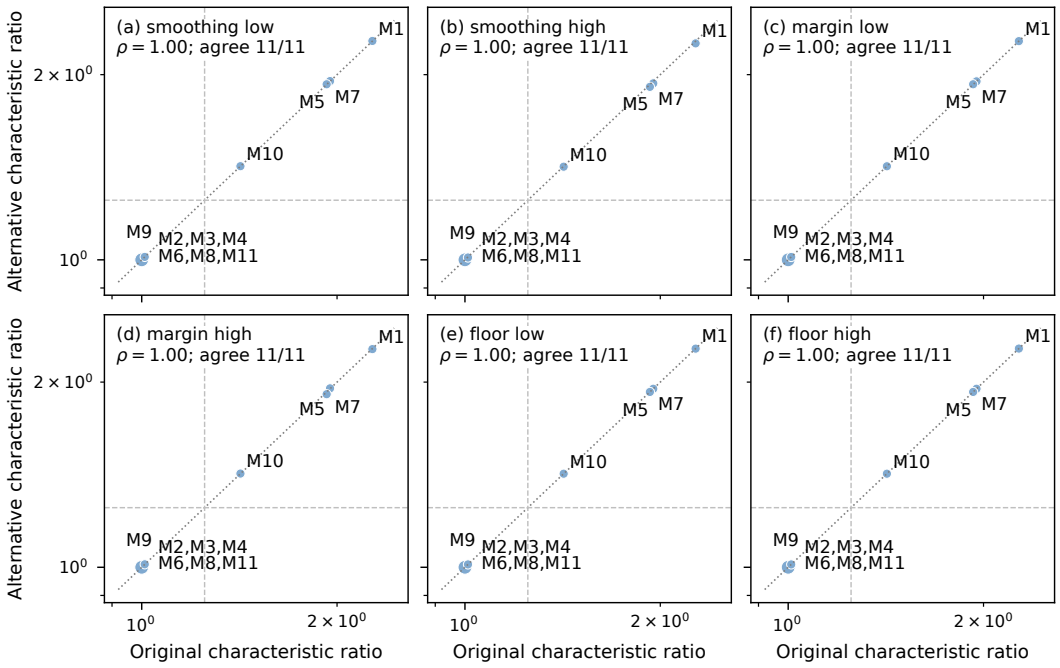


Figure D.6: **Characteristic-ratio stability** across six finite-Θ perturbations; pool rankings and family choices remain unchanged.

D.7 CALIBRATION-DATA REQUIREMENT

The characteristic design is constructed from a historical full-feedback calibration sample, so we measure how quickly this geometry stabilizes as the calibration set is reduced. For each frozen pool and calibration size $m \in \{32, 64, 128, 256, 512, \text{full}\}$, we draw independent calibration-only subsamples, rebuild the smoothed 64-state law P_0 and the complete finite class with the same frozen construction pipeline, and recompute $R = C^\star_{\mathrm{S}}/C^\star_{\mathrm{A}}$. The family threshold remains frozen at $R \geq 1.25$ throughout. Calibration draws exclusively from the historical full-feedback sample. The study contains 6,600 pool–size–replicate rows, and all reconstructed action families remain identifiable.

Table D.4 shows that the evaluation-family choice stabilizes substantially earlier than the exact serving decision. Agreement with the full-data family choice rises from 84.7% at 32 prompts to 94.7% at 128, 98.5% at 256, and 100% at 512. The median Spearman correlation of the 11 characteristic ratios with the full-data reference reaches 0.989 by 256 prompts. Exact serving decisions are more sample-sensitive, reaching 96.4% agreement at 512 prompts. Because six of the eleven full-reference ratios are exactly one, the pre-specified median absolute log-ratio error is identically zero; we therefore additionally report the mean absolute log error, which decreases from 0.132 at 32 prompts to 0.026 at 512.

Table D.4: **Calibration-geometry stability.** Family, ratio, and serving-choice agreement against the full-calibration reference.

Prompts/pool	Family agree.	Median Spearman	Mean $\lvert\Delta \log R\rvert$	Serving agree.
32	84.7%	0.861	0.132	69.0%
64	91.0%	0.938	0.096	73.8%
128	94.7%	0.973	0.058	80.1%
256	98.5%	0.989	0.047	87.8%
512	100%	0.989	0.026	96.4%
Full	100%	1.000	0	100%

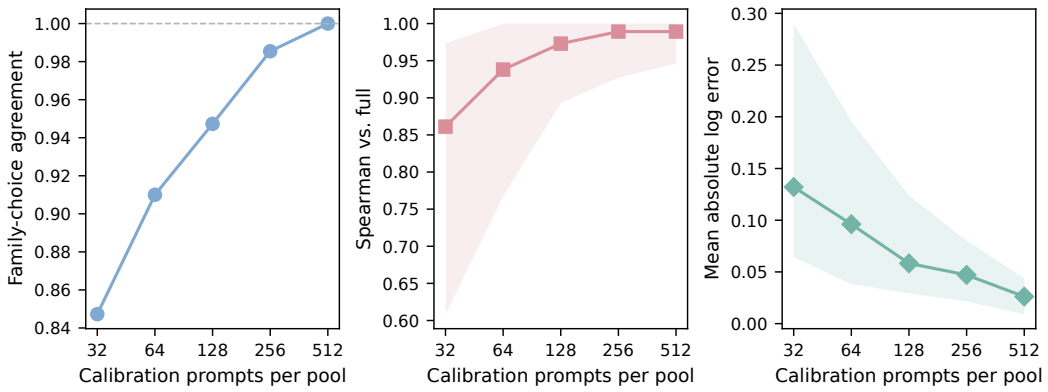


Figure D.7: **Calibration sample complexity.** Family agreement, characteristic-ratio rank correlation, and mean absolute log-ratio error.

D.8 Held-out-task evaluation-family selection

We evaluate whether the characteristic statistic can be used to choose an evaluation family before certification on a new task. The task identifier `mmlupro` is preselected by availability and reserved for TEST. The family rule is fit on the remaining DEV data and frozen before held-out MMLU-Pro replay.

Using the 100-seed mean point estimate in each pool, the frozen DEV rule chooses the lower-cost valid family on all 11 held-out-task pools; the two family costs have overlapping uncertainty for M9. The transfer study contains 2,200 seed-level runs with no censoring and measures held-out-task family selection within the same serving benchmark.

D.9 Finite-confidence scaling diagnostic

To probe whether the first-order constant is informative at practical confidence levels, we vary the target error probability over

$$\delta \in \{0.10, 0.05, 0.02, 0.01, 0.005, 0.002, 0.001\}$$

on four representative model pools chosen from frozen characteristic statistics before observing the scaling outcomes. We include an easy and a hard model pool with $R \leq 1.05$ (M8 and M2), both evaluated with TRACK-S, and an easy and a hard model pool with $R \geq 1.25$ (M10 and M1), both evaluated with TRACK-ALL. Each condition uses 100 common seeds, the same finite hypothesis class, GLR, evaluation costs, and forced-exploration rule as the main replay, and the original horizon of 500,000. None of the 2,800 runs is censored.

Because the algorithm stops at

$$\beta_\delta = \log \frac{|\Theta| - 1}{\delta},$$

Figure D.8 normalizes the observed mean cost by the exact finite-confidence characteristic scale $C^\star_\Theta(\theta)\beta_\delta$. A ratio near one indicates that the observed cost is close to the first-order characteristic

prediction after accounting for the exact GLR threshold. At $\delta = 10^{-3}$, M2, M10, and M1 reach ratios 1.053, 1.088, and 1.119, respectively. M8 behaves differently: its characteristic cost is exceptionally small, and its normalized ratio remains 9.643, although the ratio decreases steadily as confidence becomes stricter. M8 therefore exhibits substantial lower-order finite-confidence overhead relative to the characteristic term.

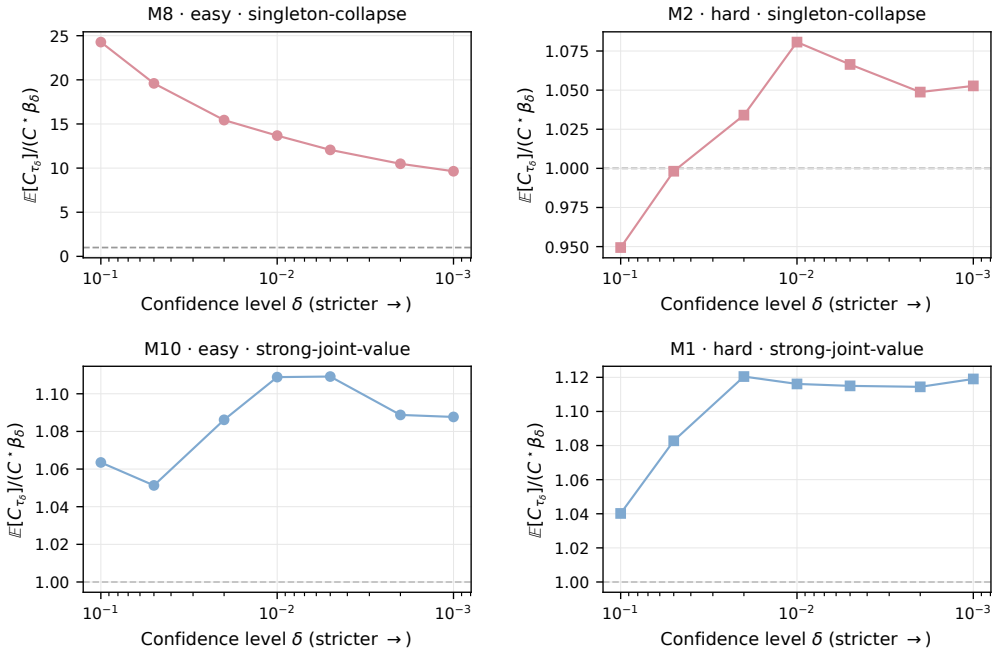


Figure D.8: **Finite-confidence scaling of TRACK-AND-CERTIFY.** Mean cost normalized by $C^\star_\Theta(\theta)\beta_\delta$ across four model pools.

Table D.5: Finite-confidence endpoint at $\delta = 10^{-3}$; 95% mean-cost intervals over 100 common-seed runs.

Model pool	Evaluation family	Mean cost [95% CI]	$\mathbb{E}[C]/(C^\star \log(1/\delta))$	$\mathbb{E}[C]/(C^\star \beta_\delta)$
M8	TRACK-S	0.142 [0.138, 0.148]	9.643	9.643
M2	TRACK-S	192.542 [176.246, 209.896]	1.298	1.053
M10	TRACK-ALL	4.504 [4.114, 4.894]	1.450	1.088
M1	TRACK-ALL	31.499 [29.042, 34.211]	1.492	1.119

Across the 400 runs at each successive confidence level, the total observed error counts are 18, 13, 3, 0, 0, 0, and 0. No model-pool-level empirical error rate exceeds its corresponding target δ. Mean cost increases as δ decreases for all four model pools, and no per-seed cost monotonicity violations are observed. For M8, $|\Theta| = 2$, so $\beta_\delta = \log(1/\delta)$ and the two normalizations in Table D.5 coincide.