DialBandit: Adaptive Sequential Search with Tunable Evaluation Fidelity

Jianping Huang[*]
Southeast University, China
jphuang@seu.edu.cn

Xiang Liu[*]
Southeast University, China
The Chinese University of Hong Kong, Hong Kong
xiangliu@seu.edu.cn

Feng Shan[†]
Southeast University, China
shanfeng@seu.edu.cn

Abstract

Model cascades reduce LLM inference costs by sequentially querying candidate models and stopping once a satisfactory response is generated. However, existing cascades largely optimize model escalation under fixed evaluation schemes, overlooking that optimal stopping fundamentally depends on the reliability of intermediate evaluations. We formalize this as an online sequential search problem with tunable-fidelity probes, in which the system jointly decides which model to query, at what fidelity to evaluate it, and when to stop. Under a fixed evaluation protocol, higher fidelity incurs greater evaluation cost but reduces stochastic observation noise through an unknown arm-specific noise function. For the oracle setting, we derive a generalized index policy that tightly couples fidelity selection with stage-wise stopping. For the online setting, we propose `DialBandit`, which learns fidelity-dependent noise functions from unbiased replicate-based labels and performs uncertainty-aware planning over probing order, evaluation fidelity, and stopping. We prove a finite-time sublinear regret bound and show that `DialBandit` consistently improves net utility over strong baselines in semi-empirical environments based on repeated judging and calibrated noisy LLM evaluations.

1 Introduction

Large language models (LLMs) deliver strong performance, but often at substantial inference cost. To mitigate this, many practical systems increasingly rely on *model cascades* [1, 2]: sequentially querying candidate models and stopping once a satisfactory response is generated. However, most existing cascades optimize mainly along the model-selection axis, using static thresholds [3] or offline-trained routing policies [4, 5], naively treating the evaluation process as fixed. In reality, the critical stop-or-continue decision depends not just on *which* model is queried, but on *how reliably* its response can be evaluated.

In modern LLM pipelines, one practical way to improve this reliability is to spend more evaluation-side compute, for example through repeated judging, larger voting pools, or evaluator ensembles [6–10]. We refer to this tunable evaluation budget as *fidelity*. When intermediate evaluation is noisy (*i.e.,* at low fidelity), a cascade may misjudge response quality, stopping prematurely on a weak answer or escalating unnecessarily and wasting compute. Figure 1(a) illustrates this phenomenon in a SummEval-based setup [11]: increasing fidelity reduces evaluation noise and ranking flips; further details are provided in Appendix C. This transforms the traditional cascade into a tightly coupled sequential decision problem: the system must jointly decide *which model to probe, at what fidelity to evaluate it, and when to stop the search*, as demonstrated in Figure 1(b).

[*]Equal contribution.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

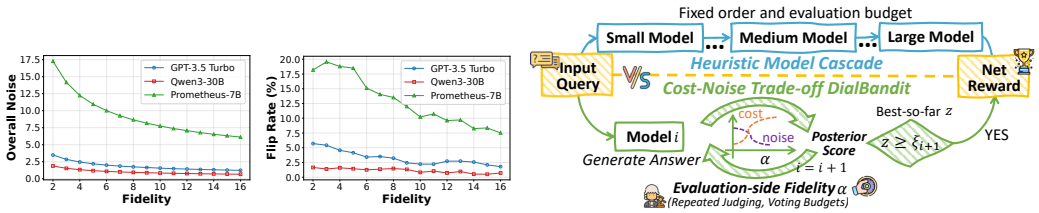


Figure 1: In (a), within a SummEval-based repeated-judging setup, increasing judge-time fidelity (e.g., more repeated evaluations) reduces effective noise and ranking flips. In (b), heuristic cascades typically fix probe order and evaluation budget, whereas our setting adapts model choice, fidelity, and stopping within each round.

These observations motivate us to formalize *online sequential search with tunable-fidelity probes*. Across repeated rounds, the learner faces this sequential decision problem and seeks to maximize *net utility* (*i.e.,* response quality minus generation and evaluation costs), continuing only when the value of further search justifies its cost. Crucially, in our model, fidelity modulates *stochastic evaluation noise* under a fixed evaluation protocol (*e.g.,* repeated sampling from a static judge or panel). The resulting task is therefore online control of information reliability around the evaluator's expected score; stable systematic evaluator bias [6, 12–14] is absorbed into that score rather than changed by fidelity (see Appendix J).

This formulation introduces three tightly coupled challenges. **First**, higher fidelity yields more reliable signals but incurs greater expense, creating a cost-noise tradeoff that is revealed only through policy-dependent interaction. **Second**, the learner must disentangle uncertainty about inherent model quality from uncertainty due to evaluation noise; otherwise, naive noise estimates can be systematically biased, since they are inferred from imperfect mean estimates on adaptively sampled observations. **Third**, these estimation errors propagate directly into sequential planning, distorting the probing order, fidelity allocation, and stopping decisions within each round.

Existing frameworks capture only parts of this setting. Pandora's box provides the closest sequential-search structure, but with a fixed observation model; once fidelity changes probe reliability, both probing values and stopping rules must change with it [15–17]. Multi-fidelity bandits usually treat fidelity as changing approximation quality, whereas here it changes the noise of a fixed target and thus the evidence for stopping [18–20]. LLM routing adapts model choice, but usually keeps evaluation fixed and therefore cannot jointly optimize evaluation effort and stopping [1, 5, 21]. To address this gap, we propose `DialBandit`. At its core, `DialBandit` combines GP-based learning of fidelity-dependent observation noise from sparse unbiased labels with uncertainty-aware planning for robust ordering and suffix-threshold stopping. Our contributions are summarized as follows:

- We formalize *online sequential search with tunable-fidelity probes*, a novel sequential-search setting in which evaluation-side compute adaptively controls observation noise rather than the underlying response quality. This addresses the inefficiency of fixed evaluation budgets, which can lead to either wasted compute or premature stopping.
- For the oracle setting, we derive a generalized index policy that couples evaluation fidelity with stage-wise stopping. Building on this structure, we propose `DialBandit`, which learns arm-specific fidelity-noise curves from sparse replicate-based unbiased labels and translates them into robust ordering, fidelity, and stopping decisions under model uncertainty, avoiding the bias of naive residual-based noise estimation under adaptive sampling.
- We prove a finite-time sublinear regret bound of $\tilde{\mathcal{O}}(\mathrm{poly}(K)T^{3/4})$ for `DialBandit`. The analysis characterizes how joint mean-and-noise learning perturbs robust ordering and stopping, and traces the $T^{3/4}$ rate to the extra exploration required to identify fidelity-dependent noise from only $\Theta(\sqrt{T})$ unbiased labels.
- We evaluate `DialBandit` on semi-empirical LLM-routing environments built from repeated judging and calibrated noisy evaluations. Across benchmarks, it consistently improves net utility up by 21.9%-31.1% over strong baselines while exhibiting adaptive fidelity allocation and efficient stopping.

2 Related Work

We position our work within four research threads, with a broader discussion deferred to Appendix D.

Online sequential search. Pandora's Box [15] and its online variants [16, 17] provide the closest sequential-search structure with inspection costs, but they rigidly assume a *fixed* observation model. We instead let evaluation fidelity control observation reliability, coupling probing value and stopping.

Multi-fidelity learning. Standard multi-fidelity methods use low-fidelity queries as cheap, typically biased approximations of a target objective [22, 19, 20]. In contrast, in our setting, fidelity alters the observation *noise* rather than the underlying response quality itself. While related GP-based methods provide confidence bounds for expensive optimization [23, 24], our approach leverages GP learning not to optimize an objective across fidelities, but as a subroutine to estimate fidelity-dependent reliability for robust sequential stopping.

LLM-as-a-Judge and evaluation-side compute. Automated judges are scalable but suffer from both systematic bias and stochastic noise [25–27]. While prior work extensively studies calibration for the former [12–14] and deploys extra evaluation-side compute (e.g., repeated sampling, ensembles) to mitigate the latter [6, 8–10], these interventions are typically static. We instead treat this compute-driven reliability as an online control variable inside a cascade policy.

Efficient LLM routing and cascades. Prior work reduces inference costs through static heuristics [1], offline routing [5], or online bandits [21]. These methods adapt model selection but usually keep evaluation fixed, whereas we jointly optimize model selection, evaluation fidelity, and stopping.

3 Problem Formulation

We begin by defining a single-round sequential search problem, and then repeat it over T rounds to obtain the online learning setting. In each round $t \in [T]$, a user query arrives and the learner sequentially probes arms (LLMs) from a candidate set $[K] = \{1, \dots, K\}$.

Observation model. Each arm i has a population mean quality $\mu_i \in [0, 1]$, representing its deployment-level average performance. On round t, its query-specific latent quality is drawn independently as $Q_{t,i} \sim \mathcal{N}(\mu_i, \kappa_i^2)$, where $\kappa_i^2 > 0$ captures *cross-query heterogeneity*. A probe does not observe $Q_{t,i}$ directly. Instead, evaluating arm i at fidelity $\alpha \in [0, 1]$ yields a noisy measurement:

$$Y_{t,i}(\alpha) \mid Q_{t,i} \sim \mathcal{N}(Q_{t,i}, \sigma_i^2(\alpha)), \tag{1}$$

where $\sigma_i^2 : [0, 1] \to [\sigma_{\min}^2, \Sigma_{\max}]$ is an unknown arm-specific *observation-noise* function, monotonically decreasing in α. This function captures stochastic evaluation instability across different models' output distributions; crucially, allocating higher fidelity (*e.g.,* via repeated judging or larger voting pools) reduces this noise without altering the underlying response quality. Here, we focus on the online learning of μ_i and the runtime fidelity-dependent noise curve $\sigma_i^2(\alpha)$, while treating κ_i^2 as an offline-calibrated arm profile (its online learning of κ_i^2 is deferred to Appendix K). Marginally, the observation follows $Y_{t,i}(\alpha) \sim \mathcal{N}(\mu_i, \kappa_i^2 + \sigma_i^2(\alpha))$.

Posterior score. Upon observing a realization y of $Y_{t,i}(\alpha)$, which is only a noisy measurement of the latent quality $Q_{t,i}$, we use the posterior mean of $Q_{t,i}$ as the decision score. With a scaling factor $\lambda > 0$ to convert quality into utility, this score is

$$S_i(y, \alpha) = \lambda \, \mathbb{E}[Q_{t,i} \mid Y_{t,i}(\alpha) = y] = \lambda \left(w_i(\alpha) y + (1 - w_i(\alpha)) \mu_i \right), \tag{2}$$

where the shrinkage factor $w_i(\alpha) = \kappa_i^2 / (\kappa_i^2 + \sigma_i^2(\alpha))$ measures how much the observation should be trusted relative to the arm-level prior mean. Higher fidelity decreases $\sigma_i^2(\alpha)$, increases $w_i(\alpha)$, and thus places more weight on the observed score y. Crucially, separating intrinsic query heterogeneity (κ_i^2) from evaluation noise ($\sigma_i^2(\alpha)$) prevents *over-escalation*: if a model's underlying response is inherently poor, allocating more compute merely yields a highly confident score. This ensures the cascade stops efficiently instead of wasting compute on unpromising probes.

Sequential policy and search costs. Within round t, the learner executes a sequential search policy π_t over stages $m = 1, 2, \dots$. Let $\mathcal{O}_{t,m-1}$ denote the set of arms probed before stage m. At stage m, the policy either *STOP*s or *PROBE*s an unobserved arm $i_m \in [K] \setminus \mathcal{O}_{t,m-1}$ at a chosen fidelity $\alpha_m \in [0, 1]$. Probing arm i_m incurs a deterministic, known total cost $c_{i_m}^{\mathrm{g}} + c^{\mathrm{e}}(\alpha_m)$, where c_i^{g} is

the arm-specific generation cost and $c^{\mathrm{e}}(\cdot)$ is a shared evaluation-side fidelity cost. The cascade tracks the best-so-far posterior score as its baseline utility $z_{t,m-1} = \max_{j \in \mathcal{O}_{t,m-1}} S_j(y_{t,j}, \alpha_{t,j})$, with $z_{t,0} = -\infty$ enforcing at least one probe per query.

Single-round utility. When π_t issues *STOP*, let $\mathcal{O}_t \subset [K]$ be the set of probed arms and let $C_t = \sum_{i \in \mathcal{O}_t} \big(c_i^{\mathrm{g}} + c^{\mathrm{e}}(\alpha_{t,i})\big)$ be the total probing cost. The round-t net utility is

$$U_t(\pi_t) = \max_{j \in \mathcal{O}_t} S_j(y_{t,j}, \alpha_{t,j}) - C_t, \tag{3}$$

and its expectation is $W_t = \mathbb{E}[U_t(\pi_t)]$. Given the observations, maximizing posterior score is equivalent to maximizing posterior expected latent quality under the evaluation protocol.

Repeated over T rounds, the learner seeks to maximize cumulative expected utility $\sum_{t=1}^{T} W_t$.

Definition 3.1 (Regret)**.** Let π_t^* denote the oracle optimal policy for round t under the true parameters $\{\mu_i, \kappa_i^2, \sigma_i^2(\cdot)\}_{i=1}^K$, and let $W_t^* = \mathbb{E}[U_t(\pi_t^*)]$. The cumulative regret is

$$\mathrm{Reg}_T = \sum_{t=1}^{T} (W_t^* - W_t). \tag{4}$$

Standing regularity assumptions. We assume $\sigma_i^2(\cdot)$ is L_σ-Lipschitz on $[0,1]$ and belongs to a known Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ with $\|\sigma_i^2\|_{\mathcal{H}_k} \le B$ for some known constant $B > 0$. We further assume $0 < \kappa_{\min}^2 \le \kappa_i^2 \le \kappa_{\max}^2$, $c_i^{\mathrm{g}} \in [c_{\min}^{\mathrm{g}}, c_{\max}^{\mathrm{g}}]$, and $c^{\mathrm{e}}(\cdot)$ is L_c-Lipschitz and non-decreasing on $[0,1]$. Equivalently, the total probing cost is uniformly bounded in $[c_{\min}, c_{\max}]$.

4 Oracle Index-Based Policy

To benchmark the online learning problem, we consider an *oracle case* in which the true parameters $\{\mu_i, \sigma_i^2(\cdot)\}_{i=1}^K$ and the shared evaluation-cost curve $c^{\mathrm{e}}(\cdot)$ are known. Each round then reduces to a single-round planning problem with sequential probing and fidelity selection. This problem admits an *index* structure that generalizes Weitzman's rule [15] by optimizing over fidelity at each probe.

Single-arm net gain. Fix arm i and best-so-far posterior score baseline z. Define $\mathcal{A} \subseteq [0,1]$ as the feasible set of fidelity. Let $X_i(\alpha) = S_i(Y_{t,i}(\alpha), \alpha)$ denote the random posterior score from probing arm i at fidelity α. The net gain of the probing arm i at fidelity $\alpha \in \mathcal{A}$ is given by:

$$V_i(z, \alpha) = \mathbb{E}_{X_i(\alpha)}[\max(X_i(\alpha), z)] - z - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha). \tag{5}$$

Let $V_{i,\mathcal{A}}^*(z) = \max_{\alpha \in \mathcal{A}} V_i(z, \alpha)$ be the optimal gain. Thus, fidelity is optimized locally per probe, while the maximized index determines the global search order and stopping threshold. In a bounded baseline domain $z \in Z$, we assume $V_{i,\mathcal{A}}^*(z_1) - V_{i,\mathcal{A}}^*(z_2) \ge L_z(z_2 - z_1)$ for all $z_1 < z_2$ in Z, where $L_z \in (0,1)$, ensuring that the break-even threshold is Lipschitz-stable in the model parameters.

Index interpretation. Under the standing regularity assumptions, $V_{i,\mathcal{A}}^*(z)$ is continuous and strictly decreasing in z, so each arm admits a unique break-even *index* $\xi_{i,\mathcal{A}}$ satisfying $V_{i,\mathcal{A}}^*(\xi_{i,\mathcal{A}}) = 0$; we interpret this as the reservation value of arm i (see Appendix E.1). Intuitively, the resulting policy follows a generalized Weitzman's rule: sort arms by index and continue probing until the current best score exceeds the indices of all remaining arms.

Theorem 4.1 (The Generalized Weitzman's Rule [15])**.** *Given $\{\mu_i, \sigma_i^2(\cdot)\}_{i=1}^K$, let ξ_i be the index defined above. Then an optimal policy is: for any state $(z, \mathcal{U})$, stop iff $z \ge \max_{i \in \mathcal{U}} \xi_i$; otherwise probe any arm $i^* \in \arg\max_{i \in \mathcal{U}} \xi_i$ and choose any $\alpha \in \arg\max_{\alpha \in \mathcal{A}} V_{i^*}(z, \alpha)$. A full description (named `IdxPolicy`) and the proof are deferred to Appendix E.2.*

5 `DialBandit` for Online Learning

Moving from the oracle benchmark to the online setting, the learner must jointly estimate arm quality and fidelity-dependent noise while planning under uncertainty. We propose `DialBandit` (Figure 2 and Algorithm 1), which maintains empirical mean estimates and GP priors over the arm-specific noise functions. Each subsequent round operates in three phases: (i) *Model Construction*: assembling

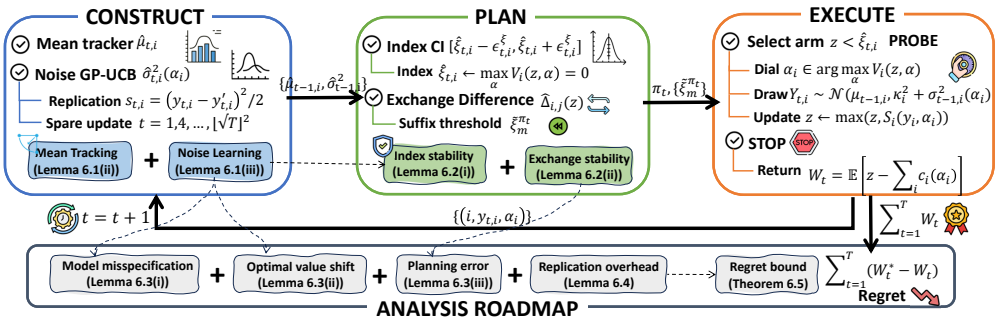


Figure 2: Overview of `DialBandit` framework and roadmap of regret analysis.

Algorithm 1 `DialBandit`

Input: Arms $[K]$, Horizon T, Grids $\mathcal{A}, \mathcal{Z}$, Schedules $\{\beta_n\}, \mathcal{I}(t)$.
Initialize empirical means and GP priors with $\hat{\mu}_{0,i} = 0$ and $\mathcal{D}^{\sigma}_{0,i} = \emptyset$ for all arms.
for $t = 1, \ldots, T$ **do**
  Compute $\hat{\sigma}^2_{t-1,i}(\cdot)$ via Eq. (7) and form model $\hat{\mathcal{M}}_t$; ▷ *Phase 1: Model Construction*
  Compute planning radii $\epsilon^{\xi}_{t,i}$ and $\epsilon^{\Delta}_{t,i,j}$; ▷ *Phase 2: Robust Planning*
  $(\{\xi_{t,i}\}, \pi_t) \leftarrow$ `SeqPolicy`$(\hat{\mathcal{M}}_t, \{\epsilon^{\xi}_{t,i}\}, \{\epsilon^{\Delta}_{t,i,j}\}, \mathcal{A}, \mathcal{Z})$;
  Execute $\mathcal{X}_t \leftarrow$ `IdxPolicy`$(\pi_t, \{\xi_{t,i}\})$, and update $\hat{\mu}_{t,i}$ via Eq. (6); ▷ *Phase 3: Execution & Update*
  if $\mathcal{I}(t) = 1$ **then**
    Use replication for $(i, \alpha) \in \mathcal{X}_t$ and the explore-only probe $(i_t^{\exp}, \alpha_t^{\exp})$ to update GP $(m^{\sigma}_{t,i}, v^{\sigma}_{t,i})$;

an observation model $\hat{\mathcal{M}}_t$ from empirical means and GP-based noise confidence bounds; (ii) *Planning*: computing confidence radii to construct a robust probing order π_t via exchange-difference tests, alongside stage-wise stopping thresholds $\{\hat{\xi}_{t,i}\}$; and (iii) *Execution & Update*: running the policy to update mean estimates, while conditionally performing sparse replication probes to acquire unbiased noise labels for GP updates. Crucially, estimation errors propagate into planning via corrupted exchange differences (affecting ordering) and inaccurate index estimates (affecting stopping).

5.1 Phase 1: uncertainty-aware model construction

At each round t, the uncertainty-aware observation model $\hat{\mathcal{M}}_t = \{(\hat{\mu}_{t-1,i}, \hat{\sigma}^2_{t-1,i}(\cdot))\}_{i=1}^{K}$ is constructed from data collected up to round $t-1$. The main difficulty is endogenous sampling: naive squared-residual proxies $(y - \hat{\mu})^2$ mix cross-query heterogeneity, mean-estimation error, and observation noise, adding a $(\mu - \hat{\mu})^2$ bias term. We therefore use sparse *replication* to obtain unbiased labels for evaluation noise. Specifically, define an update indicator $\mathcal{I}(t) = 1$ if $t \in \{1^2, 2^2, \ldots, \lfloor\sqrt{T}\rfloor^2\}$, and let $\mathcal{O}_t$ be the probed arm set in round t. After executing the round, if $\mathcal{I}(t) = 1$, then for each arm $i \in \mathcal{O}_t$ probed at fidelity $\alpha_{t,i}$, we request a second conditionally independent evaluation sample $y'_{t,i}(\alpha_{t,i}) \sim \mathcal{N}(Q_{t,i}, \sigma_i^2(\alpha_{t,i}))$ on the same query while reusing the same generated response.

Mean tracking for arm quality. μ_i denotes the deployment-time average quality of arm i. This phase uses the pre-round estimate $\hat{\mu}_{t-1,i}$, while after collecting the primary observations in round t, $\hat{\mu}_{t,i}$ will be updated as:

$$\hat{\mu}_{t,i} = \frac{\sum_{\tau \le t : i \in \mathcal{O}_\tau} y_{\tau,i}}{N^{\mu}_{t,i}}, \tag{6}$$

where $N^{\mu}_{t,i} = |\{\tau \le t : i \in \mathcal{O}_\tau\}|$, and $\hat{\mu}_{t,i}$ becomes available for constructing $\hat{\mathcal{M}}_{t+1}$.

GP-UCB learning for evaluation noise. We learn the observation-noise function $\sigma_i^2(\cdot)$ via GP regression using replicate-based labels $\tilde{s}_{t,i} = (y_{t,i} - y'_{t,i})^2/2$. Conditioned on $Q_{t,i}$, these labels satisfy $\mathbb{E}[\tilde{s}_{t,i}] = \sigma_i^2(\alpha_{t,i})$. These labels isolate evaluation noise from both the arm mean μ_i and the cross-query heterogeneity κ_i^2. We append $(\alpha_{t,i}, \tilde{s}_{t,i})$ to a replication dataset $\mathcal{D}^{\sigma}_{t,i}$ of size $N^{\sigma}_{t,i}$ and use the GP posterior mean $m^{\sigma}_{t-1,i}(\alpha)$ and variance $v^{\sigma}_{t-1,i}(\alpha)$ to form the upper confidence estimate

$$\hat{\sigma}^2_{t-1,i}(\alpha) = m^{\sigma}_{t-1,i}(\alpha) + \sqrt{\beta_{N^{\sigma}_{t-1,i}}}\sqrt{v^{\sigma}_{t-1,i}(\alpha)}, \tag{7}$$

Algorithm 2 `SeqPolicy`

Input: $\hat{\mathcal{M}}_t$, $\{\epsilon^{\xi}_{t,i}\}$, $\{\epsilon^{\Delta}_{t,i,j}\}$, $\mathcal{A}$, $\mathcal{Z}$.
for each arm $i \in [K]$ **do**
  Compute $\hat{\xi}_{t,i}$ by $V^*_{i,\mathcal{A}}(\hat{\xi}_{t,i}) = 0$; Set $\mathrm{CI}_{t,i} \leftarrow [\hat{\xi}_{t,i} - \epsilon^{\xi}_{t,i}, \hat{\xi}_{t,i} + \epsilon^{\xi}_{t,i}]$;
Define $\succ^{\mathrm{base}}_t$ by sorting $\hat{\xi}_{t,i}$, and set $\prec_t \leftarrow \emptyset$;
for each pair (i, j) with $\mathrm{CI}_{t,i} \cap \mathrm{CI}_{t,j} \neq \emptyset$ **do**
  Compute $\underline{\Delta}_{ij} \leftarrow \min_{g \in \mathcal{Z}} \hat{\Delta}_{ij}(g) - 2\eta_z$;
  if $\underline{\Delta}_{ij} \leq -(\epsilon^{\Delta}_{t,i,j} + 4\eta_z)$ **then** add $(j \prec_t i)$; **else if** $\underline{\Delta}_{ij} \geq \epsilon^{\Delta}_{t,i,j} + 4\eta_z$ **then** add $(i \prec_t j)$;
$\pi_t \leftarrow \mathrm{LinearExtension}(\prec_t)$ tie-broken by $\succ^{\mathrm{base}}_t$;
for $m = K, \ldots, 1$ **do**
  Compute $\widetilde{\xi}^{\pi_t,\mathcal{A}}_m$ by $\widetilde{V}^{\pi_t,\mathcal{A}}_m(z) = 0$; Set $\xi_{t,\pi_{t,m}} \leftarrow \widetilde{\xi}^{\pi_t,\mathcal{A}}_m$;
return $(\{\xi_{t,\pi_{t,m}}\}_{m=1}^{K}, \pi_t)$;

following standard GP-UCB analysis [23]. To accelerate posterior contraction, we additionally issue an auxiliary explore-only probe selecting $(i^{\mathrm{exp}}_t, \alpha^{\mathrm{exp}}_t) \in \arg\max_{(i,\alpha)\in[K]\times\mathcal{A}} \sqrt{v^{\sigma}_{t-1,i}(\alpha)}$; its overhead is bounded in Lemma 6.4 and detailed GP updates are deferred to Appendix F.1.

5.2 Phase 2 & 3: robust planning and execution

Given the estimated model $\hat{\mathcal{M}}_t$, we plan on a fidelity grid $\mathcal{A}$ and a baseline grid $\mathcal{Z}$, whose covering radii η_α and η_z are defined in Appendix Definitions B.1 and B.2. Algorithm 2 (`SeqPolicy`) then performs two coupled tasks: robust ordering and stage-wise thresholding.

Robust ordering. For each arm, we solve $V^*_{i,\mathcal{A}}(\hat{\xi}_{t,i}) = 0$ to obtain an estimated stopping index $\hat{\xi}_{t,i}$ and its uncertainty radius $\epsilon^{\xi}_{t,i}$, yielding the confidence interval $\mathrm{CI}_{t,i}$ (Lemma 6.2(i)). When two arms have overlapping intervals, index information alone does not certify an order, so we compare them through the *exchange difference* $\Delta_{ij}(z) = R_{ij}(z) - R_{ji}(z)$ (formalized in Appendix H.2), which measures the preference for probing i before j at baseline utility z. Evaluating $\hat{\Delta}_{ij}$ on $\mathcal{Z}$ gives the conservative lower bound $\underline{\Delta}_{ij} = \min_{g\in\mathcal{Z}} \hat{\Delta}_{ij}(g) - 2\eta_z$ (Appendix H.4). Here $\epsilon^{\Delta}_{t,i,j}$ is the estimation radius for this pairwise comparison (Lemma 6.2(ii)). If $\underline{\Delta}_{ij} > \epsilon^{\Delta}_{t,i,j} + 4\eta_z$, we set $i \prec_t j$; if $\underline{\Delta}_{ij} < -(\epsilon^{\Delta}_{t,i,j} + 4\eta_z)$, we set $j \prec_t i$. A topological sort then produces a valid linear extension π_t.

Thresholding and execution. With π_t fixed, `SeqPolicy` computes stage-wise suffix stopping thresholds by backward induction. Let $W^{\pi_t}_{K+1}(z) = z$, and the suffix net gain at stage m for arm $i = \pi_{t,m}$ is $\widetilde{V}^{\pi_t,\mathcal{A}}_m(z) = \max_{\alpha\in\mathcal{A}}\Big(\mathbb{E}_{\hat{\mathcal{M}}_t}[W^{\pi_t}_{m+1}(\max(z, S_i(Y_{t,i}(\alpha), \alpha)))] - c^{\mathrm{g}}_i - c^{\mathrm{e}}(\alpha)\Big) - z$. The threshold $\xi_{t,i}$ is the root of $\widetilde{V}^{\pi_t,\mathcal{A}}_m(z) = 0$. In this phase, `IdxPolicy` executes π_t and stops once the realized baseline utility satisfies $z \geq \xi_{t,i}$.

6 Theoretical Analysis

Model assumptions. Recall assumptions in Section 3: independent arms, known κ^2_i, bounded costs, Lipschitz $c^{\mathrm{e}}(\cdot)$, and monotone RKHS noise curves $\sigma^2_i(\cdot)$ with $\|\sigma^2_i\|_{\mathcal{H}_k} \leq B$. Primary observations are sub-Gaussian with proxy $V_{\max} = \kappa^2_{\max} + \Sigma_{\max}$; replicate-label noise is bounded on the clean event, enabling GP-UCB confidence bounds. Let $\mathcal{M}$ and $\hat{\mathcal{M}}_t$ denote true and estimated models.

Analysis roadmap. We establish the sublinear regret of `DialBandit` in three stages: (i) bounding the statistical estimation errors of arm parameters (Lemma 6.1), (ii) tracing the propagation of these errors through the Lipschitz properties of the planning indices (Lemma 6.2), and (iii) aggregating the induced single-round suboptimality alongside the explicit exploration overhead (Lemmas 6.3–6.4). This yields the final cumulative regret in Theorem 6.5.

6.1 Estimation and planning stability

Here, we formalize the clean event $\mathcal{E}$, and give the planning stability lemma.

Lemma 6.1 (Concentration on the Clean Event)**.** *Under the model assumptions, for any $\rho \in (0,1)$, there exists a clean event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1-\rho$ such that, for all $i \in [K]$: (i)* ***Observation Boundedness:*** *All realized observations are bounded by a threshold $Y_{\max}(\rho) = \mathcal{O}(\sqrt{\log(T)})$. (ii)* ***Mean Tracking:*** *Empirical means concentrate as $|\hat{\mu}_{t-1,i} - \mu_i| \leq B_{N^{\mu}_{t-1,i}} = \mathcal{O}(1/\sqrt{N^{\mu}_{t-1,i}})$. (iii)* ***Noise Learning:*** *The GP posterior bounds the true noise as $\sup_{\alpha\in[0,1]} |\hat{\sigma}^2_{t-1,i}(\alpha) - \sigma^2_i(\alpha)| \leq B_{N^{\sigma}_{t-1,i}}$, where $B_{N^{\sigma}_{t-1,i}}$ abbreviates the uniform GP radius $2\sqrt{\beta_{N^{\sigma}_{t-1,i}}}\sup_{\alpha}\sqrt{v^{\sigma}_{t-1,i}(\alpha)}$.*

Proof is deferred to Appendices G.1–G.4.

By Lemma 6.1, we have $Z = [-S_{\max}, S_{\max}]$ with $S_{\max} = \lambda(Y_{\max}+1)$. Within this, estimation errors and discrete grid approximations (with covering radii η_α and η_z) propagate into the sequential planning via two channels: the stopping indices ξ_i and the pairwise exchange differences Δ_{ij}.

Lemma 6.2 (Planning Stability)**.** *Fix pair $\{i,j\}$, round t, and covering radius η_α. On $\mathcal{E}$, we have: (i)* ***Index Stability:*** *$|\xi_i - \hat{\xi}_{t,i}| \leq \epsilon^{\xi}_{t,i} = \mathcal{O}\big(B_{N^{\mu}_{t-1,i}} + B_{N^{\sigma}_{t-1,i}} + \eta_\alpha\big)$. (ii)* ***Exchange Stability:*** *$\sup_{z\in Z} |\Delta_{ij}(z;\mathcal{M}) - \hat{\Delta}_{ij}(z;\hat{\mathcal{M}}_t)| \leq \epsilon^{\Delta}_{t,i,j} = \mathcal{O}\big(\max_{k\in\{i,j\}} B_{N^{\mu}_{t-1,k}} + \max_{k\in\{i,j\}} B_{N^{\sigma}_{t-1,k}} + \eta_\alpha\big)$.*

See Appendices H.2–H.5 for the proof.

6.2 Regret bound

For notational convenience, define $\Psi^{\mu}_t = \sum_{j=1}^{K} B_{N^{\mu}_{t-1,j}}$ and $\Psi^{\sigma}_t = \sum_{j=1}^{K} B_{N^{\sigma}_{t-1,j}}$.

Lemma 6.3 (Single-Round Suboptimality)**.** *Fix round t. Let $\hat{\pi}_t$ be our executed policy and $W^*(\mathcal{M})$ the oracle optimal value. On event $\mathcal{E}$, the suboptimality decomposes into: (i)* ***Model Misspecification:*** *$|W(\mathcal{M};\pi) - W(\hat{\mathcal{M}}_t;\pi)| \leq \mathcal{O}(K(\Psi^{\mu}_t + \Psi^{\sigma}_t))$ for any fixed π. (ii)* ***Optimal Value Shift:*** *$|W^*(\mathcal{M}) - W^*(\hat{\mathcal{M}}_t)| \leq \mathcal{O}(K(\Psi^{\mu}_t + \Psi^{\sigma}_t))$. (iii)* ***Planning Error:*** *Positioning arms in $\hat{\pi}_t$ via adjacent swaps [16] incurs a bounded sorting penalty $W^*(\hat{\mathcal{M}}_t) - W(\hat{\mathcal{M}}_t;\hat{\pi}_t) \leq \mathcal{O}\big(K^2(\Psi^{\mu}_t + \Psi^{\sigma}_t + \eta_\alpha + \eta_z)\big)$.*

See Appendices I.2–I.4 for the induction over the value recursion.

Lemma 6.4 (Replication Overhead)**.** *Let $\varrho_t \geq 0$ be the extra cost in round t from sparse replication, including forced noise-exploration. Then $\sum_{t=1}^{T} \mathbb{E}[\varrho_t] \leq (\lfloor\sqrt{T}\rfloor + 1)(K+1)c_{\max}$.*

See Appendix I.5 for the proof.

Theorem 6.5 (Regret Bound)**.** *Let $M = \lfloor\sqrt{T}\rfloor$ and $\bar{\gamma}_M = \max_i \gamma_{M,i}$. With probability at least $1-\rho$, the clean event $\mathcal{E}$ holds, and on this event the regret of* `DialBandit` *satisfies*

$$\mathrm{Reg}_T \leq \tilde{\mathcal{O}}\Big(K^4\sqrt{T} + K^4 T^{3/4}\sqrt{\bar{\gamma}_M} + K^2 T(L_\alpha\eta_\alpha + \eta_z) + Kc_{\max}\sqrt{T}\Big). \tag{8}$$

In particular, taking $\eta_\alpha = \eta_z = T^{-1/4}$ gives $\mathrm{Reg}_T \leq \tilde{\mathcal{O}}\big(\mathrm{poly}(K)T^{3/4}(1+\sqrt{\bar{\gamma}_M})\big)$.

Proof sketch. On $\mathcal{E}$, estimation errors in $\hat{\mu}_i$ and $\hat{\sigma}^2_i(\cdot)$ perturb the oracle policy only through index and exchange-difference. Lemma 6.3 converts these perturbations into a one-round policy gap by the mean and GP noise radii. Summing the mean radii gives the standard $\tilde{\mathcal{O}}(K\sqrt{T})$ term. For noise learning, updates occur only at $M = \lfloor\sqrt{T}\rfloor$ sparse replication epochs; each epoch affects at most $\tilde{\mathcal{O}}(\sqrt{T})$ rounds, while GP information gain gives $\sum_{m\leq M} u_m \leq \tilde{\mathcal{O}}(\sqrt{M\bar{\gamma}_M})$, yielding the $\tilde{\mathcal{O}}(KT^{3/4}\sqrt{\bar{\gamma}_M})$ accumulated noise-radius term. The additional K^3 comes from stage-wise propagation of ordering and stopping errors. The remaining terms are the discretization error $T(L_\alpha\eta_\alpha + \eta_z)$ and the replication overhead $\mathcal{O}(Kc_{\max}\sqrt{T})$. See Appendix I.6 for the full decomposition.

Discussion and limitations. While the sublinear regret bound confirms the efficiency of `DialBandit`, practical deployment involves several key considerations. First, the Gaussian observation model gives a closed-form posterior score; the concentration arguments mainly require bounded or sub-Gaussian tails. Second, sparse noise updates create an overhead-regret tradeoff: aggressive sparsity minimizes auxiliary probe costs but delays GP posterior contraction, leading to the $T^{3/4}$ rate. Third, replicate-based labels avoid the mean-induced bias of residual noise proxies $(y-\hat{\mu})^2$. Fourth, although κ^2_i is

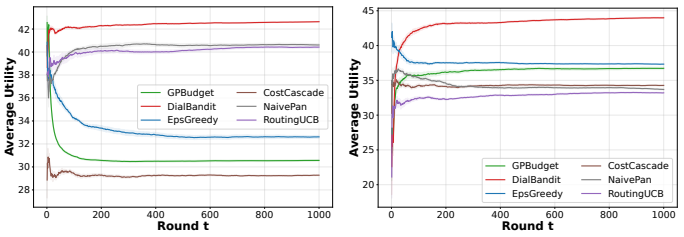


Figure 3: Average net utility (AlpacaEval and MMLU).

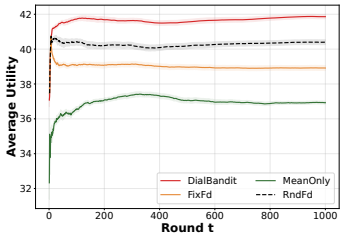


Figure 4: Ablation.

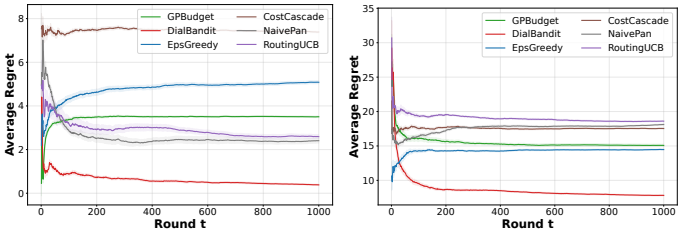


Figure 5: Average regret (AlpacaEval and MMLU).

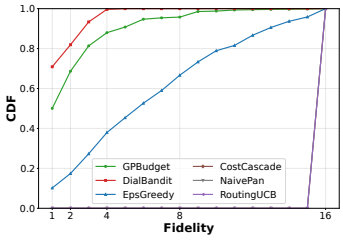


Figure 6: Adaptive fidelity.

treated as an offline-calibrated profile, Appendix K extends the analysis to online learning of κ_i^2 while preserving the $\tilde{\mathcal{O}}(T^{3/4})$ regret. Finally, our analysis is stationary and non-contextual, isolating how fidelity-dependent evaluation noise should be learned and used for sequential stopping. Contextual routing would require query-conditioned quality models and contextual regret. Separately, under non-stationary query distributions (*e.g.,* shifting prompt difficulty), our replicate-based labels remain unbiased for instantaneous evaluation noise, though full analysis would require dynamic regret and time-local estimators. See Appendix I.7 and K for extended discussions.

7 Experiments

We evaluate `DialBandit` on two semi-empirical environments: one based on repeated judging (AlpacaEval) and one based on calibrated noisy logits (MMLU). We aim to answer: **(Q1)** Does `DialBandit` improve overall net utility and reduce regret over strong baselines, and are these gains supported by its key algorithmic components? **(Q2)** Does the learned policy exhibit adaptive fidelity allocation and efficient within-round stopping? **(Q3)** Does the sparse noise-learning schedule both sharpen the planning model over time and lead to vanishing asymptotic overhead? **(Q4)** Does `DialBandit` achieve a competitive cost-quality tradeoff relative to baselines?

7.1 Experimental setup

Our environments use seven open-weight LLMs as candidate arms (Mistral-7B [28], Gemma-2-9B [29], Llama3.1-8B [30], and Qwen2.5 suite [31], and Phi-3.5 [32]). In each round, the learner sequentially probes candidate arms, chooses a fidelity from a discretized grid, and stops adaptively. We run $T = 1000$ rounds over 50 random seeds, using a 16-point repeat-count grid $n \in \{1, \ldots, 16\}$ with normalized fidelity $\alpha = n/16$; unless stated otherwise, $\lambda = 54$. Details are provided in Appendix L.1.

Datasets & observation models. On AlpacaEval [33], we score offline-generated candidate responses using Prometheus-7B-v2.0 [34]. Fidelity is implemented by repeated judging: a higher level triggers more independent judge calls, reducing observation noise at higher evaluation cost. On MMLU [35], we use a controlled observation channel to study joint probing, fidelity selection, and stopping. To isolate the evaluation-side reliability-cost tradeoff, we fit arm-specific (μ_i, κ_i^2) from offline traces and then synthesize fidelity-dependent noisy observations using calibrated monotone noise curves. Utilities are reported in terms of the Bayes-consistent posterior score S_i.

Baselines. We compare against five representative strategies: (1) `GPBudget` [23]: GP-UCB over fidelity with sequential budget allocation; (2) `NaivePan`: Pandora's index [15, 36] executed at a

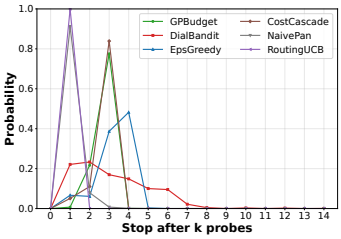


Figure 7: Stopping behavior.

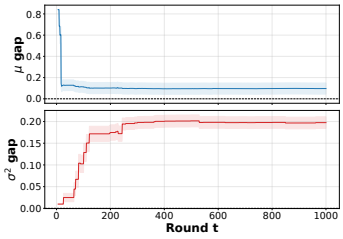


Figure 8: Dynamic learning.

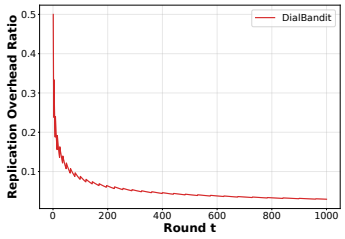


Figure 9: Replication overhead.

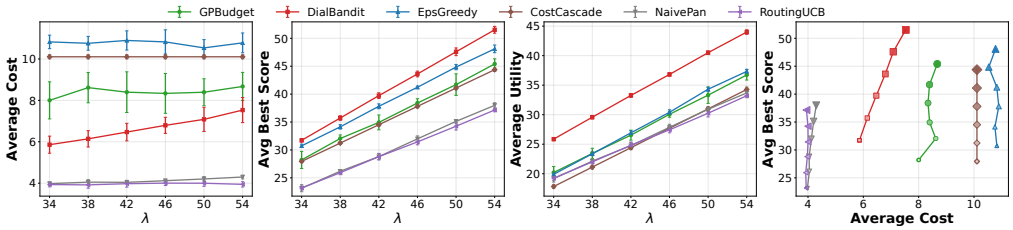


Figure 10: Cost-quality tradeoff across λ.

fixed fidelity; (3) `RoutingUCB`: one-step UCB routing without sequential search; (4) `CostCas` [1]: a fixed-order cascade with predefined stopping heuristics; (5) `EpsGreedy` [37]: standard ϵ-greedy exploration over (arm, fidelity) pairs. For ablations, `MeanOnly` removes noise-aware learning, `FixFd` uses a fixed fidelity, and `RndFd` replaces adaptive fidelity selection with random choices.

7.2 Results

Overall performance and component contribution (Q1). Figure 3 tracks the running net utility on AlpacaEval and MMLU. `DialBandit` consistently outperforms across both environments, improving final average net utility by 31.1% on AlpacaEval and 21.9% on MMLU over strong baselines. These gains emerge early, indicating superior within-round joint decisions (order, fidelity, and stopping) rather than merely larger probing budgets. It also tracks the calibrated per-round oracle closely, reducing average regret by 87.0% and 53.5% on AlpacaEval and MMLU, respectively, as shown in Figure 5. Ablation results further show that these gains rely on the joint design, where removing fidelity control or noise-aware learning degrades net utility by up to 13.3% (Figure 4).

Adaptive fidelity and stopping (Q2). On both AlpacaEval and MMLU, `DialBandit` concentrates most of its evaluation budget on low fidelities, escalating only when the evidence is ambiguous (Figure 6 and Appendix Figure 12). On MMLU, the stopping distribution places mainly on shallow searches, indicating query-dependent stopping depth rather than a fixed probing budget (Figure 7). On AlpacaEval, the policy usually stops after one or two probes, consistent with efficient early stopping in an easier environment (Appendix Figure 13).

Dynamic learning and overhead decay (Q3). Figures 8 and 9 highlight the two effects of the sparse schedule. The mean estimates stabilize quickly, while the noise estimates improve more gradually with the expected staircase pattern induced by updates on rounds $t \in \{1^2, 2^2, \ldots, \lfloor\sqrt{T}\rfloor^2\}$. At the same time, the replication-overhead ratio decays over time, matching the sparse-update prediction that auxiliary costs become lower order.

Cost-quality tradeoff (Q4). Figure 10 shows that as the quality preference λ increases, `DialBandit` moves smoothly toward moderately higher-cost operating points while delivering consistent gains in both best-score quality and average utility. In the cost-quality plane, its operating points occupy a favorable upper-left region relative to the baselines, indicating that it achieves stronger quality without requiring the highest probing cost.

8 Conclusion

This paper formalizes online sequential search with tunable-fidelity probes, inspired by LLM cascades in which evaluation-side compute can be traded for more reliable intermediate assessments under a fixed evaluation protocol. For the oracle setting, we derive a generalized index policy that couples fidelity selection with sequential stopping. For the online setting, we propose `DialBandit`, which leverages sparse auxiliary replications to learn fidelity-dependent noise curves and performs uncertainty-aware planning over probing order, evaluation fidelity, and stopping. We prove a finite-time sublinear regret bound, and show that `DialBandit` improves end-to-end net utility in semi-empirical evaluations while exhibiting adaptive fidelity selection and efficient early stopping.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China Grants 62472090, 62402102, the National Natural Science Foundation of Jiangsu Grants BK20242026, BK20241275, Key Laboratory of Computer Network and Information Integration of the Ministry of Education of China Grant 93K-9, the Jiangsu Provincial Key Laboratory of Network and Information Security Grant BM2003201, the Collaborative Innovation Center of Novel Software Technology.

Competing interests. The authors declare no competing interests.

References

[1] 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.

[2] Youhe Jiang, Fangcheng Fu, Wanru Zhao, Stephan Rabanser, Nicholas D Lane, and Binhang Yuan. Cascadia: A cascade serving system for large language models. *arXiv preprint arXiv:2506.04203*, 2025.

[3] Lei Li, Yankai Lin, Deli Chen, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. Cascadebert: Accelerating inference of pre-trained language models via calibrated complete models cascade. *arXiv preprint arXiv:2012.14682*, 2020.

[4] Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, et al. Automix: Automatically mixing language models. *Advances in Neural Information Processing Systems*, 37:131000–131034, 2024.

[5] Xuechen Zhang, Zijian Huang, Ege Onur Taga, Carlee Joe-Wong, Samet Oymak, and Jiasi Chen. Efficient contextual llm cascades through budget-constrained policy learning. *Advances in Neural Information Processing Systems*, 37:91691–91722, 2024.

[6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[7] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[8] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[9] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

[10] Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in llm-as-a-judge frameworks. *arXiv preprint arXiv:2510.27106*, 2025.

[11] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

[12] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, 2024.

[13] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, 2025.

[14] Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16537–16552, 2025.

[15] Martin Weitzman. *Optimal search for the best alternative*, volume 78. Department of Energy, 1978.

[16] Khashayar Gatmiry, Thomas Kesselheim, Sahil Singla, and Yifan Wang. Bandit algorithms for prophet inequality and pandora's box. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 462–500. SIAM, 2024.

[17] Hossein Esfandiari, MohammadTaghi HajiAghayi, Brendan Lucier, and Michael Mitzenmacher. Online pandora's boxes and bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1885–1892, 2019.

[18] Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, and Barnabas Poczos. Multi-fidelity gaussian process bandit optimisation. *Journal of Artificial Intelligence Research*, 66:151–196, 2019.

[19] Xuchuang Wang, Qingyun Wu, Wei Chen, and John Lui. Multi-fidelity multi-armed bandits revisited. *Advances in Neural Information Processing Systems*, 36:31570–31600, 2023.

[20] Justin Whitehouse, Aaditya Ramdas, and Steven Z Wu. On the sublinear regret of gp-ucb. *Advances in Neural Information Processing Systems*, 36:35266–35276, 2023.

[21] Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. Adaptive llm routing under budget constraints. *arXiv preprint arXiv:2508.21141*, 2025.

[22] Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. Multi-fidelity bayesian optimisation with continuous approximations. In *International conference on machine learning*, pages 1799–1808. PMLR, 2017.

[23] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[24] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

[25] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

[26] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[27] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, 2025.

[28] Mistral AI. Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2023.

[29] Gemma Team. Gemma. 2024.

[30] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[31] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[32] Microsoft. Phi-3.5-instruct. https://huggingface.co/microsoft/Phi-3.5-mini-instruct, 2024.

[33] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An Automatic Evaluator of Instruction-following Models, May 2023.

[34] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[35] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[36] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pages 767–776. PMLR, 2015.

[37] Nicolo Cesa-Bianchi and Paul Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *Proceedings of the fifteenth international conference on machine learning*, pages 100–108, 1998.

[38] OpenAI. https://openai.com/, 2026.

[39] Qwen Team. Qwen3 technical report, 2025.

[40] Arpit Agarwal, Rohan Ghuge, and Viswanath Nagarajan. Semi-bandit learning for monotone stochastic optimization. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1260–1274. IEEE, 2024.

[41] Danny Segev and Sahil Singla. Efficient approximation schemes for stochastic probing and prophet problems. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 793–794, 2021.

[42] Yusuf Kalayci, Vinod Raman, and Shaddin Dughmi. Optimal stopping vs best-of-n for inference time optimization. *arXiv preprint arXiv:2510.01394*, 2025.

[43] Guangya Wan, Zixin Stephen Xu, Sasa Zorc, Manel Baucells, Mengxuan Hu, Hao Wang, and Sheng Li. Beacon: Bayesian optimal stopping for efficient llm sampling. *arXiv preprint arXiv:2510.15945*, 2025.

[44] Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R Collins, Jeff Schneider, Barnabas Poczos, and Eric P Xing. Tuning hyperparameters without grad students: Scalable and robust bayesian optimisation with dragonfly. *Journal of Machine Learning Research*, 21(81):1–27, 2020.

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

[46] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 232–238, 2013.

[47] Semih Cayci, Atilla Eryilmaz, and Rayadurgam Srikant. Budget-constrained bandits over general cost and reward distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4388–4398. PMLR, 2020.

[48] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522, 2023.

[49] Ferdi Kossmann, Ziniu Wu, Alex Turk, Nesime Tatbul, Lei Cao, and Samuel Madden. Cascadeserve: Unlocking model cascades for inference serving. *arXiv preprint arXiv:2406.14424*, 2024.

A Broader impacts

By adaptively allocating high-fidelity evaluations only when they are most useful, `DialBandit` can reduce evaluation-side compute, latency, and energy use in LLM pipelines. At the same time, our framework optimizes variance reduction under a fixed evaluation protocol; if the evaluator has systematic biases, the policy may make biased judgments more confidently. Deployments, especially in high-stakes settings, should therefore pair fidelity control with bias audits, calibration, and other evaluator-safety checks.

B Supplementary definitions

Definition B.1 (Fidelity Grid)**.** A finite set $\mathcal{A} \subset [0,1]$ has a covering radius η_α if for each $\alpha \in [0,1]$ there exists $g(\alpha) \in \mathcal{A}$ such that $|\alpha - g(\alpha)| \le \eta_\alpha$. In the analysis, we take $\eta_\alpha = T^{-1/4}$, equivalently $|\mathcal{A}| = \tilde{\mathcal{O}}(T^{1/4})$.

Definition B.2 (Baseline Grid)**.** A finite set $\mathcal{Z} \subset Z$ has a covering radius η_z if for each $z \in Z$ there exists $g(z) \in \mathcal{Z}$ such that $|z - g(z)| \le \eta_z$. In the analysis, we take $\eta_z = T^{-1/4}$, equivalently $|\mathcal{Z}| = \tilde{\mathcal{O}}(T^{1/4})$.

C Details of motivating example

This section provides the empirical setup supporting Figure 1 (a), demonstrating that evaluation-side compute acts as a controllable knob for response reliability.

Environment and judge protocol. We use the SummEval dataset [11] to instantiate a judge-based cascade environment. For a random subset of 100 source documents, we evaluate 5 fixed candidate models' summaries. To expose the intrinsic instability of automated evaluation, each candidate-query pair is scored multiple times by a judge model (GPT-3.5 Turbo [38], Qwen3-30B [39], or Prometheus-7B [34]) using a consistent pointwise prompt with structured score extraction.

Fidelity and evaluation metrics. We operationalize *fidelity* n as the number of independent, repeated judgments for a given query-candidate pair (q, m). Higher n directly incurs greater evaluation-side compute (e.g., API calls, tokens, latency). Let $s_{q,m,r}$ denote the r-th stochastic judge score, with the aggregated score $\bar{s}^{(n)}_{q,m} = \frac{1}{n}\sum_{r=1}^{n} s_{q,m,r}$. We quantify evaluation reliability via two metrics: (i) *overall noise*, measured as the bootstrap standard deviation of the sample mean $\bar{s}^{(n)}_{q,m}$ aggregated across all pairs, capturing score uncertainty; and (ii) *flip rate*, defined as the frequency with which the pairwise ranking induced by $\bar{s}^{(n)}_{q,m}$ contradicts a high-fidelity reference ranking ($n_{\mathrm{ref}} = 24$, excluding ties), capturing decision instability.

Empirical insights and implications. As shown in Figure 1(a), increasing fidelity consistently drives down both overall noise and flip rate across all judges, corroborating the existence of a fidelity-noise tradeoff. Crucially, the curves reveal heterogeneity across judges. For instance, Qwen3-30B exhibits a low and flat noise where additional fidelity yields minimal benefit, whereas Prometheus-7B shows a steep curve where low fidelity causes high ranking flip but scales well with compute. This reveals two key premises of our work: (1) noise is fidelity-dependent ($\sigma^2(\alpha)$ is decreasing), and (2) the cost-noise tradeoff is strictly arm-specific ($\sigma_i^2(\alpha)$ varies drastically by model i). Rather than a limitation, this observation motivates our sequential decision framework, where we actively treat fidelity as an online control variable to dynamically balance evaluation cost against noise reduction.

D Supplementary related work in Section 2

We discuss four lines of related work that inform our problem setting: online sequential search, multi-fidelity learning, evaluation-side compute in LLM systems, and efficient LLM routing. These threads motivate complementary aspects of our formulation, but none jointly addresses tunable evaluation fidelity, online learning of observation reliability, and sequential stopping within a unified cascade policy.

Online sequential search. The Pandora's Box problem [15] is a classical model of sequential search with inspection costs, in which the decision maker opens boxes sequentially and stops when further search is no longer worthwhile. Its central insight is the optimality of index-based policies under known reward distributions. A growing literature studies online or learning-based variants, where arm distributions are initially unknown and must be learned from interaction [17, 16, 40, 41]. These works provide the closest structural foundation for our setting because they combine sequential probing, inspection costs, and stopping decisions. Recent papers have also begun to connect Pandora-style ideas to inference-time optimization in LLM systems [42, 43].

While the key difference is that these formulations assume a fixed observation model: probing an arm reveals a sample from a static distribution. In our setting, probing includes an additional control variable, *evaluation fidelity*, which changes the reliability of the observation at extra cost. This changes both the planning problem and the learning problem. On the planning side, the value of probing an arm depends not only on the arm identity and current baseline, but also on the chosen fidelity. On the learning side, the learner must estimate an unknown fidelity-dependent noise function under policy-dependent sampling, rather than only learning arm values under a fixed observation law. Thus, our work extends the sequential search from fixed observations to controllable, learnable observation reliability.

Learning with tunable fidelity and multi-fidelity evaluation. Tunable-fidelity and multi-fidelity learning have been studied extensively in Bayesian optimization, bandits, and hyperparameter optimization [22, 44, 45, 19, 20]. In these settings, lower-fidelity evaluations typically serve as cheaper but less accurate approximations to a target high-fidelity objective. The learner exploits these approximations to accelerate optimization, identify promising regions, or reduce the number of expensive high-fidelity queries. Our setting differs in both the statistical target and the downstream decision problem. Here fidelity controls the noise level of the observation used for within-round sequential stopping under a fixed evaluation protocol. Classical multi-fidelity methods use fidelity to observe *different approximations of an objective*, whereas we use fidelity to obtain *different-noise observations of the same latent quality*. This changes the learning target from an unknown mean objective to an unknown reliability function, and changes the decision problem from global optimization to sequential search-and-stop planning.

Related kernelized and GP-based methods develop confidence bounds and regret guarantees for optimizing expensive unknown objectives [23, 24]. In our framework, GP-style confidence estimates are used only as a subroutine for learning a fidelity-dependent noise function from sparse replicate-based labels, not as a direct optimizer over arm-fidelity pairs. Budgeted exploration and active querying under explicit costs [46, 47] are also related at a high level, but in our formulation the auxiliary probes used for noise learning are treated as algorithmic overhead rather than benchmark actions available to the oracle policy.

LLM-as-a-Judge and evaluation-side compute. Recent work extensively studies the use of LLMs as automated evaluators for open-ended generation tasks [25–27]. This literature highlights two distinct challenges: systematic bias (e.g., position, length, or self-enhancement biases) [12, 13] and

stochastic noise (variance). To combat stochastic noise, a main trend is to scale *evaluation-side compute*. Techniques such as repeated judging, self-consistency voting, multi-agent debate, and evaluator ensembles have proven highly effective at stabilizing scores and improving alignment with human preferences [6, 8–10, 48]. These findings motivate our treatment of evaluation fidelity as a meaningful and controllable resource.

Under a fixed evaluator and rubric, repeated judgments reduce stochastic variation around the evaluator's expected score, while stable systematic bias remains part of that score. Calibration methods address the relation between this score and external quality [14]. We study how to allocate evaluation fidelity online within a cascade, where its cost and noise level jointly shape probing, escalation, and stopping. In our formulation, judge-based evaluation induces a fidelity-dependent noise curve, forcing the algorithm to learn how noise reduction affects optimal stopping and escalation decisions within a cascade policy.

Efficient LLM inference, routing, and cascades. A large literature studies how to reduce LLM inference cost through routing, cascades, and adaptive compute. Early cascade systems such as FrugalGPT [1] and CascadeBERT [3] establish the basic sequential-escalation paradigm with fixed scoring rules, while systems such as Cascadia [2] and CascadeServe [49] focus on scaling multi-model serving in production settings. Beyond static heuristics, learning-based routers aim to adapt model selection to the query, including offline-trained routing policies [4, 5] and online routing methods [21]. These approaches differ in supervision and learning style, but they share a common assumption: the evaluation protocol is fixed, so the observed quality signal is treated as an exogenous input rather than a controllable decision variable.

Our work fundamentally differs in the evaluation signal. While existing routing frameworks, whether trained offline or online, assume the observation of model quality comes from a fixed protocol, we elevate the evaluation step itself to a tunable decision variable. Consequently, our framework goes beyond model selection: it jointly optimizes model choice, *evaluation fidelity*, and stopping within each query. By learning the fidelity-reliability tradeoff interactively, the system dynamically decides not just which model to query, but exactly how much evaluation-side compute to spend to reliably assess its response.

E Supplementary Section 4

E.1 Properties of net gain

Lemma E.1 (Properties of Net Gain)**.** *Fix arm i and any fidelity $\alpha \in [0,1]$. With standing regularity assumptions, for any α, $V_i(z,\alpha)$ and $V^*_{i,\mathcal{A}}(z)$ are both continuous and strictly decreasing in z, with* $\arg\max_{\alpha\in\mathcal{A}} V_i(z,\alpha)$ *existing for all z.*

Proof. Fix an arm i and a fidelity $\alpha \in [0,1]$. Let $Y \sim \mathcal{N}(\mu_i, \kappa_i^2 + \sigma_i^2(\alpha))$ and define $X_i(\alpha) = S_i(Y,\alpha)$. Then $X_i(\alpha) \sim \mathcal{N}(m, s^2)$ with $m = \lambda\mu_i$ and $s = \lambda w_i(\alpha)\sqrt{\kappa_i^2 + \sigma_i^2(\alpha)}$. Recall $V_i(z,\alpha) = \mathbb{E}[\max(X_i(\alpha), z)] - z - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha)$.

Strict monotonicity in z. For any real z, we have $\mathbb{E}[\max(X_i(\alpha), z)] = z\mathbb{P}(X_i(\alpha) \le z) + \mathbb{E}[X_i(\alpha)\mathbf{1}\{X_i(\alpha) > z\}]$. Using the well-known closed-form formula, $\mathbb{E}\left[X_i(\alpha)\mathbf{1}\{X_i(\alpha) > z\}\right] = m\big(1 - \Phi(a)\big) + s\phi(a)$, where $a = \frac{z-m}{s}$, Φ and ϕ are the standard normal cumulative distribution function (CDF) and the probability distribution function (PDF). Therefore,

$$\mathbb{E}[\max(X_i(\alpha), z)] = z\Phi(a) + m\big(1 - \Phi(a)\big) + s\phi(a). \tag{9}$$

Then, $\frac{d}{dz}\mathbb{E}[\max(X_i(\alpha), z)] = \Phi(a) = \mathbb{P}(X_i(\alpha) \le z)$. Hence, $\frac{\partial}{\partial z}V_i(z,\alpha) = \mathbb{P}(X_i(\alpha) \le z) - 1$. Since $X_i(\alpha)$ has a continuous density, for any finite z we have $0 < \mathbb{P}(X_i(\alpha) \le z) < 1$. Thus, $-1 < \frac{\partial}{\partial z}V_i(z,\alpha) < 0$, and $V_i(z,\alpha)$ is strictly decreasing in z.

Continuity in z. The derivative $\frac{\partial}{\partial z}V_i(z,\alpha)$ exists for all z and is bounded in $[-1,0]$. Therefore, $V_i(\cdot,\alpha)$ is 1-Lipschitz and in particular continuous in z.

Properties of $V^*_{i,\mathcal{A}}(z)$. Let $\mathcal{A} \subseteq [0,1]$ and recall $V^*_{i,\mathcal{A}}(z) = \max_{\alpha\in\mathcal{A}} V_i(z,\alpha)$. First, if $z_1 < z_2$, then for each α we have $V_i(z_1,\alpha) > V_i(z_2,\alpha)$. Taking maxima over $\alpha \in \mathcal{A}$ on both sides gives $V^*_{i,\mathcal{A}}(z_1) > V^*_{i,\mathcal{A}}(z_2)$, so $V^*_{i,\mathcal{A}}$ is strictly decreasing. Second, assume $c^{\mathrm{e}}(\cdot)$ and $\sigma_i^2(\cdot)$ are continuous on

$[0, 1]$, then Eq. (9) implies $V_i(z, \alpha)$ is continuous in (z, α), which indicates that $V^*_{i,\mathcal{A}}(z)$ is continuous in z and $\arg\max_{\alpha\in\mathcal{A}} V_i(z, \alpha)$ is non-empty. □

E.2 Algorithm description and proof of Theorem 4.1

Algorithm 3 `IdxPolicy` (full version)

Input: $\{\mu_i, \kappa_i^2, \sigma_i^2(\cdot), c_i^{\mathrm{g}}\}_{i=1}^K$, $c^{\mathrm{e}}(\cdot)$, ξ_i, π, $\mathcal{A}$.
if π is not provided **then** Set π by descending-sorting $\{\xi_i\}$;
$z \leftarrow -\infty$, $\mathcal{X} \leftarrow \emptyset$;
for each arm i in π
  if $z \geq \xi_i$ **then break**
  Choose $\alpha_i \in \arg\max_{\alpha\in\mathcal{A}} V_i(z, \alpha)$; Draw $Y_i(\alpha_i) \sim \mathcal{N}(\mu_i, \kappa_i^2 + \sigma_i^2(\alpha_i))$ and observe y_i;
  $z \leftarrow \max(z, S_i(y_i, \alpha_i))$, $\mathcal{X} \leftarrow \mathcal{X} \cup \{(i, y_i, \alpha_i)\}$;
Return $\mathcal{X}$.

Proof. Fix an arm i. By Lemma E.1, $V_i^*(z)$ is continuous and strictly decreasing in z and $\arg\max_{\alpha\in[0,1]} V_i(z, \alpha)$ is non-empty. Moreover, $V_i^*(z) \to +\infty$ as $z \to -\infty$, and by the no-free-probe assumption there exists $c_{\min} > 0$ such that for all α, $V_i(z, \alpha) \to -(c_i^{\mathrm{g}} + c^{\mathrm{e}}(\alpha)) \leq -c_{\min}$ as $z \to +\infty$. Hence, $V_i^*(z)$ crosses zero exactly once; let ξ_i be the unique root of $V_i^*(z) = 0$. Let $W(z,\mathcal{U})$ denote the optimal expected net utility of the state $(z,\mathcal{U})$ in the oracle problem. For the single-arm case $\mathcal{U} = \{i\}$, stopping yields the current baseline z, while probing the control α yields $\mathbb{E}[\max(X_i(\alpha), z)] - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha) = z + V_i(z, \alpha)$. Therefore,

$$W(z, \{i\}) = \max\Big\{z, \max_{\alpha\in[0,1]} \big(z + V_i(z, \alpha)\big)\Big\} = z + \max\{0, V_i^*(z)\}, \tag{10}$$

so probing is optimal iff $V_i^*(z) > 0$, equivalently $z < \xi_i$, and an optimal control is any $\alpha \in \arg\max_{\alpha\in[0,1]} V_i(z, \alpha)$.

For multiple arms, the oracle problem is a Pandora's Box instance in the sense of [15]. Each arm can be probed at most once; probing arm i requires choosing α and paying the known inspection cost $c_i^{\mathrm{g}} + c^{\mathrm{e}}(\alpha)$ before observing the prize; conditional on (i, α) the prize $X_i(\alpha) = S_i(Y_{t,i}(\alpha), \alpha)$ has a fixed marginal distribution independent of the other arms (since latent qualities $Q_{t,i}$ are drawn independently across arms); and the terminal payoff is $\max\{z, \max X\}$ minus the accumulated probing costs. Therefore, Weitzman's theorem applies and the reservation value of arm i is precisely the unique ξ_i satisfying $V_i^*(\xi_i) = 0$. The optimal policy is to stop iff $z \geq \max_{i\in\mathcal{U}} \xi_i$, and otherwise probe any $i^* \in \arg\max_{i\in\mathcal{U}} \xi_i$. When probing i^* at baseline z, choosing any $\alpha \in \arg\max_{\alpha\in[0,1]} V_{i^*}(z, \alpha)$ is optimal by the definition of $V_{i^*}^*(z)$. Algorithm 3 implements this policy when run with $\{\xi_i\}_{i=1}^K$. □

F Supplementary Subsection 5

F.1 GP-UCB formulation

Recall that for each arm i, we collect replicate-based labels $\tilde{s}_{t,i} = (y_{t,i} - y'_{t,i})^2/2$ in noise-update rounds ($\mathcal{I}(t) = 1$). We maintain the replication dataset $\mathcal{D}^{\sigma}_{t,i} = \{(\alpha_{\tau,i}, \tilde{s}_{\tau,i}) : \tau \leq t,\ \mathcal{I}(\tau) = 1\}$, and let $N^{\sigma}_{t,i} = |\mathcal{D}^{\sigma}_{t,i}|$ be the number of noise-learning samples. Conditioned on $\mathcal{D}^{\sigma}_{t-1,i}$, the GP posterior yields a mean $m^{\sigma}_{t-1,i}(\alpha)$ and variance $v^{\sigma}_{t-1,i}(\alpha)$ via standard closed-form conjugate updates:

$$m^{\sigma}_{t-1,i}(\alpha) = \mathbf{k}_i(\alpha)^\top (\mathbf{K}_i + \eta_n^2 \mathbf{I})^{-1} \tilde{\mathbf{s}}_i, \tag{11}$$

$$v^{\sigma}_{t-1,i}(\alpha) = k(\alpha, \alpha) - \mathbf{k}_i(\alpha)^\top (\mathbf{K}_i + \eta_n^2 \mathbf{I})^{-1} \mathbf{k}_i(\alpha), \tag{12}$$

where $\tilde{\mathbf{s}}_i$ is the label vector in $\mathcal{D}^{\sigma}_{t-1,i}$, $\mathbf{k}_i(\cdot)$ and $\mathbf{K}_i$ are the kernel vector and the kernel matrix evaluated at the queried fidelities under kernel k, and η_n^2 is the regularization parameter.

The upper confidence bound $\hat{\sigma}^2_{t-1,i}(\alpha)$ in Eq. (7) is then constructed using the schedule $\{\beta_n\}$ derived from the kernel's information gain [23], which we analyze further in Lemma G.1.

F.2 Suffix thresholds for a fixed order

Lemma F.1 (Suffix Thresholds for a Fixed Order)**.** *Fix an order π and a finite fidelity grid $\mathcal{A} \subseteq [0,1]$. For a fixed model, define the order-constrained suffix value functions by $W_{K+1}^{\pi,\mathcal{A}}(z) = z$ and for $m = K, \dots, 1$ (arm $i = \pi_m$),*

$$W_m^{\pi,\mathcal{A}}(z) = \max\Big\{z,\ \max_{\alpha\in\mathcal{A}}\big(\mathbb{E}[W_{m+1}^{\pi,\mathcal{A}}(\max\{z, X_i(\alpha)\})] - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha)\big)\Big\}, \tag{13}$$

and the corresponding suffix net gain

$$\widetilde{V}_m^{\pi,\mathcal{A}}(z) = \max_{\alpha\in\mathcal{A}}\big(\mathbb{E}[W_{m+1}^{\pi,\mathcal{A}}(\max\{z, X_i(\alpha)\})] - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha)\big) - z. \tag{14}$$

Assume (i) for each fixed $\alpha \in \mathcal{A}$, the single-arm gain $V_i(z,\alpha)$ is continuous and strictly decreasing in z (Lemma E.1), and (ii) $\min_{\alpha\in\mathcal{A}}\big(c_i^{\mathrm{g}} + c^{\mathrm{e}}(\alpha)\big) \geq c_{\min} > 0$. Then for each m, $\widetilde{V}_m^{\pi,\mathcal{A}}(z)$ is continuous and strictly decreasing in z and admits a unique root $\widetilde{\xi}_m^{\pi,\mathcal{A}}$ such that $\widetilde{V}_m^{\pi,\mathcal{A}}(\widetilde{\xi}_m^{\pi,\mathcal{A}}) = 0$. Moreover, the optimal order-constrained policy probes at stage m iff $z < \widetilde{\xi}_m^{\pi,\mathcal{A}}$, and any optimal control can be chosen from $\arg\max_{\alpha\in\mathcal{A}}\Big(\mathbb{E}[W_{m+1}^{\pi,\mathcal{A}}(\max\{z, X_{\pi_m}(\alpha)\})] - c_{\pi_m}^{\mathrm{g}} - c^{\mathrm{e}}(\alpha)\Big)$.

Proof. Define $g_m^{\pi,\mathcal{A}}(z) = W_m^{\pi,\mathcal{A}}(z) - z \geq 0$. We claim by backward induction that each $g_m^{\pi,\mathcal{A}}$ is continuous and non-increasing in z, and $\lim_{z\to+\infty} g_m^{\pi,\mathcal{A}}(z) = 0$. The base case holds since $W_{K+1}^{\pi,\mathcal{A}}(z) = z$. Assume the claim holds for $m+1$ and fix stage m (arm $i = \pi_m$). For $\alpha \in \mathcal{A}$, let $X(\alpha) = X_i(\alpha)$ and write

$$H_\alpha(z) = \mathbb{E}\Big[W_{m+1}^{\pi,\mathcal{A}}(\max\{z, X(\alpha)\})\Big] - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha) - z = V_i(z,\alpha) + \mathbb{E}\Big[g_{m+1}^{\pi,\mathcal{A}}(\max\{z, X(\alpha)\})\Big]. \tag{15}$$

By assumption, $V_i(z,\alpha)$ is continuous and strictly decreasing in z. Since $g_{m+1}^{\pi,\mathcal{A}}$ is continuous and non-increasing and $z \mapsto \max\{z, X(\alpha)\}$ is non-decreasing and continuous, the map $z \mapsto \mathbb{E}[g_{m+1}^{\pi,\mathcal{A}}(\max\{z, X(\alpha)\})]$ is continuous and non-increasing. Hence $H_\alpha(z)$ is continuous and strictly decreasing. Because $\mathcal{A}$ is finite, $\widetilde{V}_m^{\pi,\mathcal{A}}(z) = \max_{\alpha\in\mathcal{A}} H_\alpha(z)$ is also continuous and strictly decreasing. As $z \to -\infty$, for any fixed α, $\mathbb{E}[\max\{z, X(\alpha)\}] - z \to +\infty$ and $g_{m+1}^{\pi,\mathcal{A}} \geq 0$, so $H_\alpha(z) \to +\infty$ and thus $\widetilde{V}_m^{\pi,\mathcal{A}}(z) \to +\infty$. As $z \to +\infty$, we have $V_i(z,\alpha) \to -(c_i^{\mathrm{g}} + c^{\mathrm{e}}(\alpha)) \leq -c_{\min}$ and $\mathbb{E}[g_{m+1}^{\pi,\mathcal{A}}(\max\{z, X(\alpha)\})] \leq g_{m+1}^{\pi,\mathcal{A}}(z) \to 0$, so $\limsup_{z\to+\infty} H_\alpha(z) \leq -c_{\min} < 0$ uniformly over $\alpha \in \mathcal{A}$, implying $\widetilde{V}_m^{\pi,\mathcal{A}}(z) \to -c_{\min} < 0$. Therefore, by continuity and strict monotonicity, there exists a unique $\widetilde{\xi}_m^{\pi,\mathcal{A}}$ with $\widetilde{V}_m^{\pi,\mathcal{A}}(\widetilde{\xi}_m^{\pi,\mathcal{A}}) = 0$.

Finally, from Eq. (13) and Eq. (14), $W_m^{\pi,\mathcal{A}}(z) = z + \big(\widetilde{V}_m^{\pi,\mathcal{A}}(z)\big)_+$, so probing is optimal iff $\widetilde{V}_m^{\pi,\mathcal{A}}(z) > 0$, equivalently $z < \widetilde{\xi}_m^{\pi,\mathcal{A}}$. □

G Supplementary for Lemma 6.1: concentration and clean event

For a fixed arm i, each subsection below establishes the corresponding event with failure probability $\rho' = \rho/(3K)$. Taking the intersection over the three event types and all $i \in [K]$ yields the global clean event $\mathcal{E}$ in Lemma 6.1, with $\mathbb{P}(\mathcal{E}) \geq 1 - \rho$ by a union bound.

G.1 Proof of Lemma 6.1(i): observation boundedness

Proof. Let $\{\mathcal{F}_t\}_{\geq 0}$ denote the natural filtration generated by the algorithm's history up to time t. Under the model assumptions, $\varepsilon_{t,i}$ is conditionally $\Sigma_{\max}$-sub-Gaussian given $\mathcal{F}_{t-1}$, *i.e.,* $\mathbb{E}[\exp(\lambda\varepsilon_{t,i}) \mid \mathcal{F}_{t-1}] \leq \exp(\lambda^2\Sigma_{\max}/2)$ for all $\lambda \in \mathbb{R}$ and $t \in [T]$, and the same holds for an independent replicate $\varepsilon'_{t,i}$. Fix any $u > 0$. For each $t \in [T]$, since $Y_{t,i}(\alpha)$ is marginally Gaussian with noise at most $V_{\max} = \kappa_{\max}^2 + \Sigma_{\max}$, we may write $y_{t,i} = \mu_i + \bar{\varepsilon}_{t,i}$ where $\bar{\varepsilon}_{t,i}$ is conditionally $V_{\max}$-sub-Gaussian; thus

$$\mathbb{P}(|y_{t,i}| \geq 1 + u \mid \mathcal{F}_{t-1}) \leq \mathbb{P}(|\varepsilon_{t,i}| \geq u \mid \mathcal{F}_{t-1}) \leq 2\exp\left(-\frac{u^2}{2V_{\max}}\right), \tag{16}$$

and similarly, $\mathbb{P}\big(|y'_{t,i}| \geq 1+u \mid \mathcal{F}_{t-1}\big) \leq 2\exp\Big(-\frac{u^2}{2V_{\max}}\Big)$. Taking expectation over $\mathcal{F}_{t-1}$ yields the same bounds unconditionally. Taking a union bound over $t \in [T]$ for both $\{y_{t,i}\}$ and $\{y'_{t,i}\}$ yields

$$\mathbb{P}\left(\max_{t\in[T]} \max\{|y_{t,i}|, |y'_{t,i}|\} \geq 1+u\right) \leq 4T\exp\left(-\frac{u^2}{2V_{\max}}\right). \tag{17}$$

Setting the RHS at ρ' gives $u = \sqrt{2V_{\max}\log(4T/\rho')}$, with probability at least $1-\rho'$, hence

$$\max_{t\in[T]} \max\{|y_{t,i}|, |y'_{t,i}|\} \leq 1 + \sqrt{2V_{\max}\log(4T/\rho')} = Y_{\max}(\rho'). \tag{18}$$

□

G.2 Proof of Lemma 6.1(ii): mean tracking

Proof. Let $\{\mathcal{F}_t\}_{\geq 0}$ denote the natural filtration generated by the algorithm's history up to time t. Under the model assumptions, the marginal observation noise is conditionally $V_{\max}$-sub-Gaussian given $\mathcal{F}_{t-1}$, where $V_{\max} = \kappa_{\max}^2 + \Sigma_{\max}$. Fix any $u > 0$. For each $t \in [T]$, by the conditional sub-Gaussian tail bound for the sample mean, we have

$$\mathbb{P}(|\hat{\mu}_{t,i} - \mu_i| \geq u) \leq 2\exp\left(-\frac{N_{t,i}^{\mu} u^2}{2V_{\max}}\right) \leq 2\exp\left(-\frac{\max\{1, N_{t,i}^{\mu}\}u^2}{2V_{\max}}\right). \tag{19}$$

Taking a union bound over $t \in [T]$ yields

$$\mathbb{P}(\exists t \in [T] : |\hat{\mu}_{t,i} - \mu_i| \geq u_t) \leq \sum_{t=1}^{T} 2\exp\left(-\frac{\max\{1, N_{t,i}^{\mu}\}u_t^2}{2V_{\max}}\right). \tag{20}$$

Choosing $u_t = \sqrt{\frac{2V_{\max}\log(2T/\rho')}{\max\{1, N_{t,i}^{\mu}\}}}$ makes each summand at most ρ'/T, hence the RHS is at most ρ'. Therefore, with probability at least $1-\rho'$, for all $t \in [T]$,

$$|\hat{\mu}_{t,i} - \mu_i| \leq \sqrt{\frac{2V_{\max}\log(2T/\rho')}{\max\{1, N_{t,i}^{\mu}\}}} = B_{N_{t,i}^{\mu}}. \tag{21}$$

□

G.3 Kernelized self-normalized concentration

Theorem G.1 (Kernelized self-normalized concentration for adaptive designs [23, 24])**.** *Let $\{\mathcal{F}_t\}_{t\geq 0}$ be the natural filtration of the algorithm's history. We consider an adaptive sequence $\{x_t\}_{t\geq 1}$ with x_t being $\mathcal{F}_{t-1}$-measurable, and observations $y_t = f(x_t) + \varepsilon_t$, where $\{\varepsilon_t\}$ is a conditionally R-sub-Gaussian martingale difference sequence given $\mathcal{F}_{t-1}$. Let k be a positive semidefinite kernel on $\mathcal{X}$ with $k(x,x) \leq 1$ and $f \in \mathcal{H}_k$ with $\|f\|_{\mathcal{H}_k} \leq B$. Fix a regularization parameter $\eta_n^2 > 0$.*

Consider an adaptive sequence $\{x_t\}_{t\geq 1}$ such that x_t is $\mathcal{F}_{t-1}$-measurable, and observations

$$y_t = f(x_t) + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\exp(\lambda\varepsilon_t) \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right) \ \ \forall \lambda \in \mathbb{R}. \tag{22}$$

For any $n \geq 1$, define the kernel matrix $\mathbf{K}_n = [k(x_s, x_{s'})]_{s,s'\leq n}$, the kernel vector $\mathbf{k}_n(x) = (k(x_1,x), \ldots, k(x_n,x))^\top$, and the kernel ridge / GP posterior quantities

$$m_n(x) = \mathbf{k}_n(x)^\top (\mathbf{K}_n + \eta_n^2 \mathbf{I})^{-1} \mathbf{y}_{1:n},$$
$$v_n(x) = k(x,x) - \mathbf{k}_n(x)^\top (\mathbf{K}_n + \eta_n^2 \mathbf{I})^{-1} \mathbf{k}_n(x),$$

where $\mathbf{y}_{1:n} = (y_1, \ldots, y_n)^\top$. Let the (maximum) information gain $\gamma_n = \frac{1}{2}\log\det\big(\mathbf{I} + \eta_n^{-2}\mathbf{K}_n\big)$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $n \geq 1$ and all $x \in \mathcal{X}$ simultaneously,

$$|f(x) - m_n(x)| \leq \Big(B + R\sqrt{2(\gamma_n + 1 + \log(1/\delta))}\Big)\sqrt{v_n(x)}. \tag{23}$$

Equivalently, letting $\beta_n(\delta) = \Big(B + R\sqrt{2(\gamma_n + 1 + \log(1/\delta))}\Big)^2$, the bound reads $|f(x) - m_n(x)| \leq \sqrt{\beta_n(\delta)}\sqrt{v_n(x)}$ for all n, x on the same event.

G.4 Proof of Lemma 6.1(iii): noise learning

Proof. Fix arm i and let $f_i(\alpha) = \sigma_i^2(\alpha)$. On each noise-update round t, conditional on the latent quality $Q_{t,i}$, we observe a primary sample and an independent replicate $y_{t,i} = Q_{t,i} + \varepsilon_{t,i}$ and $y'_{t,i} = Q_{t,i} + \varepsilon'_{t,i}$, where conditional on $(\mathcal{F}_{t-1}, Q_{t,i}, \alpha_t)$, $\varepsilon_{t,i}$ and $\varepsilon'_{t,i}$ are independent, mean-zero, and satisfy $\mathbb{E}[(\varepsilon_{t,i})^2 \mid \mathcal{F}_{t-1}, Q_{t,i}, \alpha_t] = \mathbb{E}[(\varepsilon'_{t,i})^2 \mid \mathcal{F}_{t-1}, Q_{t,i}, \alpha_t] = f_i(\alpha_t)$. We form the unbiased label

$$\tilde{s}_{t,i} = \frac{(y_{t,i} - y'_{t,i})^2}{2} = \frac{(\varepsilon_{t,i} - \varepsilon'_{t,i})^2}{2}. \tag{24}$$

Taking conditional expectation given $(\mathcal{F}_{t-1}, Q_{t,i}, \alpha_t)$ and using independence gives

$$\mathbb{E}[\tilde{s}_{t,i} \mid \mathcal{F}_{t-1}, Q_{t,i}, \alpha_t] = f_i(\alpha_t), \tag{25}$$

and hence also $\mathbb{E}[\tilde{s}_{t,i} \mid \mathcal{F}_{t-1}, \alpha_t] = f_i(\alpha_t)$. Define the centered noise term $\tilde{\zeta}_t = \tilde{s}_{t,i} - f_i(\alpha_t)$, so that $\mathbb{E}[\tilde{\zeta}_t \mid \mathcal{F}_{t-1}, \alpha_t] = 0$. By observation boundedness, on the event $\max_{t \le T} \max\{|y_{t,i}|, |y'_{t,i}|\} \le Y_{\max}(\rho')$, we have $|y_{t,i} - y'_{t,i}| \le 2Y_{\max}(\rho')$, and hence

$$0 \le \tilde{s}_{t,i} \le 2Y_{\max}(\rho')^2 \quad \Rightarrow \quad |\tilde{\zeta}_t| \le 2Y_{\max}(\rho')^2 + \Sigma_{\max}. \tag{26}$$

Let $\widetilde{M}(\rho') = 2Y_{\max}(\rho')^2 + \Sigma_{\max}$. Therefore, $\{\tilde{\zeta}_t\}$ is a bounded martingale difference sequence, making it conditionally sub-Gaussian with a parameter $R = \widetilde{M}(\rho')$ depending only on $(Y_{\max}(\rho'), \Sigma_{\max})$.

Now consider GP regression based on the dataset $\mathcal{D}^{\sigma}_{t-1,i} = \{(\alpha_t, \tilde{s}_{t,i})\}_{t=1}^{N^{\sigma}_{t-1,i}}$. We invoke Theorem G.1 with $\mathcal{X} = [0,1]$, $x_t = \alpha_t$, $f(\alpha) = \sigma_i^2(\alpha)$, and observations $\tilde{s}_{t,i} = f(x_t) + \tilde{\zeta}_t$. On the bounded-observation event, we have $|\tilde{\zeta}_t| \le \widetilde{M}(\rho')$, hence the condition $\mathbb{E}[\exp(\lambda \varepsilon_t) \mid \mathcal{F}_{t-1}] \le \exp(\lambda^2 R^2/2)$ in Theorem G.1 holds with $R = \widetilde{M}(\rho')$.

To obtain a confidence sequence over n, set $\delta_n = \frac{6\rho'}{\pi^2 n^2}$ which makes $\sum_{n=1}^{\infty} \delta_n = \rho'$. Then let $\beta_n = \beta_n(\delta_n) = (B + R\sqrt{2(\gamma_n + 1 + \log(1/\delta_n))})^2$. Applying Theorem G.1, $\mathbb{P}(\{\forall x \in \mathcal{X}, |f(x) - m_n(x)| \le \sqrt{\beta_n(\delta)}\sqrt{v_n(x)}\}) \ge 1 - \delta_n$, and using the union bound over n, we have

$$\mathbb{P}(\{\forall n \ge 1, \forall x \in \mathcal{X}, |f(x) - m_n(x)| \le \sqrt{\beta_n(\delta)}\sqrt{v_n(x)}\}) \ge 1 - \sum_{n \ge 1} \delta_n = 1 - \rho'. \tag{27}$$

Consequently, with regard to our sparse noise update, we can set $n = N^{\sigma}_{t-1,i}$ as the number of samples on $\mathcal{D}^{\sigma}_{t-1,i}$. Then with probability at least $1 - \rho'$, for all $\alpha \in [0,1]$,

$$\left| f_i(\alpha) - m^{\sigma}_{t-1,i}(\alpha) \right| \le \sqrt{\beta_{N^{\sigma}_{t-1,i}}} \sqrt{v^{\sigma}_{t-1,i}(\alpha)}. \tag{28}$$

By the definition of the optimistic noise estimate $\hat{\sigma}^2_{t-1,i}(\alpha) = m^{\sigma}_{t-1,i}(\alpha) + \sqrt{\beta_{N^{\sigma}_{t-1,i}}} \sqrt{v^{\sigma}_{t-1,i}(\alpha)}$, we obtain

$$\sup_{\alpha \in [0,1]} \left| \hat{\sigma}^2_{t-1,i}(\alpha) - f_i(\alpha) \right| \le 2\sqrt{\beta_{N^{\sigma}_{t-1,i}}} \sup_{\alpha \in [0,1]} \sqrt{v^{\sigma}_{t-1,i}(\alpha)} = B_{N^{\sigma}_{t-1,i}}, \tag{29}$$

which proves the claimed bound. □

H Proof of Lemma 6.2: planning stability

Lemma H.1 (Distribution Shift Bound)**.** *Fix round t, arm j, and control α_j. Let $Y_j \sim \mathcal{N}(\mu_j, \kappa_j^2 + \sigma_j^2(\alpha_j))$ and $\hat{Y}_j \sim \mathcal{N}(\hat{\mu}_{t-1,j}, \kappa_j^2 + \hat{\sigma}^2_{t-1,j}(\alpha_j))$ be the observation distributions under true and estimated models, respectively. Let $f : \mathbb{R} \to \mathbb{R}$ be a function that is 1-Lipschitz in its argument. Under the clean event, the distribution shift error is bounded:*

$$\left| \mathbb{E}_{Y_j}[f(X_j(\alpha_j))] - \mathbb{E}_{\hat{Y}_j}[f(\hat{X}_j(\alpha_j))] \right| \le C^{\mu}_{shift} B_{N^{\mu}_{t-1,j}} + C^{\sigma}_{shift} B_{N^{\sigma}_{t-1,j}}, \tag{30}$$

where $\hat{X}_j$ is the posterior score under the estimated model, and $C^{\mu}_{shift}, C^{\sigma}_{shift}$ are constants dependent on $\lambda, \kappa_j^2, \sigma^2_{\min}$.

Proof. Since f is 1-Lipschitz, the mapping $y \mapsto f(S_j(y, \alpha))$ is Lipschitz continuous with a constant bounded by $\lambda w_j \leq \lambda$. The result follows by bounding the Wasserstein distance between the true observation distribution $\mathcal{N}(\mu_j, \kappa_j^2 + \sigma_j^2(\alpha_j))$ and the estimated one $\mathcal{N}(\hat{\mu}_{t-1,j}, \kappa_j^2 + \hat{\sigma}_{t-1,j}^2(\alpha_j))$, combined with the Lipschitz property of the posterior score mapping under the clean event. □

H.1 Exchange difference

Lemma H.2 (Exchange Difference)**.** *Fix a baseline $z \in \mathbb{R}$ (in the same λ-weighted scale as the state variable used in the recursion $z' = \max\{z, X_k\}$). For each arm $k \in \{i, j\}$, at baseline z let $\alpha_k^*(z) \in \arg\max_{\alpha \in [0,1]} V_k(z, \alpha)$, and define $X_k = S_k(Y_k, \alpha_k^*(z))$ with $Y_k \sim \mathcal{N}\big(\mu_k, \kappa_k^2 + \sigma_k^2(\alpha_k^*(z))\big)$. Let F_k denote the CDF of X_k (under $\alpha_k^*(z)$), and write $c_k = c_k^{\mathrm{g}} + c^{\mathrm{e}}(\alpha_k^*(z))$. Let $V_k^*(z)$ denote the optimized single-arm net gain at baseline z:*

$$V_k^*(z) = \mathbb{E}\big[\max\{X_k, z\}\big] - z - c_k. \tag{31}$$

Then the exchange difference between probing i then j versus j then i at baseline z satisfies

$$\Delta_{ij}(z) = V_i^*(z)\big(1 - F_j(z)\big) - V_j^*(z)\big(1 - F_i(z)\big). \tag{32}$$

Proof. For $k \in \{i, j\}$, write F_k for the CDF of X_k in control $\alpha_k^*(z)$, and write c_k for the corresponding cost. Consider probing arm i before j. If $X_i > z$ (probability $1 - F_i(z)$), the policy stops and obtains payoff X_i. If $X_i \leq z$ (probability $F_i(z)$), then the baseline remains z (since $\max\{z, X_i\} = z$) and the policy probes j, obtaining payoff $\max\{X_j, z\}$. Therefore,

$$\begin{aligned} W_{ij}(z) &= \mathbb{E}\big[X_i \mathbf{1}\{X_i > z\}\big] + F_i(z)\mathbb{E}\big[\max\{X_j, z\}\big] - c_i - c_j F_i(z) \\ &= \int_{x>z} x \mathrm{d}F_i(x) + F_i(z)\left(\int_{x>z} x \mathrm{d}F_j(x) + zF_j(z)\right) - c_i - c_j F_i(z). \end{aligned} \tag{33}$$

By symmetry,

$$\begin{aligned} W_{ji}(z) = &\int_{x>z} x \mathrm{d}F_j(x) + F_j(z)\left(\int_{x>z} x \mathrm{d}F_i(x) + zF_i(z)\right) \\ &- c_j - c_i F_j(z). \end{aligned} \tag{34}$$

Taking the difference and grouping terms yields

$$\begin{aligned} \Delta_{ij}(z) &= W_{ij}(z) - W_{ji}(z) \\ &= (1 - F_j(z))\Big(\int_{x>z} x \mathrm{d}F_i(x) - c_i\Big) - (1 - F_i(z))\Big(\int_{x>z} x \mathrm{d}F_j(x) - c_j\Big). \end{aligned} \tag{35}$$

Finally, relate the tail integral to $V_k(z)$. Since

$$\mathbb{E}\big[\max\{X_k, z\}\big] = \int_{x>z} x \mathrm{d}F_k(x) + zF_k(z), \tag{36}$$

we have

$$V_k(z) = \mathbb{E}\big[\max\{X_k, z\}\big] - z - c_k = \int_{x>z} x \mathrm{d}F_k(x) - z(1 - F_k(z)) - c_k, \tag{37}$$

equivalently,

$$\int_{x>z} x \mathrm{d}F_k(x) - c_k = V_k(z) + z(1 - F_k(z)). \tag{38}$$

Substituting this identity for $k = i, j$ into the expression for $\Delta_{ij}(z)$, leads to cancellation of $z(1 - F_i(z))(1 - F_j(z))$ terms, thus, we obtain the claimed formula. □

H.2 Proof of Lemma H.3: fidelity grid bounds

Lemma H.3 (Fidelity Grid Bounds)**.** *Fix arm i and baseline z. With $L_\alpha = L_c + \frac{\lambda\kappa_i^2}{2(\kappa_i^2+\sigma_{\min}^2)^{3/2}\sqrt{2\pi}} L_\sigma$, for any fidelity grid $\mathcal{A}$ (in Definition B.1) with covering radius η_α, $|V_i^*(z) - V_{i,\mathcal{A}}^*(z)| \le L_\alpha \eta_\alpha$.*

Proof. Under model assumptions, $\sigma_i^2(\cdot)$ is L_σ-Lipschitz on $[0,1]$ with $\sigma_i^2(\alpha) \ge \sigma_{\min}^2 > 0$, and $c^{\mathrm{e}}(\cdot)$ is L_c-Lipschitz. Fix z and let $Y(\alpha) \sim \mathcal{N}(\mu_i, \kappa_i^2 + \sigma_i^2(\alpha))$ and $X(\alpha) = S_i(Y(\alpha), \alpha)$. Since $S_i(y,\alpha) = \lambda(w_i(\alpha)y + (1-w_i(\alpha))\mu_i)$ with $w_i(\alpha) = \frac{\kappa_i^2}{\kappa_i^2+\sigma_i^2(\alpha)}$, we have $X(\alpha) \sim \mathcal{N}(\lambda\mu_i, s(\alpha)^2)$ with $s(\alpha) = \lambda\kappa_i^2(\kappa_i^2 + \sigma_i^2(\alpha))^{-1/2}$. Define $L_\alpha = L_c + \frac{\lambda\kappa_i^2}{2(\kappa_i^2+\sigma_{\min}^2)^{3/2}\sqrt{2\pi}} L_\sigma$. Recall $V_i(z,\alpha) = \mathbb{E}[\max(X(\alpha), z)] - z - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha)$. Hence,

$$|V_i(z,\alpha) - V_i(z,\alpha')| \le \big|\mathbb{E}[\max(X(\alpha), z)] - \mathbb{E}[\max(X(\alpha'), z)]\big| + |c^{\mathrm{e}}(\alpha) - c^{\mathrm{e}}(\alpha')|. \tag{39}$$

The cost term is bounded by $L_c|\alpha - \alpha'|$.

Using the closed form in Eq. (9), one can verify that $\frac{\partial}{\partial s}\mathbb{E}[\max(X,z)] = \phi(a)$ where $a = \frac{z-m}{s}$ and ϕ is the standard normal PDF. Therefore, $\big|\frac{\partial}{\partial s}\mathbb{E}[\max(X,z)]\big| \le \frac{1}{\sqrt{2\pi}}$. By the mean value theorem and the expression for $s(\alpha)$,

$$\big|\mathbb{E}[\max(X(\alpha), z)] - \mathbb{E}[\max(X(\alpha'), z)]\big| \le \frac{\lambda\kappa_i^2 L_\sigma}{2(\kappa_i^2 + \sigma_{\min}^2)^{3/2}\sqrt{2\pi}} |\alpha - \alpha'|. \tag{40}$$

Combining the cost and value bounds gives $|V_i(z,\alpha) - V_i(z,\alpha')| \le L_\alpha|\alpha - \alpha'|$, so $V_i(z,\cdot)$ is L_α-Lipschitz on $[0,1]$. Since $\mathcal{A} \subset [0,1]$, we have $V_{i,\mathcal{A}}^*(z) \le V_i^*(z)$; it remains to show $V_{i,\mathcal{A}}^*(z) \ge V_i^*(z) - L_\alpha\eta_\alpha$. Let $\alpha^* \in \arg\max_{\alpha\in[0,1]} V_i(z,\alpha)$ and choose $g \in \mathcal{A}$ with $|\alpha^* - g| \le \eta_\alpha$. Hence, $V_i(z,g) \ge V_i(z,\alpha^*) - L_\alpha\eta_\alpha = V_i^*(z) - L_\alpha\eta_\alpha$. Taking $\max_{\alpha\in\mathcal{A}}$ on the left yields $V_{i,\mathcal{A}}^*(z) \ge V_i^*(z) - L_\alpha\eta_\alpha$, so $|V_i^*(z) - V_{i,\mathcal{A}}^*(z)| \le L_\alpha\eta_\alpha$. □

H.3 Lemma: baseline grid bound

Lemma H.4 (Baseline Grid Bound)**.** *Fix arm pair $\{i,j\}$. For any baseline grid $\mathcal{Z} \subset Z$ (in Definition B.2) with covering radius η_z, $\inf_{z\in Z}\Delta_{ij}(z) \ge \min_{g\in\mathcal{Z}}\Delta_{ij}(g) - 2\eta_z$; the same holds for $\hat{\Delta}_{ij}$.*

Proof. Fix arm pair $\{i,j\}$. We prove $\Delta_{ij}(z)$ is 2-Lipschitz in z on Z, and the stated grid bound for Δ_{ij} and $\hat{\Delta}_{ij}$.

Recall $\Delta_{ij}(z) = R_{ij}(z) - R_{ji}(z)$, and that $R_{ij}(z)$ is the expected net utility of probing i then j from the baseline z, which we write as $W(z, \{i,j\}, (i,j))$ in the W-notation used in Lemma H.1. For any $a \in \mathbb{R}$, $z \mapsto \max\{z,a\}$ is 1-Lipschitz thus $z \mapsto \mathbb{E}[\max\{z,X\}]$ is 1-Lipschitz for any random variable X.

The two-stage recursion is $W(z,\{i,j\},(i,j)) = \max_{\alpha_i}\big(\mathbb{E}[W_j(z_i)] - c_i^{\mathrm{g}} - c^{\mathrm{e}}(\alpha_i)\big)$ with $z_i = \max\{z, X_i(\alpha_i)\}$, where $W_j(z)$ denotes the optimal value when only arm j remains to be probed from baseline z. For each fixed α_i, $z \mapsto \mathbb{E}[W_j(\max\{z, X_i(\alpha_i)\})]$ is 1-Lipschitz in z. Taking $\max$ over α_i preserves the 1-Lipschitz constant, so $W(\cdot,\{i,j\},(i,j))$ is 1-Lipschitz; the same holds for $W(\cdot,\{i,j\},(j,i))$. Hence,

$$\begin{aligned}|\Delta_{ij}(z) - \Delta_{ij}(z')| &\le |W(z,\{i,j\},(i,j)) - W(z',\{i,j\},(i,j))| \\ &\quad + |W(z,\{i,j\},(j,i)) - W(z',\{i,j\},(j,i))| \\ &\le 2|z - z'|. \end{aligned}\tag{41}$$

For each $z \in Z$ there exists $g \in \mathcal{Z}$ with $|z-g| \le \eta_z$. Then

$$\Delta_{ij}(z) \ge \Delta_{ij}(g) - 2|z-g| \ge \Delta_{ij}(g) - 2\eta_z \ge \min_{g'\in\mathcal{Z}}\Delta_{ij}(g') - 2\eta_z. \tag{42}$$

Taking $\inf_{z\in Z}$ on the left gives $\inf_{z\in Z}\Delta_{ij}(z) \ge \min_{g\in\mathcal{Z}}\Delta_{ij}(g) - 2\eta_z$. The same reasoning applies to $\hat{\Delta}_{ij}$ under the estimated model. □

H.4 Proof of Lemma 6.2(i): index stability

Proof. Fix arm i and the fidelity grid $\mathcal{A}$. Let $V_i^*(z;\mu,\sigma^2;\kappa_i^2) = \max_{\alpha\in[0,1]} V_i(z,\alpha;\mu,\sigma^2;\kappa_i^2)$, $V_{i,\mathcal{A}}^*(z;\mu,\sigma^2;\kappa_i^2) = \max_{\alpha\in\mathcal{A}} V_i(z,\alpha;\mu,\sigma^2;\kappa_i^2)$. Then ξ_i is the unique root of $V_i^*(z;\mu,\sigma^2;\kappa_i^2)=0$ and $\xi_{i,\mathcal{A}}$ is the unique root of $V_{i,\mathcal{A}}^*(z;\mu,\sigma^2;\kappa_i^2)=0$. Similarly, $\hat{\xi}_{t,i}$ is the algorithmic index in the estimated model, *i.e.,* the unique root of $V_{i,\mathcal{A}}^*(z;\hat{\mu}_{t-1,i},\hat{\sigma}^2_{t-1,i};\kappa_i^2)=0$.

Grid approximation. By Lemma H.3 for all $z\in Z$, we have

$$|V_i^*(z;\mu,\sigma^2;\kappa_i^2) - V_{i,\mathcal{A}}^*(z;\mu,\sigma^2;\kappa_i^2)| \le L_\alpha\eta_\alpha. \tag{43}$$

As $\mathcal{A}\subset[0,1]$, $V_{i,\mathcal{A}}^*(z;\mu,\sigma^2;\kappa_i^2) \le V_i^*(z;\mu,\sigma^2;\kappa_i^2)$, hence $0 \le V_i^*(z;\mu,\sigma^2;\kappa_i^2) - V_{i,\mathcal{A}}^*(z;\mu,\sigma^2;\kappa_i^2) \le L_\alpha\eta_\alpha$. Setting $z=\xi_i$ and using $V_i^*(\xi_i;\mu,\sigma^2;\kappa_i^2)=0$ gives

$$-L_\alpha\eta_\alpha \le V_{i,\mathcal{A}}^*(\xi_i;\mu,\sigma^2) \le 0. \tag{44}$$

Since $V_{i,\mathcal{A}}^*(\cdot;\mu,\sigma^2)$ is strictly decreasing and L_z-strongly monotone on Z by the standing assumption in Section 4, and $V_{i,\mathcal{A}}^*(\xi_{i,\mathcal{A}};\mu,\sigma^2)=0$, we have $\xi_{i,\mathcal{A}}\le\xi_i$ and

$$|\xi_i-\xi_{i,\mathcal{A}}| = \xi_i-\xi_{i,\mathcal{A}} \le \frac{1}{L_z}\Big(V_{i,\mathcal{A}}^*(\xi_{i,\mathcal{A}};\mu,\sigma^2) - V_{i,\mathcal{A}}^*(\xi_i;\mu,\sigma^2)\Big) \le \frac{L_\alpha}{L_z}\eta_\alpha. \tag{45}$$

Estimation on the grid. Fix any $z\in Z$ and let $\alpha^*(z)\in\arg\max_{\alpha\in\mathcal{A}} V_i(z,\alpha;\mu,\sigma^2)$, $\hat{\alpha}^*(z)\in\arg\max_{\alpha\in\mathcal{A}} V_i(z,\alpha;\hat{\mu},\hat{\sigma}^2)$. Define the two-point error

$$\omega(z) = |\mu-\hat{\mu}| + |\sigma^2(\alpha^*(z)) - \hat{\sigma}^2(\alpha^*(z))| + |\sigma^2(\hat{\alpha}^*(z)) - \hat{\sigma}^2(\hat{\alpha}^*(z))|. \tag{46}$$

For fixed (z,α) under the profile κ_i^2, let $Y\sim\mathcal{N}(\mu,\kappa_i^2+\sigma^2(\alpha))$. The map $V_i(z,\alpha;\mu,\sigma^2;\kappa_i^2)$ is Lipschitz in (μ,σ^2), with $|\partial V_i/\partial\mu|\le\lambda$ and $|\partial V_i/\partial\sigma^2|\le C_{V,i}^\sigma$ for a finite constant $C_{V,i}^\sigma$ depending only on $(\lambda,\kappa_i^2,\sigma_{\min})$. Hence for $L_V=\max\{\lambda,C_{V,i}^\sigma\}$,

$$|V_i(z,\alpha;\mu,\sigma^2) - V_i(z,\alpha;\hat{\mu},\hat{\sigma}^2)| \le L_V\Big(|\mu-\hat{\mu}| + |\sigma^2(\alpha)-\hat{\sigma}^2(\alpha)|\Big). \tag{47}$$

Evaluating the grid maximizers at $\alpha^*(z)$ and $\hat{\alpha}^*(z)$ yields

$$|V_{i,\mathcal{A}}^*(z;\mu,\sigma^2) - V_{i,\mathcal{A}}^*(z;\hat{\mu},\hat{\sigma}^2)| \le L_V\,\omega(z). \tag{48}$$

By the strong monotonicity of $V_{i,\mathcal{A}}^*(\cdot;\hat{\mu},\hat{\sigma}^2)$ on the two roots $V_{i,\mathcal{A}}^*(\xi_{i,\mathcal{A}};\mu,\sigma^2)=0$ and $V_{i,\mathcal{A}}^*(\hat{\xi}_{t,i};\hat{\mu},\hat{\sigma}^2)=0$, it has:

$$L_z|\xi_{i,\mathcal{A}}-\hat{\xi}_{t,i}| \le \big|V_{i,\mathcal{A}}^*(\xi_{i,\mathcal{A}};\hat{\mu},\hat{\sigma}^2) - V_{i,\mathcal{A}}^*(\xi_{i,\mathcal{A}};\mu,\sigma^2)\big| \le L_V\,\omega(\xi_{i,\mathcal{A}}). \tag{49}$$

On the clean event $\mathcal{E}$, we have $|\mu-\hat{\mu}|\le B_{N_{t-1,i}^\mu}$ and $\sup_{\alpha\in[0,1]}|\sigma^2(\alpha)-\hat{\sigma}^2(\alpha)|\le B_{N_{t-1,i}^\sigma}$, hence also for all $\alpha\in\mathcal{A}$. Therefore, $\omega(\xi_{i,\mathcal{A}})\le B_{N_{t-1,i}^\mu}+2B_{N_{t-1,i}^\sigma}$, and Eq. (49) gives

$$|\xi_{i,\mathcal{A}}-\hat{\xi}_{t,i}| \le \frac{L_V}{L_z}\Big(B_{N_{t-1,i}^\mu}+2B_{N_{t-1,i}^\sigma}\Big). \tag{50}$$

Finally, by the triangle inequality, combining Eq. (45) and Eq. (50) yields

$$|\xi_i-\hat{\xi}_{t,i}| \le |\xi_i-\xi_{i,\mathcal{A}}| + |\xi_{i,\mathcal{A}}-\hat{\xi}_{t,i}| \le \frac{L_V}{L_z}\Big(B_{N_{t-1,i}^\mu}+2B_{N_{t-1,i}^\sigma}\Big) + \frac{L_\alpha}{L_z}\eta_\alpha. \tag{51}$$

□

H.5 Proof of Lemma 6.2(ii): exchange stability

Proof. For a fixed order $(a,b)\in\{(i,j),(j,i)\}$, let $W_{a\to b}(z;\mathcal{M})$ denote the expected optimal value when probing arm a first and then arm b, with optimal control choices at each step in model $\mathcal{M}$.

With grid restriction $\alpha \in \mathcal{A}$, we denote $W^{\mathcal{A}}_{a\to b}(z;\mathcal{M})$ as the analogous two-arm value. Define the exchange difference as

$$\Delta_{ij}(z;\mathcal{M}) = W_{i\to j}(z;\mathcal{M}) - W_{j\to i}(z;\mathcal{M}), \Delta^{\mathcal{A}}_{ij}(z;\mathcal{M}) = W^{\mathcal{A}}_{i\to j}(z;\mathcal{M}) - W^{\mathcal{A}}_{j\to i}(z;\mathcal{M}). \quad (52)$$

Similarly, let $\hat{\Delta}_{ij}(z;\hat{\mathcal{M}}_t)$ be the exchange difference under the estimated model on $\mathcal{A}$.

Grid approximation. Fix an order $(a,b) \in \{(i,j),(j,i)\}$. Write the two-arm recursion as

$$W_{a\to b}(z;\mathcal{M}) = \max_{\alpha_a\in[0,1]} \Big(\mathbb{E}\big[W_b(z_a;\mathcal{M})\big] - c^{\mathrm{g}}_a - c^{\mathrm{e}}(\alpha_a)\Big), \qquad z_a = \max\{z, X_a(\alpha_a)\}, \quad (53)$$

where the second-stage value is $W_b(z;\mathcal{M}) = \max\Big\{z,\ z + \max_{\alpha_b\in[0,1]} V_b(z,\alpha_b)\Big\}$.

By Lemma H.3, for any baseline z and any arm $k \in \{a,b\}$, we have $V^*_k(z) - V^*_{k,\mathcal{A}}(z) \le L_\alpha \eta_\alpha$. Therefore,

$$\big|W_b(z;\mathcal{M}) - W^{\mathcal{A}}_b(z;\mathcal{M})\big| \le L_\alpha \eta_\alpha. \quad (54)$$

By the same grid-approximation control as used for single-arm maximizations,

$$\sup_{z\in Z} \big|W_{a\to b}(z;\mathcal{M}) - W^{\mathcal{A}}_{a\to b}(z;\mathcal{M})\big| \le 2L_\alpha \eta_\alpha. \quad (55)$$

Finally, apply Eq. (53) to both orders (i,j) and (j,i) and use triangle inequality:

$$\begin{aligned}\sup_{z\in Z}\Big|\Delta_{ij}(z;\mathcal{M}) - \Delta^{\mathcal{A}}_{ij}(z;\mathcal{M})\Big| &\le \sup_{z\in Z}\big|W_{i\to j}(z;\mathcal{M}) - W^{\mathcal{A}}_{i\to j}(z;\mathcal{M})\big| \\ &\quad + \sup_{z\in Z}\big|W_{j\to i}(z;\mathcal{M}) - W^{\mathcal{A}}_{j\to i}(z;\mathcal{M})\big| \\ &\le 2L_\alpha\eta_\alpha + 2L_\alpha\eta_\alpha = 4L_\alpha\eta_\alpha. \end{aligned} \quad (56)$$

Estimation on the grid. We now bound the stability of the grid-restricted two-arm values under model estimation errors.

(i) One-arm errors. For any arm $k \in \{a,b\}$, let $z_k = \max\{z, X_k(\alpha)\}$ with $Y_k \sim \mathcal{N}(\mu_k, \kappa_k^2 + \sigma_k^2(\alpha))$ and $X_k(\alpha) = S_k(Y_k,\alpha)$ and $\hat{z}_k = \max\{z, \hat{X}_k(\alpha)\}$ with $\hat{Y}_k \sim \mathcal{N}(\hat{\mu}_{t-1,k}, \kappa_k^2 + \hat{\sigma}^2_{t-1,k}(\alpha))$ and $\hat{X}_k(\alpha) = \hat{S}_k(\hat{Y}_k,\alpha)$. Define the one-arm value functions with grid approximation:

$$W^{\mathcal{A}}_k(z;\mathcal{M}) = \max_{\alpha\in\mathcal{A}} \mathbb{E}[z_k] - c^{\mathrm{g}}_k - c^{\mathrm{e}}(\alpha), \qquad W^{\mathcal{A}}_k(z;\hat{\mathcal{M}}_t) = \max_{\alpha\in\mathcal{A}} \mathbb{E}[\hat{z}_k] - c^{\mathrm{g}}_k - c^{\mathrm{e}}(\alpha). \quad (57)$$

For each fixed $\alpha \in \mathcal{A}$, the map $x \mapsto \max\{z,x\}$ is 1-Lipschitz, and hence Lemma H.1 applies uniformly over $\alpha \in \mathcal{A}$. Therefore,

$$\begin{aligned}\sup_{z\in Z}\Big|W^{\mathcal{A}}_k(z;\mathcal{M}) - W^{\mathcal{A}}_k(z;\hat{\mathcal{M}}_t)\Big| &\le \sup_{z\in Z}\max_{\alpha\in\mathcal{A}}\Big|\mathbb{E}[z_k] - \mathbb{E}[\hat{z}_k]\Big| \\ &\le C^{\mu}_{\mathrm{shift}} B_{N^{\mu}_{t-1,k}} + C^{\sigma}_{\mathrm{shift}} B_{N^{\sigma}_{t-1,k}}. \end{aligned} \quad (58)$$

(ii) Two-arm order value under a fixed first-stage control. Fix an order $(a,b) \in \{(i,j),(j,i)\}$. For fixed $\alpha_a \in \mathcal{A}$, denote

$$Q^{\mathcal{A}}_{a\to b}(z,\alpha_a;\mathcal{M}) = \mathbb{E}\Big[W^{\mathcal{A}}_b(z_a;\mathcal{M})\Big] - c^{\mathrm{g}}_a - c^{\mathrm{e}}(\alpha_a), \qquad z_a = \max\{z, X_a(\alpha_a)\}, \quad (59)$$

and $Q^{\mathcal{A}}_{a\to b}(z,\alpha_a;\hat{\mathcal{M}}_t)$ is defined similarly to $\hat{z}_a = \max\{z, \hat{X}_a(\alpha_a)\}$. Then, we have

$$\begin{aligned}&\Big|Q^{\mathcal{A}}_{a\to b}(z,\alpha_a;\mathcal{M}) - Q^{\mathcal{A}}_{a\to b}(z,\alpha_a;\hat{\mathcal{M}}_t)\Big| \\ &\quad\le \underbrace{\Big|\mathbb{E}\Big[W^{\mathcal{A}}_b(z_a;\mathcal{M}) - W^{\mathcal{A}}_b(z_a;\hat{\mathcal{M}}_t)\Big]\Big|}_{\text{(A) one-arm error for arm } b} + \underbrace{\Big|\mathbb{E}\Big[W^{\mathcal{A}}_b(z_a;\hat{\mathcal{M}}_t)\Big] - \mathbb{E}\Big[W^{\mathcal{A}}_b(\hat{z}_a;\hat{\mathcal{M}}_t)\Big]\Big|}_{\text{(B) baseline distribution shift for arm } a}. \end{aligned} \quad (60)$$

For (A) term, by Eq. (58) with $k=b$, (A) $\le C^{\mu}_{\mathrm{shift}} B_{N^{\mu}_{t-1,b}} + C^{\sigma}_{\mathrm{shift}} B_{N^{\sigma}_{t-1,b}}$; For (B) term, note that for any fixed model, the one-arm value $W^{\mathcal{A}}_b(\cdot;\hat{\mathcal{M}}_t)$ is 1-Lipschitz in its baseline. Thus, we can also apply

Lemma H.1 to the pair $(z_a, \hat{z}_a)$ generated by arm a, yielding $(\mathrm{B}) \le C^{\mu}_{\mathrm{shift}} B_{N^{\mu}_{t-1,a}} + C^{\sigma}_{\mathrm{shift}} B_{N^{\sigma}_{t-1,a}}$. Therefore,

$$\left| Q^{\mathcal{A}}_{a\to b}(z, \alpha_a; \mathcal{M}) - Q^{\mathcal{A}}_{a\to b}(z, \alpha_a; \hat{\mathcal{M}}_t) \right| \le 2C^{\mu}_{\mathrm{shift}} \max_{k\in\{a,b\}} B_{N^{\mu}_{t-1,k}} + 2C^{\sigma}_{\mathrm{shift}} \max_{k\in\{a,b\}} B_{N^{\sigma}_{t-1,k}}. \tag{61}$$

Taking $\sup_{\alpha_a \in \mathcal{A}}$, we have

$$\left| W^{\mathcal{A}}_{a\to b}(z; \mathcal{M}) - W^{\mathcal{A}}_{a\to b}(z; \hat{\mathcal{M}}_t) \right| \le 2C^{\mu}_{\mathrm{shift}} \max_{k\in\{a,b\}} B_{N^{\mu}_{t-1,k}} + 2C^{\sigma}_{\mathrm{shift}} \max_{k\in\{a,b\}} B_{N^{\sigma}_{t-1,k}}. \tag{62}$$

Applying Eq. (62) to both orders (i, j) and (j, i) yields, for all $z \in Z$,

$$\left| \Delta^{\mathcal{A}}_{ij}(z; \mathcal{M}) - \hat{\Delta}_{ij}(z; \hat{\mathcal{M}}_t) \right| \le 4C^{\mu}_{\mathrm{shift}} \max_{k\in\{i,j\}} B_{N^{\mu}_{t-1,k}} + 4C^{\sigma}_{\mathrm{shift}} \max_{k\in\{i,j\}} B_{N^{\sigma}_{t-1,k}}. \tag{63}$$

By triangle inequality, combining Eq. (56) and Eq. (63) yields

$$\begin{aligned} \left| \Delta_{ij}(z; \mathcal{M}) - \hat{\Delta}_{ij}(z; \hat{\mathcal{M}}_t) \right| &\le \left| \Delta_{ij}(z; \mathcal{M}) - \Delta^{\mathcal{A}}_{ij}(z; \mathcal{M}) \right| + \left| \Delta^{\mathcal{A}}_{ij}(z; \mathcal{M}) - \hat{\Delta}_{ij}(z; \hat{\mathcal{M}}_t) \right| \\ &\le 4C^{\mu}_{\mathrm{shift}} \max_{k\in\{i,j\}} B_{N^{\mu}_{t-1,k}} + 4C^{\sigma}_{\mathrm{shift}} \max_{k\in\{i,j\}} B_{N^{\sigma}_{t-1,k}} + 4L_{\alpha}\eta_{\alpha}, \end{aligned} \tag{64}$$

which completes the proof. □

I Supplementary for Section 6: regret bound

I.1 Exchanging adjacent inverted pairs

Lemma I.1 (Exchanging Adjacent Inverted Pairs)**.** *Fix a set of arms $\mathcal{U}$ with $n = |\mathcal{U}|$ and a strict partial order $\prec$ on $\mathcal{U}$, which is irreflexive and transitive. Let π^* be any linear extension of $\prec$ (*i.e., *a permutation consistent with $\prec$). Then for any permutation π of $\mathcal{U}$, there exists a sequence of permutations $\pi^{(0)} = \pi, \pi^{(1)}, \ldots, \pi^{(M)} = \pi^*$ such that each $\pi^{(m)}$ is obtained from $\pi^{(m-1)}$ by swapping an adjacent inverted pair with respect to $\prec$. Moreover, $M \le \binom{n}{2}$. Assume that each such adjacent swap changes the value by at most ϵ uniformly over $z \in Z$,* i.e., *$W(z, \mathcal{U}, \pi^{(m)}) - W(z, \mathcal{U}, \pi^{(m-1)}) \le \epsilon$ then we*

$$W(z, \mathcal{U}, \pi^*) - W(z, \mathcal{U}, \pi) \le M\epsilon \le \binom{n}{2}\epsilon. \tag{65}$$

Proof. For a permutation π, define its inversion set (with respect to $\prec$) as

$$\mathrm{Inv}(\pi) = \{(u, v) \in \mathcal{U}^2 : \ v \prec u, \ \pi^{-1}(u) < \pi^{-1}(v)\}. \tag{66}$$

Then π is a linear extension of $\prec$ iff $\mathrm{Inv}(\pi) = \emptyset$. If $\mathrm{Inv}(\pi) \neq \emptyset$, pick an inversion $(u, v) \in \mathrm{Inv}(\pi)$ minimizing the distance $\pi^{-1}(v) - \pi^{-1}(u)$. We claim u and v are adjacent in π. Otherwise, there exists an element w with $\pi^{-1}(u) < \pi^{-1}(w) < \pi^{-1}(v)$. By transitivity of $\prec$, at least one of $v \prec w$ or $w \prec u$ must hold; in either case, (u, w) or (w, v) is an inversion with strictly smaller distance, contradicting minimality. Hence, an adjacent inverted pair exists.

Let π' be obtained from π by swapping this adjacent inverted pair (u, v). Since u and v are adjacent, swapping only changes their relative order; for any $w \in \mathcal{U} \setminus \{u, v\}$, the relative order between w and each of u, v is unchanged. Therefore, $\mathrm{Inv}(\pi') = \mathrm{Inv}(\pi) \setminus \{(u, v)\}$ and in particular $|\mathrm{Inv}(\pi')| = |\mathrm{Inv}(\pi)| - 1$. By iteratively swapping adjacent inverted pairs, we obtain a sequence $\pi^{(0)}, \pi^{(1)}, \ldots$ that strictly decreases $|\mathrm{Inv}(\cdot)|$ at each step until reaching a permutation with no inversions, *i.e.,* a linear extension π^*. Since $|\mathrm{Inv}(\pi)| \le \binom{n}{2}$ for any π, the number of swaps satisfies $M \le \binom{n}{2}$. Finally, under the per-swap bound, we telescope:

$$W(z, \mathcal{U}, \pi^*) - W(z, \mathcal{U}, \pi) = \sum_{m=1}^{M} \left(W(z, \mathcal{U}, \pi^{(m)}) - W(z, \mathcal{U}, \pi^{(m-1)}) \right) \le M\epsilon \le \binom{n}{2}\epsilon, \tag{67}$$

where the last inequality follows from the value loss bound ϵ. The resulting bound is uniform over $z \in Z$. □

I.2 Proof of Lemma 6.3(i): model misspecification

Proof. We prove the following stronger induction statement. For each $n \geq 0$, there exist constants $L^{\pi}_{n,\mu}$ and $L^{\pi}_{n,\sigma}$ such that for any arm set $\mathcal{U}$ with $|\mathcal{U}| = n$ and any baseline z,

$$\left|W(z,\mathcal{U},\mathcal{M};\pi) - W(z,\mathcal{U},\hat{\mathcal{M}}_t;\pi)\right| \leq L^{\pi}_{n,\mu} \sum_{j\in\mathcal{U}} B_{N^{\mu}_{t-1,j}} + L^{\pi}_{n,\sigma} \sum_{j\in\mathcal{U}} B_{N^{\sigma}_{t-1,j}}. \tag{68}$$

Base case ($n = 0$): $W(z,\emptyset,\cdot;\pi) = z$, so $L^{\pi}_{0,\mu} = 0$ and $L^{\pi}_{0,\sigma} = 0$ suffice.

Inductive step: Assume that the induction hypothesis holds for all sets of size $n-1$ with constants $L^{\pi}_{n-1,\mu}$ and $L^{\pi}_{n-1,\sigma}$. At state $(z,\mathcal{U})$ with $|\mathcal{U}| = n$, policy π prescribes a specific action.

Case 1: Policy π stops. Then $W(z,\mathcal{U},\mathcal{M};\pi) = W(z,\mathcal{U},\hat{\mathcal{M}}_t;\pi) = z$. The difference is 0, which requires $L^{\pi}_{n,\mu} \geq 0$ and $L^{\pi}_{n,\sigma} \geq 0$.

Case 2: Policy π probes arm j with control α^{π}_j. We decompose the value difference:

$$\begin{aligned}|W(z,\mathcal{U},\mathcal{M};\pi) - W(z,\mathcal{U},\hat{\mathcal{M}}_t;\pi)| \leq \underbrace{\left|\mathbb{E}_{Y_j}[W(z'_j,\mathcal{U}',\mathcal{M};\pi)] - \mathbb{E}_{\hat{Y}_j}[W(z'_j,\mathcal{U}',\mathcal{M};\pi)]\right|}_{\text{(A) distribution shift}} \\ + \underbrace{\mathbb{E}_{\hat{Y}_j}\left[\left|W(z'_j,\mathcal{U}',\mathcal{M};\pi) - W(z'_j,\mathcal{U}',\hat{\mathcal{M}}_t;\pi)\right|\right]}_{\text{(B) continuation error}},\end{aligned} \tag{69}$$

where $z'_j = \max(z, X_j(\alpha))$ (or $\hat{X}_j(\alpha)$ respectively), and $\mathcal{U}' = \mathcal{U} \setminus \{j\}$.

For (A) term, by Lemma H.1, since $W(\cdot,\mathcal{U}',\mathcal{M};\pi)$ is 1-Lipschitz in its first argument, we have:

$$\text{(A)} \leq C^{\mu}_{\text{shift}} B_{N^{\mu}_{t-1,j}} + C^{\sigma}_{\text{shift}} B_{N^{\sigma}_{t-1,j}}. \tag{70}$$

For (B) term, by the induction hypothesis applied to the continuation value function:

$$\text{(B)} \leq L^{\pi}_{n-1,\mu} \sum_{j'\in\mathcal{U}'} B_{N^{\mu}_{t-1,j'}} + L^{\pi}_{n-1,\sigma} \sum_{j'\in\mathcal{U}'} B_{N^{\sigma}_{t-1,j'}}. \tag{71}$$

Combining Eq. (70) and Eq. (71), it has:

$$\begin{aligned}&|W(z,\mathcal{U},\mathcal{M};\pi) - W(z,\mathcal{U},\hat{\mathcal{M}}_t;\pi)| \\ &\leq C^{\mu}_{\text{shift}} B_{N^{\mu}_{t-1,j}} + C^{\sigma}_{\text{shift}} B_{N^{\sigma}_{t-1,j}} + L^{\pi}_{n-1,\mu} \sum_{j'\in\mathcal{U}'} B_{N^{\mu}_{t-1,j'}} + L^{\pi}_{n-1,\sigma} \sum_{j'\in\mathcal{U}'} B_{N^{\sigma}_{t-1,j'}} \\ &\leq \left(C^{\mu}_{\text{shift}} + L^{\pi}_{n-1,\mu}\right) \sum_{j\in\mathcal{U}} B_{N^{\mu}_{t-1,j}} + \left(C^{\sigma}_{\text{shift}} + L^{\pi}_{n-1,\sigma}\right) \sum_{j\in\mathcal{U}} B_{N^{\sigma}_{t-1,j}}.\end{aligned} \tag{72}$$

This requires $L^{\pi}_{n,\mu} \geq C^{\mu}_{\text{shift}} + L^{\pi}_{n-1,\mu}$ and $L^{\pi}_{n,\sigma} \geq C^{\sigma}_{\text{shift}} + L^{\pi}_{n-1,\sigma}$.

Induction conclusion. Since $L^{\pi}_{0,\mu} = 0$ and $L^{\pi}_{0,\sigma} = 0$, by induction:

$$L^{\pi}_{n,\mu} = n \cdot C^{\mu}_{\text{shift}}, L^{\pi}_{n,\sigma} = n \cdot C^{\sigma}_{\text{shift}}. \tag{73}$$

For $n = K$ and the initial state ($z_0 = -\infty, \mathcal{U} = [K]$), we obtain $L^{\pi}_{K,\mu} = KC^{\mu}_{\text{shift}}$ and $L^{\pi}_{K,\sigma} = KC^{\sigma}_{\text{shift}}$. Substituting $\Psi^{\mu}_t = \sum_{j\in[K]} B_{N^{\mu}_{t-1,j}}$, $\Psi^{\sigma}_t = \sum_{j\in[K]} B_{N^{\sigma}_{t-1,j}}$, it gives the claimed bound $\mathcal{O}\big(K(\Psi^{\mu}_t + \Psi^{\sigma}_t)\big)$. □

I.3 Proof of Lemma 6.3(ii): optimal value shift

Proof. We prove by induction on $n = |\mathcal{U}|$. For each $n \geq 0$, there exist constants $L_{n,\mu}$ and $L_{n,\sigma}$ such that for any arm set $\mathcal{U}$ with $|\mathcal{U}| = n$ and any baseline z,

$$\left|W^*(z,\mathcal{U};\mathcal{M}) - W^*(z,\mathcal{U};\hat{\mathcal{M}}_t)\right| \leq L_{n,\mu} \sum_{j\in\mathcal{U}} B_{N^{\mu}_{t-1,j}} + L_{n,\sigma} \sum_{j\in\mathcal{U}} \sup_{\alpha\in[0,1]} B^{\sigma}_{t-1,j}(\alpha). \tag{74}$$

Base case ($n = 0$): $W^*(z, \emptyset, \cdot) = z$, so $L_{0,\mu} = 0$ and $L_{0,\sigma} = 0$ suffice.

Inductive step: Assume the bound holds for all sets of size $n-1$ with constants $L_{n-1,\mu}$ and $L_{n-1,\sigma}$. Let $j^* \in \arg\max_{j\in\mathcal{U}} Q(j; z, \mathcal{U}, \mathcal{M})$ and $\hat{j}^* \in \arg\max_{j\in\mathcal{U}} Q(j; z, \mathcal{U}, \hat{\mathcal{M}}_t)$. We analyze four exhaustive cases:

Case 1: Both policies stop. Then $W^*(z, \mathcal{U}, \mathcal{M}_t) = W^*(z, \mathcal{U}, \hat{\mathcal{M}}_t) = z$. The difference is 0, requiring $L_{n,\mu} \geq 0$ and $L_{n,\sigma} \geq 0$.

Case 2: Both policies continue with same arm ($j^* = \hat{j}^*$). Let $\alpha_j^* = \arg\max_\alpha Q(j^*; \mathcal{M}_t)$ and $\hat{\alpha}_j^* = \arg\max_\alpha Q(j^*; \hat{\mathcal{M}}_t)$. By optimality:

$$Q(j^*; \mathcal{M}_t) = Q(j^*, \alpha_j^*; \mathcal{M}_t) \geq Q(j^*, \hat{\alpha}_j^*; \mathcal{M}_t), Q(j^*; \hat{\mathcal{M}}_t) = Q(j^*, \hat{\alpha}_j^*; \hat{\mathcal{M}}_t) \geq Q(j^*, \alpha_j^*; \hat{\mathcal{M}}_t). \tag{75}$$

Therefore:

$$\begin{aligned} &|Q(j^*; \mathcal{M}_t) - Q(j^*; \hat{\mathcal{M}}_t)| \\ &\leq \max\Big(|Q(j^*, \alpha_j^*; \mathcal{M}_t) - Q(j^*, \alpha_j^*; \hat{\mathcal{M}}_t)|, |Q(j^*, \hat{\alpha}_j^*; \mathcal{M}_t) - Q(j^*, \hat{\alpha}_j^*; \hat{\mathcal{M}}_t)|\Big) \end{aligned} \tag{76}$$

For each control choice $\alpha \in \{\alpha_j^*, \hat{\alpha}_j^*\}$, we apply a triangle decomposition:

$$\begin{aligned} \big|Q(j^*, \alpha; \mathcal{M}_t) - Q(j^*, \alpha; \hat{\mathcal{M}}_t)\big| \leq &\underbrace{\Big|\mathbb{E}_{Y_{j^*}}\big[W(z'_{j^*}, \mathcal{U}', \mathcal{M}_t)\big] - \mathbb{E}_{\hat{Y}_{j^*}}\big[W(z'_{j^*}, \mathcal{U}', \mathcal{M}_t)\big]\Big|}_{\text{(A)distribution shift}} \\ &+ \underbrace{\mathbb{E}_{\hat{Y}_{j^*}}\Big[\big|W(z'_{j^*}, \mathcal{U}', \mathcal{M}_t) - W(z'_{j^*}, \mathcal{U}', \hat{\mathcal{M}}_t)\big|\Big]}_{\text{(B)continuation error}}, \end{aligned} \tag{77}$$

where $\mathcal{U}' = \mathcal{U} \setminus \{j^*\}$ and $z'_{j^*} = \max\{z, X_{j^*}(\alpha)\}$ under $\mathcal{M}_t$ (and $\hat{z}'_{j^*} = \max\{z, \hat{X}_{j^*}(\alpha)\}$ under $\hat{\mathcal{M}}_t$). By Lemma H.1, (A) term is bounded by:

$$\text{(A)} \leq C^\mu_{\text{shift}} B_{N^\mu_{t-1,j^*}} + C^\sigma_{\text{shift}} B_{N^\sigma_{t-1,j^*}}. \tag{78}$$

By the induction hypothesis, (B) term is bounded by:

$$\text{(B)} \leq L_{n-1,\mu} \sum_{j'\in\mathcal{U}\setminus\{j^*\}} B_{N^\mu_{t-1,j'}} + L_{n-1,\sigma} \sum_{j'\in\mathcal{U}\setminus\{j^*\}} B_{N^\sigma_{t-1,j'}}. \tag{79}$$

Taking the maximum over both control choices, which introduces a factor of 2 in the worst case, we have:

$$\begin{aligned} &|W^*(z, \mathcal{U}, \mathcal{M}_t) - W^*(z, \mathcal{U}, \hat{\mathcal{M}}_t)| \\ &\leq 2\lambda B_{N^\mu_{t-1,j^*}} + \frac{2\lambda}{2\sigma_{\min}} B_{N^\sigma_{t-1,j^*}} + L_{n-1,\mu} \sum_{j'\in\mathcal{U}'} B_{N^\mu_{t-1,j'}} + L_{n-1,\sigma} \sum_{j'\in\mathcal{U}'} B_{N^\sigma_{t-1,j'}} \\ &\leq (2\lambda + L_{n-1,\mu}) \sum_{j\in\mathcal{U}} B_{N^\mu_{t-1,j}} + \left(\frac{\lambda}{\sigma_{\min}} + L_{n-1,\sigma}\right) \sum_{j\in\mathcal{U}} B_{N^\sigma_{t-1,j}}. \end{aligned} \tag{80}$$

This requires:

$$L_{n,\mu} \geq 2\lambda + L_{n-1,\mu}, \qquad L_{n,\sigma} \geq \frac{\lambda}{\sigma_{\min}} + L_{n-1,\sigma}. \tag{81}$$

Case 3: Both continue with different arms ($j^* \neq \hat{j}^*$). By triangle inequality:

$$\begin{aligned} &|Q(j^*; z, \mathcal{U}, \mathcal{M}) - Q(\hat{j}^*; z, \mathcal{U}, \hat{\mathcal{M}}_t)| \\ &\leq |Q(j^*; z, \mathcal{U}, \mathcal{M}) - Q(j^*; z, \mathcal{U}, \hat{\mathcal{M}}_t)| + |Q(\hat{j}^*; z, \mathcal{U}, \mathcal{M}) - Q(\hat{j}^*; z, \mathcal{U}, \hat{\mathcal{M}}_t)|. \end{aligned} \tag{82}$$

Each term is bounded similarly to Case 2, requiring:

$$L_{n,\mu} \geq 2\lambda + L_{n-1,\mu}, L_{n,\sigma} \geq \frac{\lambda}{\sigma_{\min}} + L_{n-1,\sigma}. \tag{83}$$

Case 4: Stopping decisions diverge. Without loss of generality, suppose $\mathcal{M}$ stops (so $z \geq Q(j; z, \mathcal{U}, \mathcal{M})$ for all j) but $\hat{\mathcal{M}}_t$ continues with arm $\hat{j}^*$ (so $Q(\hat{j}^*; z, \mathcal{U}, \hat{\mathcal{M}}_t) > z$). Then:

$$\begin{aligned} |W^*(z, \mathcal{U}, \mathcal{M}) - W^*(z, \mathcal{U}, \hat{\mathcal{M}}_t)| &= |z - Q(\hat{j}^*; z, \mathcal{U}, \hat{\mathcal{M}}_t)| \\ &\leq |Q(\hat{j}^*; z, \mathcal{U}, \mathcal{M}) - Q(\hat{j}^*; z, \mathcal{U}, \hat{\mathcal{M}}_t)|, \end{aligned} \tag{84}$$

which reduces to Case 2, requiring the same bounds.

Induction conclusion. Taking the maximum over all cases, and since $L_{0,\mu} = 0$ and $L_{0,\sigma} = 0$, by induction we have the following:

$$L_{n,\mu} = L_{n,\sigma} = n \cdot \max(2\lambda, \frac{\lambda}{\sigma_{\min}}). \tag{85}$$

For $n = K$ and the initial state ($z_0 = -\infty, \mathcal{U} = [K]$), we obtain $L^*_{K,\mu} = L^*_{K,\sigma} = K \max(2\lambda, \frac{\lambda}{\sigma_{\min}})$. Substituting $\Psi^\mu_t = \sum_{j\in[K]} B_{N^\mu_{t-1,j}}$, $\Psi^\sigma_t = \sum_{j\in[K]} B_{N^\sigma_{t-1,j}}$, it gives the claimed bound $K \max(2\lambda, \lambda/\sigma_{\min})\Psi^\mu_t + K \max(2\lambda, \lambda/\sigma_{\min})\Psi^\sigma_t$. □

I.4 Proof of Lemma 6.3(iii): planning error

Proof. Fix a state $(z, \mathcal{U})$ with $n = |\mathcal{U}|$ remaining arms. We first define

$$Q(j; z, \mathcal{U}, \hat{\mathcal{M}}_t) = \max_{\alpha\in\mathcal{A}} \Big(\mathbb{E}\big[W(\max\{z, \lambda\hat{Y}_j(\alpha)\}, \mathcal{U} \setminus \{j\}; \hat{\mathcal{M}}_t)\big] - c^{\mathrm{g}}_j - c^{\mathrm{e}}(\alpha) \Big). \tag{86}$$

Thus, $W^*(z, \mathcal{U}; \hat{\mathcal{M}}_t) = \max\{z, \max_{j\in\mathcal{U}} Q(j; z, \mathcal{U}, \hat{\mathcal{M}}_t)\}$. Let $j^* \in \arg\max_{j\in\mathcal{U}} Q(j; z, \mathcal{U}, \hat{\mathcal{M}}_t)$ be an optimal first arm under $\hat{\mathcal{M}}_t$. If $\hat{\pi}_t$ continues probing arms from $(z, \mathcal{U})$, let $\hat{j}$ be the arm it probes first. We prove by induction on n that there exist constants $L^\epsilon_{n,\mu}, L^\epsilon_{n,\sigma}$ such that

$$\begin{aligned} &W^*(z, \mathcal{U}; \hat{\mathcal{M}}_t) - W(z, \mathcal{U}; \hat{\mathcal{M}}_t, \hat{\pi}_t) \\ &\leq L^\epsilon_{n,\mu} \sum_{j\in\mathcal{U}} B_{N^\mu_{t-1,j}} + L^\epsilon_{n,\sigma} \sum_{j\in\mathcal{U}} \sup_{\alpha\in[0,1]} B^\sigma_{t-1,j}(\alpha) + 4n(L_\alpha\eta_\alpha + \eta_z). \end{aligned} \tag{87}$$

Base case ($n = 0$). $W^*(z, \emptyset, \cdot) = z$. The difference is 0, so $L^\epsilon_{0,\mu} = 0$ and $L^\epsilon_{0,\sigma} = 0$ suffice.

Inductive step. Assume the induction hypothesis holds for all sets of size $n - 1$.

Case 1: Both policies stop. Then $W^*(z, \mathcal{U}, \hat{\mathcal{M}}_t) = W(z, \mathcal{U}, \hat{\mathcal{M}}_t, \hat{\pi}_t) = z$, which requires $L^\epsilon_{n,\mu} \geq 0$ and $L^\epsilon_{n,\sigma} \geq 0$.

Case 2: Both continue with the same arm ($j^* = \hat{j}$). Condition on the first arm being the same. The remaining difference comes from possible mismatch in the chosen control at the first step, and continuation suboptimality on the suffix. Then, we have:

$$\begin{aligned} &|W^*(z, \mathcal{U}, \hat{\mathcal{M}}_t) - W(z, \mathcal{U}, \hat{\mathcal{M}}_t, \hat{\pi}_t)| \\ &\leq \big(2\lambda + L^\epsilon_{n-1,\mu}\big) \sum_{j\in\mathcal{U}} B_{N^\mu_{t-1,j}} + \left(\frac{\lambda}{\sigma_{\min}} + L^\epsilon_{n-1,\sigma}\right) \sum_{j\in\mathcal{U}} B_{N^\sigma_{t-1,j}}. \end{aligned} \tag{88}$$

This requires:

$$L^\epsilon_{n,\mu} \geq 2\lambda + L^\epsilon_{n-1,\mu}, \qquad L^\epsilon_{n,\sigma} \geq \frac{\lambda}{\sigma_{\min}} + L^\epsilon_{n-1,\sigma}. \tag{89}$$

Case 3: Both policies continue but with different first arms ($j^* \neq \hat{j}$). Suppose $\hat{j}$ appears at position ℓ in an optimal probing order $\pi^* = (j^* = \pi^*_1, \pi^*_2, \ldots, \pi^*_\ell = \hat{j}, \ldots, \pi^*_n)$ under $\hat{\mathcal{M}}_t$. Define a strict partial order $\prec_{\hat{j}}$ on $\mathcal{U}$, where $\hat{j} \prec_{\hat{j}} u$ for all $u \in \mathcal{U} \setminus \{\hat{j}\}$, and for any $u, v \in \mathcal{U} \setminus \{\hat{j}\}$, set

$u \prec_{\hat{j}} v$ iff u appears before v in π^*. Let $\pi^\#$ be a linear extension of $\prec_{\hat{j}}$, then $\pi^\#$ is π^* with $\hat{j}$ moved to the first position and can be obtained from π^* by $\ell - 1$ adjacent swaps. By Lemma I.1, there exists a sequence of permutations $\{\pi^{(m)}\}_{m=0}^{\ell-1}$ with $\pi^{(0)} = \pi^*$ and $\pi^{(\ell-1)} = \pi^\#$, where each $\pi^{(m)}$ is obtained from $\pi^{(m-1)}$ by swapping one adjacent pair. Thus,

$$\begin{aligned}|Q(j^*; z, \mathcal{U}, \hat{\mathcal{M}}_t) - Q(\hat{j}; z, \mathcal{U}, \hat{\mathcal{M}}_t)| &= |W(z, \mathcal{U}, \pi^{(0)}; \hat{\mathcal{M}}_t) - W(z, \mathcal{U}, \pi^{(\ell-1)}; \hat{\mathcal{M}}_t)| \\ &\leq \sum_{m=1}^{\ell-1} |W(z, \mathcal{U}, \pi^{(m-1)}; \hat{\mathcal{M}}_t) - W(z, \mathcal{U}, \pi^{(m)}; \hat{\mathcal{M}}_t)|. \quad (90)\end{aligned}$$

Consider an adjacent swap in Eq. (90) that exchanges arms b then a into a then b. Let z_k be any current baseline after k probes. The swap loss is bounded by the following:

$$|W(\ldots, a, b, \ldots; \hat{\mathcal{M}}_t) - W(\ldots, b, a, \ldots; \hat{\mathcal{M}}_t)| \leq \mathbb{E}_{z_k}\big[(-\Delta_{ab}(z_k; \hat{\mathcal{M}}_t))_+\big]. \quad (91)$$

Under `SeqPolicy`, a precedence constraint $(b \prec_t a)$ is added only when it is high-confidence: either via CI separation, or via the robust exchange test with threshold $\tau_{t-1,a,b}$. For overlapping-CI pairs where the exchange test applies, this implies

$$\underline{\Delta}_{ab} = \min_{g \in \mathcal{G}_{t-1,a,b}} \hat{\Delta}_{ab}(g) - 2\eta_z \; > \; -\tau_{t-1,a,b}. \quad (92)$$

By Lemma H.4, $\inf_{z \in \mathcal{Z}} \hat{\Delta}_{ab}(z) \geq \underline{\Delta}_{ab}$, hence

$$\mathbb{E}_{z_k}\big[(-\hat{\Delta}_{ab}(z_k; \hat{\mathcal{M}}_t))_+\big] \leq \sup_{z \in \mathcal{Z}}(-\hat{\Delta}_{ab}(z; \hat{\mathcal{M}}_t))_+ \leq \tau_{t-1,a,b}. \quad (93)$$

Consider the worst case where $\ell = n$ (i.e., $\hat{j}$ is at the end of the optimal sequence), requiring $n - 1$ exchanges. Since each $\tau_{t-1,a,b}$ involves at most 2 arms and $\tau_{t-1,a,b} = \epsilon^{\Delta}_{t-1,a,b} + 4\eta_z$, where Lemma 6.2 gives the explicit bound:

$$\epsilon^{\Delta}_{t-1,a,b} \leq C^{\mu}_{\Delta} \max_{k \in \{a,b\}} B_{N^{\mu}_{t-1,k}} + C^{\sigma}_{\Delta} \max_{k \in \{a,b\}} B_{N^{\sigma}_{t-1,k}} + 4L_\alpha \eta_\alpha. \quad (94)$$

Therefore,

$$\begin{aligned}&|Q(j^*; z, \mathcal{U}, \hat{\mathcal{M}}_t) - Q(\hat{j}; z, \mathcal{U}, \hat{\mathcal{M}}_t)| \\ &\leq (n-1)\left(C^{\mu}_{\Delta} \max_{k \in \{a,b\}} B_{N^{\mu}_{t-1,k}} + C^{\sigma}_{\Delta} \max_{k \in \{a,b\}} B_{N^{\sigma}_{t-1,k}} + 4L_\alpha \eta_\alpha\right). \quad (95)\end{aligned}$$

This requires:

$$L^{\epsilon}_{n,\mu} \geq 2n \max\{2\lambda, \frac{\lambda}{\sigma_{\min}}\} + L^{\epsilon}_{n-1,\mu}, \qquad L^{\epsilon}_{n,\sigma} \geq 2n \max\{2\lambda, \frac{\lambda}{\sigma_{\min}}\} + L^{\epsilon}_{n-1,\sigma}. \quad (96)$$

Case 4: Stopping decisions diverge. This case occurs when $W^*(z, \mathcal{U}; \hat{\mathcal{M}}_t) > z$ and $W(z, \mathcal{U}, \hat{\mathcal{M}}_t, \hat{\pi}_t) = z$, which can be reduced to a pure ordering loss under the same model, bounded by the adjacent-swap argument in Case 3.

Solving the recurrence. Starting from $L^{\epsilon}_{0,\mu} = 0$ and $L^{\epsilon}_{0,\sigma} = 0$, we obtain:

$$L^{\epsilon}_{n,\mu} = L^{\epsilon}_{n,\sigma} = n(n+1) \max\{2\lambda, \frac{\lambda}{\sigma_{\min}}\}. \quad (97)$$

For $n = K$ and the initial state $(z_0 = -\infty, \mathcal{U} = [K])$, it has $L^{\epsilon}_{K,\mu} = L^{\epsilon}_{K,\sigma} = \Theta(K^2) \max\{2\lambda, \frac{\lambda}{\sigma_{\min}}\}$. Moreover, the α-grid term contributes at most $\sum_{n=1}^{K} 4nL_\alpha\eta_\alpha = \Theta(K^2)L_\alpha\eta_\alpha$ and the z-grid term contributes at most $\sum_{n=1}^{K} 4n\eta_z = \Theta(K^2)\eta_z$ across the recursion. Substituting $\Psi^{\mu}_t = \sum_{j \in [K]} B_{N^{\mu}_{t-1,j}}$, $\Psi^{\sigma}_t = \sum_{j \in [K]} B_{N^{\sigma}_{t-1,j}}$, it gives the claimed bound $C^{\mathrm{err}} K^2 (\max(4\lambda, \frac{2\lambda}{\sigma_{\min}})(\Psi^{\mu}_t + \Psi^{\sigma}_t) + L_\alpha\eta_\alpha + \eta_z)$, with a constant $C^{\mathrm{err}} > 0$. This completes the proof. □

I.5 Proof of Lemma 6.4

Proof. Recall that the noise-update rounds are those with $\mathcal{I}(t) = 1$, where $\mathcal{I}(t) = 1$ iff $t \in \{1^2, 2^2, \ldots, \lfloor\sqrt{T}\rfloor^2\}$. On such rounds, the algorithm may incur additional cost due to extra replicate queries used only for noise labels, and an auxiliary forced noise-exploration probe as described in Section 5.1. Let $\varrho_t \geq 0$ denote the (random) additional cost incurred in round t. Clearly, $\varrho_t = 0$ if $\mathcal{I}(t) = 0$. Then, we consider $\mathcal{I}(t) = 1$.

Replicate-query overhead. For each arm $i \in \mathcal{O}_t$ that is probed by the within-round policy at fidelity $\alpha_{t,i}$, our algorithm may request one additional conditionally independent replicate at the same fidelity to form the unbiased noise label. Thus, the extra cost is evaluation-only, namely $c^{\mathrm{e}}(\alpha_{t,i})$, because the replicate reuses the same generated response. Since $c^{\mathrm{e}}(\alpha) \leq c_{\max}$ for all α, the total replicate-query overhead on round t is bounded by

$$\sum_{i \in \mathcal{O}_t} c^{\mathrm{e}}(\alpha_{t,i}) \leq |\mathcal{O}_t|\, c_{\max} \leq K c_{\max}. \tag{98}$$

Forced noise-exploration overhead. In addition, on update rounds the algorithm may perform at most one auxiliary forced noise-exploration probe to speed up posterior contraction uniformly across arms/fidelities. Clearly, its additional cost is at most $c_{\max}$ in each noise-update round.

Let $N_{\mathrm{upd}} = |\{t \leq T : \mathcal{I}(t) = 1\}|$ be the number of update rounds. Combining these two overheads, we have

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[\varrho_t] = \sum_{t:\mathcal{I}(t)=1} \mathbb{E}[\varrho_t] &\leq \sum_{t:\mathcal{I}(t)=1} \big(K c_{\max} + c_{\max}\big) = N_{\mathrm{upd}}\big(K c_{\max} + c_{\max}\big) \\ &\leq (\lfloor\sqrt{T}\rfloor + 1)\big(K c_{\max} + c_{\max}\big),\end{aligned} \tag{99}$$

which implies the claimed overhead bound. □

I.6 Proof of Theorem 6.5

Proof. **Single-round decomposition.** Let U_t^{alg} denote the realized net utility of `DialBandit` in round t. On noise-update rounds, the algorithm may additionally request replicate samples and one forced noise-exploration probe, which incurs the replication overhead as described in Lemma 6.4. Hence, $U_t^{\mathrm{alg}} = U_t(\hat{\pi}_t) - \varrho_t$. For each $t \in [T]$, we define the per-round regret as

$$\mathrm{Reg}_t = W_t^*(\mathcal{M}) - \mathbb{E}[U_t^{\mathrm{alg}}] = W_t^*(\mathcal{M}) - W_t(\mathcal{M}; \hat{\pi}_t) + \mathbb{E}[\varrho_t], \tag{100}$$

where $W_t^*(\mathcal{M})$ is the oracle optimum (Definition 3.1), $W_t(\mathcal{M}; \hat{\pi}_t)$ is the expected net utility of the within-round policy $\hat{\pi}_t$ of our algorithm, and $\mathbb{E}[\varrho_t]$ accounts for the additional costs incurred by noise learning.

Decompose policy error. Fix round t. We first decompose the policy error $W_t^*(\mathcal{M}) - W_t(\mathcal{M}; \hat{\pi}_t)$ into three stability terms as follows:

$$\begin{aligned}W_t^*(\mathcal{M}) - W_t(\mathcal{M}; \hat{\pi}_t) \leq &\underbrace{\big|W_t^*(\mathcal{M}) - W_t^*(\hat{\mathcal{M}}_t)\big|}_{\text{(A) Optimal Value Shift}} + \underbrace{\big|W_t^*(\hat{\mathcal{M}}_t) - W_t(\hat{\mathcal{M}}_t; \hat{\pi}_t)\big|}_{\text{(B) Planning Error}} \\ &+ \underbrace{\big|W_t(\hat{\mathcal{M}}_t; \hat{\pi}_t) - W_t(\mathcal{M}; \hat{\pi}_t)\big|}_{\text{(C) Model Misspecification}},\end{aligned} \tag{101}$$

where (A) is the optimal value shift, (B) is the planning error, and (C) is the model misspecification gap. Condition on the realized within-round trajectory of $\hat{\pi}_t$. Let $\mathcal{O}_{t,m}$ be the set of arms probed before stage m, and let $\mathcal{O}_t$ be the final probed set when the policy stops. Define the remaining-to-be-probed set along this realized trajectory as $\mathcal{U}_{t,m} = \mathcal{O}_t \setminus \mathcal{O}_{t,m-1}$, so that $\mathcal{U}_{t,m} \subseteq \mathcal{O}_t$ and $|\mathcal{O}_t| \leq K$. Applying Lemma 6.3(ii) to (A), Lemma 6.3(iii) to (B), and Lemma 6.3(i) to (C), we obtain bounds in terms of stage-wise sums of confidence radii along the within-round recursion. Specifically, at stage m with remaining-arm set $\mathcal{U}_{t,m}$, the bounds take the form $\sum_{j \in \mathcal{U}_{t,m}} B_{N^{\mu}_{t-1,j}}$ and $\sum_{j \in \mathcal{U}_{t,m}} \max_{\alpha \in \mathcal{A}} B^{\sigma}_{t-1,j}(\alpha)$, which are up to problem-dependent constants. Since each arm in

$\mathcal{O}_t$ can belong to $\mathcal{U}_{t,m}$ for at most $|\mathcal{O}_t| \le K$ stages, we have

$$\sum_{m=1}^{|\mathcal{O}_t|} \sum_{j \in \mathcal{U}_{t,m}} B_{N^\mu_{t-1,j}} \le K \sum_{i \in \mathcal{O}_t} B_{N^\mu_{t-1,i}}, \qquad \sum_{m=1}^{|\mathcal{O}_t|} \sum_{j \in \mathcal{U}_{t,m}} \max_{\alpha \in \mathcal{A}} B^\sigma_{t-1,j}(\alpha) \le K \sum_{i \in \mathcal{O}_t} \max_{\alpha \in \mathcal{A}} B^\sigma_{t-1,i}(\alpha). \tag{102}$$

Note that although each arm is probed at most once within a round, a stage-wise analysis may sum confidence radii over the remaining set at each stage. Hence, a single arm's radius can be counted multiple times across stages until it is removed from the remaining set. Since there are at most K stages, each arm can be counted a maximum of K times. Therefore, on the clean event $\mathcal{E}$, the total contribution of (A)–(C) is bounded by K times the realized per-round sums over $\mathcal{O}_t$, yielding the following bound:

$$\text{(A)} + \text{(B)} + \text{(C)} \le C^{\mathrm{reg}} K^3 \left(\sum_{i \in \mathcal{O}_t} B_{N^\mu_{t-1,i}} + \sum_{i \in \mathcal{O}_t} \max_{\alpha \in \mathcal{A}} B^\sigma_{t-1,i}(\alpha) \right) + C^{\mathrm{reg}} K^2 (L_\alpha \eta_\alpha + \eta_z), \tag{103}$$

for a universal constant $C^{\mathrm{reg}} > 0$.

Accumulation of the mean term. Summing the mean-radius contribution over all rounds and probes follows the usual argument: each time arm i is probed, $N^\mu_{t,i}$ increments by one, hence

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{O}_t} B_{N^\mu_{t-1,i}} \le \tilde{\mathcal{O}}\Big(K\sqrt{T}\Big). \tag{104}$$

Accumulation of the noise term. We use epoch-wise accumulation to bound the noise term. Let the update times be $s_m = m^2$ for $m = 1, \dots, M$ and epochs $\mathcal{T}_m = \{t : s_m \le t < s_{m+1}\}$ with length Δ_m. Define the *global posterior-uncertainty level* at the start of epoch m as

$$u_m = \max_{i \in [K]} \max_{\alpha \in \mathcal{A}} \sqrt{v^\sigma_{s_m,i}(\alpha)}. \tag{105}$$

By the forced noise-exploration step on update rounds, the algorithm probes (with replication) a pair $(i^{\mathrm{exp}}_m, \alpha^{\mathrm{exp}}_m) \in [K] \times \mathcal{A}$ satisfying $\sqrt{v^\sigma_{s_m, i^{\mathrm{exp}}_m}(\alpha^{\mathrm{exp}}_m)} = u_m$, and appends the resulting unbiased label to the corresponding noise dataset, as described in Section 5.1. For any $t \in \mathcal{T}_m$, since $N^\sigma_{t-1,i} \le M$ and β_n is nondecreasing, $\sqrt{\beta_{N^\sigma_{t-1,i}}} \le \sqrt{\beta_M}$. Moreover, GP posterior noise is nonincreasing with more data, so $v^\sigma_{t-1,i}(\alpha) \le v^\sigma_{s_m,i}(\alpha) \le u_m^2$ for all $\alpha \in \mathcal{A}$. Recalling the uniform radius definition $B_{N^\sigma_{t-1,i}} = 2\sqrt{\beta_{N^\sigma_{t-1,i}}} \sup_{\alpha \in [0,1]} \sqrt{v^\sigma_{t-1,i}(\alpha)}$, we have $B_{N^\sigma_{t-1,i}} \le 2\sqrt{\beta_M} u_m$. Hence,

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{O}_t} B_{N^\sigma_{t-1,i}} \le \sum_{m=1}^{M} \sum_{t \in \mathcal{T}_m} \sum_{i \in \mathcal{O}_t} 2\sqrt{\beta_M} u_m \le 2K\sqrt{\beta_M} \sum_{m=1}^{M} \Delta_m u_m. \tag{106}$$

Since $\Delta_m \le \max_m \Delta_m \le 2\sqrt{T} + 1 = \tilde{O}(\sqrt{T})$, we have

$$\sum_{m=1}^{M} \Delta_m u_m \le \tilde{O}(\sqrt{T}) \sum_{m=1}^{M} u_m. \tag{107}$$

Finally, based on the standard GP information-gain aggregation to the forced-exploration sequence, it has

$$\sum_{m=1}^{M} u_m = \sum_{m=1}^{M} \sqrt{v^\sigma_{s_m, i^{\mathrm{exp}}_m}(\alpha^{\mathrm{exp}}_m)} \le \tilde{O}\left(\sqrt{M \cdot \max_i \gamma_{M,i}} \right). \tag{108}$$

Substituting $M = \lfloor \sqrt{T} \rfloor$ into Eq. (106) yields

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{O}_t} B_{N^\sigma_{t-1,i}} \le \tilde{O}\left(K T^{3/4} \sqrt{\max_i \gamma_{M,i}} \right). \tag{109}$$

Bound replication overhead. By Lemma 6.4, we have

$$\sum_{t=1}^{T}\mathbb{E}[\varrho_t] \le |\{t \le T : \mathcal{I}(t)=1\}| \cdot (K+1)c_{\max} \le (\lfloor\sqrt{T}\rfloor+1)(K+1)c_{\max}. \tag{110}$$

Combining Eq. (104) and Eq. (109), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\big(W_t^*(\mathcal{M}) - W_t(\mathcal{M};\hat{\pi}_t)\big) \\
&\le C^{\mathrm{reg}}K^3\left(\sum_{t=1}^{T}\sum_{i\in\mathcal{O}_t} B_{N^{\mu}_{t-1,i}} + \sum_{t=1}^{T}\sum_{i\in\mathcal{O}_t} B_{N^{\sigma}_{t-1,i}}\right) + C^{\mathrm{reg}}K^2T(L_\alpha\eta_\alpha+\eta_z) \\
&\le C^{\mathrm{reg}}K^3\left(\tilde{\mathcal{O}}\Big(K\sqrt{T}\Big) + \tilde{O}\left(KT^{3/4}\sqrt{\max_i \gamma_{M,i}}\right)\right) + C^{\mathrm{reg}}K^2T(L_\alpha\eta_\alpha+\eta_z).
\end{aligned} \tag{111}$$

Total regret bound. On the event $\mathcal{E}$, with $\eta_\alpha = \eta_z = T^{-1/4}$, we have

$$\begin{aligned}
\mathrm{Reg}_T &= \sum_{t=1}^{T}\big(W_t^*(\mathcal{M}) - W_t(\mathcal{M};\hat{\pi}_t)\big) + \sum_{t=1}^{T}\mathbb{E}[\varrho_t] \\
&\le C^{\mathrm{reg}}K^3\left(\tilde{\mathcal{O}}\Big(K\sqrt{T}\Big) + \tilde{O}\left(KT^{3/4}\sqrt{\max_i \gamma_{M,i}}\right)\right) \\
&\quad + C^{\mathrm{reg}}K^2T(L_\alpha\eta_\alpha+\eta_z) + (\lfloor\sqrt{T}\rfloor+1)(K+1)c_{\max} \\
&\le \tilde{\mathcal{O}}\left(\mathrm{poly}(K)\lambda T^{3/4}\sqrt{\max_i \gamma_{\lfloor\sqrt{T}\rfloor,i}}\right).
\end{aligned} \tag{112}$$

□

I.7 Discussion and limitations in Subsection 6.2

1. Assumption of fidelity and observation model. The base model treats fidelity as an evaluation-side control: increasing α changes only the probing cost $c^{\mathrm{e}}(\alpha)$ and the observation-noise level $\sigma_i^2(\alpha)$, not the latent query-specific quality $Q_{t,i}$. The monotonicity assumption that $\sigma_i^2(\alpha)$ decreases with fidelity captures settings such as repeated judging or evaluator ensembles, where spending more evaluation compute makes the same response more reliably assessed. This assumption may fail for knobs that also alter the candidate response, evaluator prompt, or sampling distribution, such as generation-side self-consistency or stochastic decoding of the candidate answer; such knobs fall outside the pure observation-noise model and would require redefining the target quality or moving to an objective-changing multi-fidelity formulation. For mild empirical nonmonotonicity within a fixed evaluation protocol, one can plan on a monotone projection of the learned noise curve on the fidelity grid, with the projection mismatch treated as approximation error. Finally, the Gaussian model gives the closed-form posterior score in Eq. (2) and simplifies constants. The concentration steps in the regret proof rely mainly on bounded/sub-Gaussian tails; extending the full formulation to bounded scores would require a calibrated score rule satisfying the same Lipschitz and concentration conditions.

2. Overhead-regret tradeoff and the gap to the continuous-control lower bound. The continuous-control lower bound $\Omega(\sqrt{T\gamma_T})$ reflects the statistical difficulty of learning an RKHS function when informative noise feedback is available in essentially every round. In `DialBandit`, however, unbiased noise information is obtained only on update rounds $\mathcal{I}(t)=1$, totaling about $M = \lfloor\sqrt{T}\rfloor$ auxiliary labels. This design keeps the additional overhead sublinear (see Lemma 6.4) but reduces the effective sample size for GP regression from T to M. Consequently, the noise-learning component scales as $\tilde{O}\big(\frac{T}{\sqrt{M}}\sqrt{\gamma_M}\big)$, yielding $\tilde{O}\big(T^{3/4}\sqrt{\gamma_{\lfloor\sqrt{T}\rfloor}}\big)$ under the default choice $M = \Theta(\sqrt{T})$. We choose $M = \Theta(\sqrt{T})$ for a clean epoch-based argument and a simple overhead-accuracy balance; other schedules (*e.g.,* $t_m = \lfloor m^p\rfloor$ or doubling) can be analyzed similarly and allow trading overhead for improved noise learning.

Importantly, the lower bound $\Omega(\sqrt{T\gamma_T})$ does not account for the overhead constraint induced by sparse noise updates. Thus, the rate gap is an explicit design tradeoff rather than a contradiction, where increasing the update frequency (larger M) pushes the regret toward $\tilde{O}(\sqrt{T\gamma_T})$, but necessarily increases overhead roughly linearly in the number of auxiliary queries. Closing the remaining gap by improving the overhead-regret tradeoff or proving a matching lower bound under sparse updates remains an interesting direction for future work.

3. Conditional lower bound under an auxiliary-replication budget. The following statement is *not* a lower bound for the base model itself. Instead, it quantifies a regret-overhead tradeoff when the learner is allowed only a limited number of auxiliary replication samples, such as independent replicate samples used to form unbiased noise labels. Formally, let $M \in \{0, 1, \ldots, T\}$ and consider algorithms that, in T rounds, require at most M auxiliary noise labels. Then the following proposition holds:

Proposition I.2 (Budgeted-auxiliary-replication lower bound (informal))**.** *Fix $T \geq 1$ and $M \in \{1, \ldots, T\}$. There exists a subclass of instances such that for any algorithm that requests at most M auxiliary noise labels over T rounds,*

$$\mathbb{E}[\mathrm{Reg}_T] \gtrsim \frac{T}{\sqrt{M}}\sqrt{\gamma_M}, \tag{113}$$

where γ_M is the maximum information gain for kernel k on $[0, 1]$. In particular, for $M = \Theta(\sqrt{T})$ this yields a term $\Omega(T^{3/4}\sqrt{\gamma_{\lfloor\sqrt{T}\rfloor}})$, which matches the noise-driven rate in Theorem 6.5.

Proof sketch. Reduce to a single-arm ($K = 1$) continuous-control subclass with $Y_t(\alpha) \sim \mathcal{N}(0, f(\alpha))$, where $f \in \mathcal{H}_k$ and $f(\alpha) \in [\sigma^2_{\min}, \Sigma_{\max}]$. Using the closed form $\mathbb{E}[\max(X(\alpha), 0)] \propto \sqrt{f(\alpha)}$ and a local perturbation $f(\alpha) = \sigma_0^2 + \varepsilon h(\alpha)$ yields an effective objective that is linear in $h(\alpha)$ up to a uniform $O(\varepsilon^2)$ remainder. In the budget, the learner can obtain at most M unbiased labels for $f(\alpha)$ (hence $h(\alpha)$ up to scaling). Standard kernelized bandit lower bounds over M informative observations give a $\Omega(\sqrt{M\gamma_M})$ difficulty in the query domain, and epoch amplification over T rounds, with only M informative rounds, yields the stated $\Omega(\frac{T}{\sqrt{M}}\sqrt{\gamma_M})$ term.

4. The rationale for using replicate-based noise labels. If no independent replicate is available, a natural noise proxy at fidelity α is the squared residual $s_{t,i} = (y_{t,i}(\alpha) - \hat{\mu}_{t-1,i})^2$. Even under the standard decomposition $y_{t,i}(\alpha) = \mu_i + \varepsilon_{t,i}(\alpha)$ with $\mathbb{E}[\varepsilon_{t,i}(\alpha) \mid \mathcal{F}_{t-1}, \alpha] = 0$ and $\mathbb{E}[\varepsilon_{t,i}(\alpha)^2 \mid \mathcal{F}_{t-1}, \alpha] = \sigma_i^2(\alpha)$, we have

$$\mathbb{E}[s_{t,i} \mid \mathcal{F}_{t-1}, \alpha] = \sigma_i^2(\alpha) + (\mu_i - \hat{\mu}_{t-1,i})^2. \tag{114}$$

Hence, the regression target is biased by the mean-tracking error term $b_{t,i} = (\mu_i - \hat{\mu}_{t-1,i})^2$. On the clean event, we have $|\mu_i - \hat{\mu}_{t-1,i}| \leq B_{N^\mu_{t-1,i}}$, hence

$$b_{t,i} \leq B^2_{N^\mu_{t-1,i}} = \mathcal{O}\left(\frac{\log(T/\rho')}{\max\{1, N^\mu_{t-1,i}\}}\right). \tag{115}$$

Crucially, this bias is *not* controlled by the noise-sample size $N^\sigma_{t-1,i}$. Even if the noise learner collects N^σ biased labels, the estimator concentrates around $\sigma_i^2(\cdot) + b_{t,i}$ rather than $\sigma_i^2(\cdot)$, unless $b_{t,i}$ is negligible at the update times.

Under sparse noise updates (*i.e.,* $\mathcal{I}(t) = 1$ on $t = m^2$), the cumulative impact of this bias on the noise-learning pipeline is governed by its accumulation over update rounds. For example, if arm i is only probed on update rounds, then $N^\mu_{t_m-1,i} \asymp m$ at $t_m = m^2$, so

$$\sum_{m \leq \sqrt{T}} b_{t_m,i} \lesssim \sum_{m \leq \sqrt{T}} \frac{\log(T/\rho')}{m} = \mathcal{O}(\log(T/\rho') \log T), \tag{116}$$

which yields an additional non-trivial term in the confidence analysis beyond the standard kernelized self-normalized radius. More importantly, there is no guarantee that $N^\mu_{t-1,i}$ grows fast enough uniformly across arms, leading to situations where rarely probed arms maintain $N^\mu_{t-1,i} = O(1)$ over

many update times. This results in $b_{t,i} = \Omega(1)$ for some arms. In these cases, noise estimates can become systematically biased, particularly on arms and fidelities with the highest uncertainty. This bias may result in repeated suboptimal probing decisions, leading to ongoing model misspecification and significantly worse regret bounds in adversarial scenarios, unless additional mechanisms are implemented.

In contrast, `DialBandit` uses a single independent replicate $y'_{t,i}(\alpha)$ at the same (i, α), so that the paired-difference label $\tilde{s}_{t,i} = \frac{(y_{t,i}(\alpha) - y'_{t,i}(\alpha))^2}{2}$ satisfies $\mathbb{E}[\tilde{s}_{t,i} \mid \mathcal{F}_{t-1}, \alpha] = \sigma_i^2(\alpha)$, which eliminates the mean-induced bias and enables the standard kernelized self-normalized confidence bound (see Lemma 6.1).

5. Robustness of replicate-based noise labels to non-stationarity. We claim that the replicate-based label $\tilde{s}_{t,i} = \frac{(y_{t,i}(\alpha) - y'_{t,i}(\alpha))^2}{2}$ is conditionally unbiased for the *instantaneous* noise $\sigma_{t,i}^2(\alpha)$ and does not require a stationary mean. Even if the latent mean drifts over time, *i.e.,* non-stationary $\mu_{t,i}$, the two samples $y_{t,i}(\alpha)$ and $y'_{t,i}(\alpha)$ drawn within the same round share the same conditional mean, so the paired-difference construction can eliminate the mean and remains unbiased. In contrast, residual-based labels $(y_{t,i}(\alpha) - \hat{\mu}_{t-1,i})^2$ incur an additional drift-induced bias term. Extending `DialBandit` to non-stationary settings would require adapting the benchmark (*e.g.,* dynamic regret) and using time-local estimators, but the replicate mechanism naturally supports such extensions by providing unbiased noise information under drift.

6. Contextual routing. Suppose the policy observes a query context x_t before probing and uses an arm-specific predictor $f_i(x_t) = \phi(x_t)^\top \theta_i$, where $\phi(x_t)$ is a bounded representation. A conditional version of the observation model is

$$Q_{t,i} \mid x_t \sim \mathcal{N}(f_i(x_t), \kappa_i^2), \qquad Y_{t,i}(\alpha) \mid Q_{t,i}, x_t \sim \mathcal{N}(Q_{t,i}, \sigma_i^2(\alpha)). \tag{117}$$

With conditionally independent arms and context-independent fidelity noise, the single-round index construction applies pointwise after conditioning on x_t: replace μ_i by $f_i(x_t)$ in the posterior prize and define $\xi_i(x_t)$ through $\max_\alpha V_i(\xi_i(x_t), \alpha; x_t) = 0$. The contextual comparator observes the same x_t and knows the true conditional quality predictors and noise curves. For a linear predictor, a candidate confidence width is $\beta_{t,i}^\mu \|\phi(x_t)\|_{A_{t,i}^{-1}}$, where $A_{t,i} = \rho I + \sum_{s<t:\, i \in \mathcal{O}_s} \phi(x_s)\phi(x_s)^\top$. A contextual regret analysis must control this width under cascade feedback, because each arm is observed only on contexts for which it was probed, and then propagate the resulting errors through ordering and stopping. Same-response replicate differences still isolate evaluation noise conditional on the query and response when the replicate errors are independent. Context-dependent noise $\sigma_i^2(\alpha, x_t)$ requires learning over the joint context–fidelity domain.

J Evaluator-aligned quality and systematic bias

The latent quality $Q_{t,i}$ in Section 3 is the expected score under a fixed evaluator, prompt, rubric, and aggregation rule. Its relation to an external quality target can be written as

$$Y_{t,i}^{(r)} = Q_{t,i}^{\text{target}} + b_i(x_t) + \varepsilon_{t,i}^{(r)}, \qquad Q_{t,i} = Q_{t,i}^{\text{target}} + b_i(x_t), \tag{118}$$

where $b_i(x_t)$ is a stable evaluator-specific mean shift and $\mathbb{E}[\varepsilon_{t,i}^{(r)} \mid Q_{t,i}, x_t] = 0$. For n conditionally independent evaluations of the same response, averaging reduces the variance of the stochastic term by a factor of n while leaving $b_i(x_t)$ unchanged. Same-response replicate differences cancel the common mean shift and identify the stochastic variance:

$$\mathbb{E}\left[\frac{(Y_{t,i}^{(1)} - Y_{t,i}^{(2)})^2}{2} \,\middle|\, Q_{t,i}, x_t\right] = \mathrm{Var}(\varepsilon_{t,i}^{(r)} \mid Q_{t,i}, x_t). \tag{119}$$

The policy and its oracle benchmark therefore share a net-utility objective defined by the evaluator-aligned quality $Q_{t,i}$. Stable evaluator bias is incorporated into this fixed target, while fidelity controls observation noise. The original formulation applies under the standing assumptions on $Q_{t,i}$.

K Online learning for heterogeneity profile

This section formalizes an extension of `DialBandit` in which the cross-query heterogeneity profile $\{\kappa_i^2\}_{i=1}^K$ is learned online together with $\{\mu_i\}$ and $\{\sigma_i^2(\cdot)\}$.

K.1 Problem setup

Additional environment assumptions. To learn κ_i^2 online, we require two structural assumptions: (1) $Q_{t,i}$ are drawn independently across rounds t. This ensures the empirical variance of observations properly converges to the true marginal variance; (2) the true heterogeneity parameters lie within a known bounded interval $[\kappa^2_{\min}, \kappa^2_{\max}]$. This is required to clip the online estimates and maintain bounded Lipschitz constants in the planning layer.

K.2 Estimation strategy and algorithm modifications

Marginal variance decomposition. Under the independent query assumption, the marginal distribution of first-shot observations for each arm i and fidelity α decomposes additively:

$$\mathrm{Var}(Y_{t,i}(\alpha)) = \kappa_i^2 + \sigma_i^2(\alpha). \tag{120}$$

Replicate probes within a single round identify the evaluation noise $\sigma_i^2(\alpha)$ but do *not* reveal the cross-query signal heterogeneity κ_i^2. To cleanly isolate κ_i^2, we fix a single *reference fidelity* $\alpha^\circ \in \mathcal{A}$ and track the empirical variance of its first-shot observations across rounds.

The reference-fidelity estimator. Let $\mathcal{T}_{t,i}^{(\circ)}$ be the set of past rounds where arm i was probed at the chosen reference fidelity α°, and let $\bar{y}_{t,i}^{(\circ)}$ be the corresponding sample mean. We define the empirical marginal variance as:

$$\widehat{V}_{t,i}(\alpha^\circ) = \frac{1}{N_{t,i}^{(\circ)} - 1} \sum_{\tau \in \mathcal{T}_{t,i}^{(\circ)}} \big(y_{\tau,i}(\alpha^\circ) - \bar{y}_{t,i}^{(\circ)}\big)^2, \qquad N_{t,i}^{(\circ)} = |\mathcal{T}_{t,i}^{(\circ)}|. \tag{121}$$

Since $\widehat{V}_{t,i}(\alpha^\circ)$ is an unbiased estimator of $\kappa_i^2 + \sigma_i^2(\alpha^\circ)$, we isolate κ_i^2 by subtracting the current GP mean estimate $m^\sigma_{t-1,i}(\alpha^\circ)$ of the evaluation noise:

$$\hat{\kappa}^2_{t,i} = \Pi_{[\kappa^2_{\min}, \kappa^2_{\max}]}\Big(\widehat{V}_{t,i}(\alpha^\circ) - m^\sigma_{t-1,i}(\alpha^\circ)\Big), \tag{122}$$

where Π clips the result to the valid domain. Let $N_{t,i}^{(\circ)}$ denote the total number of times arm i has been probed at the reference fidelity α° up to round t. For early rounds where $N_{t,i}^{(\circ)} < 2$, $\hat{\kappa}^2_{t,i}$ is simply initialized to an arbitrary value in this domain. This one-dimensional identification step decouples κ-learning from the complex cross-fidelity GP posterior.

Conservative estimation. Since *Overestimating* κ_i^2 inflates the shrinkage weight $\hat{w}_{t,i}(\alpha) = \hat{\kappa}^2_{t,i}/(\hat{\kappa}^2_{t,i} + \hat{\sigma}^2(\alpha))$ and makes the learner over-trust low-fidelity observations, we replace the posterior mean $m^\sigma_{t-1,i}(\alpha^\circ)$ in Eq. (122) with the optimistic UCB $\hat{\sigma}^2_{t-1,i}(\alpha^\circ)$. Subtracting an upper bound yields a conservatively *small* estimate of κ_i^2. This ties κ-learning to the existing high-probability event for σ-estimation, and ensures the shrinkage map systematically errors on the side of under-trusting noisy observations. Importantly, conservative estimation affects only constants, because replacing m^σ by the UCB changes the bound on $|\hat{\kappa}^2_{t,i} - \kappa_i^2|$ by at most the same order as the existing σ-estimation radius, so the added term is absorbed into $B^\kappa_{t,i}$ without changing its t-exponent.

Plug-in surrogate score. Let $\hat{w}_{t,i}(\alpha) = \hat{\kappa}^2_{t-1,i}/(\hat{\kappa}^2_{t-1,i} + \hat{\sigma}^2_{t-1,i}(\alpha))$ and define the planning surrogate

$$\hat{S}_{t,i}(y, \alpha) = \lambda\Big(\hat{w}_{t,i}(\alpha)\, y + \big(1 - \hat{w}_{t,i}(\alpha)\big)\, \hat{\mu}_{t-1,i}\Big), \tag{123}$$

which replaces $S_i(y, \alpha)$. Meanwhile, the random prize $\hat{X}_{t,i}(\alpha) = \hat{S}_{t,i}(Y_{t,i}(\alpha), \alpha)$ substitutes for $X_i(\alpha) = S_i(Y_{t,i}(\alpha), \alpha)$ inside the same Weitzman's rule.

Modified Algorithm 1. With $\hat{\kappa}^2_{t-1,i}$ from (122), the following changes are made:

- **Model object:** $\hat{\mathcal{M}}_t = \{(\hat{\mu}_{t-1,i}, \hat{\kappa}^2_{t-1,i}, \hat{\sigma}^2_{t-1,i}(\cdot))\}_{i=1}^K$.
- **Posterior score:** use $w_{t,i}(\alpha) = \hat{\kappa}^2_{t-1,i}/(\hat{\kappa}^2_{t-1,i} + \hat{\sigma}^2_{t-1,i}(\alpha))$ inside $S_i(\cdot)$.
- **Update schedule:** refresh $\hat{\kappa}^2_{t,i}$ on the same sparse rounds as GP noise updates ($\mathcal{I}(t) = 1$) *after* $\hat{\mu}_{t,i}$ and (m^σ, v^σ) are updated, so that $\widehat{V}_{t,i}$ uses the latest first-shot trace at α°.
- **Exploration guard:** To ensure the reference-fidelity estimator receives enough samples to concentrate, we enforce a low-priority forced probe (i, α°) whenever $N^{(\circ)}_{t,i} \le c^\circ\sqrt{t}$.

Choice of α° in practical considerations. While any fixed α° theoretically preserves the $\tilde{\mathcal{O}}(T^{3/4})$ regret exponent, its selection dictates a critical finite-sample trade-off between estimation signal-to-noise ratio (SNR) and exploration cost. A low fidelity α° incurs a large evaluation noise $\sigma_i^2(\alpha^\circ)$, which eclipses the underlying κ_i^2 and amplifies the variance of the conservative subtraction. Conversely, a high fidelity minimizes $\sigma_i^2(\alpha^\circ)$ to yield clean κ_i^2 estimates, but severely inflates the $\mathcal{O}(\sqrt{T})$ forced exploration penalty. Consequently, deploying a mid-to-high fidelity optimally balances estimator variance against the cumulative probing cost in practical deployments.

K.3 Theoretical analysis

Lemma K.1 (Concentration of $\hat{\kappa}^2$)**.** *Under the additional structural assumptions, there exists a constant $C_\kappa > 0$ such that, with probability at least $1 - \rho'$, for all $t \in [T]$ and $i \in [K]$ with $N^{(\circ)}_{t,i} \ge 2$,*

$$\left|\hat{\kappa}^2_{t,i} - \kappa_i^2\right| \le B^\kappa_{t,i} = C_\kappa \left(\sqrt{\frac{\log(2KT/\rho')}{N^{(\circ)}_{t,i}}} + B_{N^\sigma_{t,i}} + \frac{1}{N^{(\circ)}_{t,i}} \right). \tag{124}$$

Proof. The bound follows by decomposing the error in the estimator (122). Let $N^{(\circ)}_{t,i}$ denote the total number of times arm i has been probed at the reference fidelity α° up to round t. **(1)** The first term arises from the concentration of the empirical variance $\widehat{V}_{t,i}(\alpha^\circ)$ around the true marginal variance $\kappa_i^2 + \sigma_i^2(\alpha^\circ)$. Because the observations y are bounded within $[-Y_{\max}, Y_{\max}]$ on the clean event $\mathcal{E}$, their squares are also bounded. This allows us to apply Hoeffding's inequality, which guarantees a standard parametric decay rate of $\mathcal{O}(1/\sqrt{N^{(\circ)}_{t,i}})$. **(2)** The second term bounds the error $|m^\sigma_{t-1,i}(\alpha^\circ) - \sigma_i^2(\alpha^\circ)|$ from GP estimate for the evaluation noise. Crucially, under the *conservative estimation* rule, we replace the posterior mean m^σ with the UCB $\hat{\sigma}^2$. This introduces an additional deviation exactly equal to the GP confidence width $\beta\sigma_{t-1,i}(\alpha^\circ)$, which is controlled by $B_{N^\sigma_{t,i}}$. **(3)** The third term is a structural bias from replication samples in Eq. (121), which leads to an expected squared error, *i.e.,* $\mathcal{O}(1/N^{(\circ)}_{t,i})$. □

To integrate this bound, we augment the clean event with the κ-concentration event:

$$\mathcal{E}_\kappa = \{\forall t \in [T], \forall i \in [K] : |\hat{\kappa}^2_{t-1,i} - \kappa_i^2| \le B^\kappa_{t-1,i}\}, \qquad \mathcal{E}^{\text{ext}} = \mathcal{E} \cap \mathcal{E}_\kappa. \tag{125}$$

By a union bound, $\mathbb{P}(\mathcal{E}^{\text{ext}}) \ge 1 - (\rho + \rho')$. On $\mathcal{E}^{\text{ext}}$, because the posterior-score map $(\mu, \kappa^2, \sigma^2) \mapsto S_i(y, \alpha)$ is Lipschitz continuous on the compact domain $[\kappa^2_{\min}, \kappa^2_{\max}] \times [\sigma^2_{\min}, \Sigma_{\max}]$ uniformly over bounded y, the stability results (Lemma 6.2) immediately extend by appending an $\mathcal{O}(B^\kappa_{t-1,i})$ deviation to the existing planning radii $\epsilon^\xi_{t,i}$ and $\epsilon^\Delta_{t,i,j}$.

Theorem K.2 (Regret with Online κ Learning)**.** *With the structural assumptions for online κ-learning and Lemma K.1. On $\mathcal{E}^{\text{ext}}$, let $B^\kappa_{t,\max} = \max_{i\in[K]} B^\kappa_{t,i}$, and we have*

$$\text{Reg}_T \le \tilde{\mathcal{O}}\big(\text{poly}(K) T^{3/4} \sqrt{\gamma_{\lfloor\sqrt{T}\rfloor}}\big) + \tilde{\mathcal{O}}\left(\text{poly}(K) \sum_{t=1}^T B^\kappa_{t-1,\max} \right) + \tilde{\mathcal{O}}(K\sqrt{T}). \tag{126}$$

Proof. Conditional on the extended clean event $\mathcal{E}^{\text{ext}}$, the total regret decomposes additively into three components: (1) **Planning Regret:** the $\tilde{\mathcal{O}}(T^{3/4})$ is inherited from Theorem 6.5, representing the inherent cost of sequential search under the estimated model. (2) **Estimation Penalty:** the cumulative

Table 1: Arm-level offline calibration summary.

	AlpacaEval		MMLU		AlpacaEval
Arm	μ_i	κ_i^2	μ_i	κ_i^2	$\sigma_i^2(n_{\mathrm{ref}} = 16)$
1	0.841	0.010	0.592	0.242	0.012
2	0.853	0.008	0.711	0.206	0.011
3	0.741	0.045	0.586	0.243	0.023
4	0.838	0.012	0.684	0.216	0.018
5	0.824	0.020	0.719	0.202	0.021
6	0.737	0.031	0.684	0.216	0.041
7	0.807	0.017	0.647	0.228	0.020

error $\sum_{t=1}^{T} B^{\kappa}_{t-1,\max}$ induced by substituting $\hat{\kappa}^2_{t-1,i}$ for the true κ_i^2. Because the posterior-score map $S_i(y, \alpha)$ is Lipschitz continuous with respect to κ^2 on $[\kappa^2_{\min}, \kappa^2_{\max}]$, the κ-estimation error propagates linearly into the index and difference estimates at each round. (3) **Exploration Overhead:** the explicit cost incurred by the reference-fidelity guard. Because the guard mandates at most $\mathcal{O}(\sqrt{T})$ forced probes at α° for each of the K arms, and each probe carries a bounded cost $c(\alpha^\circ)$, this overhead contributes a strictly lower-order additive penalty of $\mathcal{O}(K\sqrt{T})$.

Crucially, the forced exploration ensures that the sample size grows as $N^{(\circ)}_{t,i} = \Omega(\sqrt{t})$. Thus, in Lemma K.1, $B^{\kappa}_{t,i} = \tilde{\mathcal{O}}(t^{-1/4})$. Integrating this over T yields $\sum_{t=1}^{T} \tilde{\mathcal{O}}(t^{-1/4}) = \tilde{\mathcal{O}}(T^{3/4})$, thereby bounding the cumulative estimation penalty. Summing the three components confirms that the $\tilde{\mathcal{O}}(T^{3/4})$ term governs the total regret. Thus, learning κ_i^2 online via reference-fidelity anchoring rigorously preserves the algorithm's asymptotic efficiency. □

L Supplementary for Section. 7

L.1 Experimental setup details

We evaluate `DialBandit` in semi-empirical routing environments calibrated from offline LLM traces on AlpacaEval and MMLU. Our code can be found in the supplementary material.

Arms (candidate models). The candidate set comprises seven open-weight models (arms 1–7): Mistral-7B, Qwen2.5-7B, Qwen2.5-1.5B, Phi-3.5, Gemma-2-9B, Llama3.1-8B, and Qwen2.5-3B. Unless otherwise specified, we use a fixed prompt template per benchmark. For AlpacaEval, we generate one candidate response per model-query pair offline and keep that response fixed across repeated judge evaluations to isolate evaluation noise from generation variance. For MMLU, the underlying per-query quality signal is the model's next-token logit on the correct answer. Figure 11 plots the offline arm cost-quality landscape as a static summary of the candidate pool.

Offline calibration. The oracle environments treat arm-specific profile parameters as known and calibrate them from offline traces. For each arm i, we estimate μ_i by pooling all available scores. On benchmarks with repeated evaluations per query, we estimate a reference evaluation-noise level by randomly pairing repeated scores within the same query and averaging $(y_a - y_b)^2/2$; aggregating these query-level estimates gives σ_i^2. Let $\bar{y}_q$ denote the within-query average score. We then estimate cross-query heterogeneity by $\kappa_i^2 = \mathrm{Var}_q(\bar{y}_q) - \sigma_i^2/n_{\mathrm{ref}}$, where $n_{\mathrm{ref}} = 16$ is a reference repeat count used to map evaluation noise to the query-mean scale. Table 1 summarizes the resulting arm-level offline calibration statistics. The table shows that both average quality and offline heterogeneity vary substantially across arms; for AlpacaEval, the reference evaluation-noise level is also highly model-dependent, supporting our arm-specific reliability profile assumption, while for MMLU the larger κ_i^2 values are naturally interpreted as cross-question difficulty heterogeneity in the underlying logit signal.

Round protocol & action space. In each round t, the environment samples a query uniformly from the benchmark pool. The policy executes a sequential search over unprobed arms, selecting an arm i and a fidelity level α for each probe, and stops adaptively. To ensure within-round planning remains a tractable one-dimensional optimization per candidate, we discretize the fidelity space into a finite grid

$\mathcal{A} = \{1/16, 2/16, \ldots, 1\}$. We set $\lambda = 54$ by default and sweep the quality-to-cost scaling factor over $\lambda \in \{34, 38, 42, 46, 50, 54\}$.

AlpacaEval (repeated judging). For a chosen model i and query q, we use Prometheus-7B-v2.0 [34] to repeatedly score the fixed offline-generated response. A probe at fidelity $\alpha = n/n_{\max}$ executes n independent judge evaluations and returns their empirical average. Larger n reduces observation noise at the expense of higher evaluation compute. We set $n \in \mathcal{A}$ with $n_{\max} = 16$. To match the problem formulation, we decompose the probing cost as $c_i(\alpha) = c_i^{\mathrm{g}} + c^{\mathrm{e}}(\alpha)$, where c_i^{g} is the arm-specific generation cost given by the relative-cost profile as shown in Figure 11, and the shared evaluation-side component is a linear function of α, *e.g.,* $c^{\mathrm{e}}(\alpha) = base_eval_cost + eval_cost \cdot \alpha$ with $base_eval_cost = 0.2$ and $eval_cost = 2$ in our setting.

MMLU (calibrated noisy scores). Let $x_{q,i} \in [0, 1]$ denote the calibrated latent quality of model i on query q, obtained from the correct-answer logit. During online evaluation, a probe at fidelity α returns a noisy observation centered on $x_{q,i}$ with variance $\sigma_i^2(\alpha)$. For normalized fidelity $\alpha = n/n_{\max}$ with $n \in \{1, \ldots, n_{\max}\}$, we set $\sigma_i^2(\alpha) = \frac{\mu_i(1-\mu_i)}{(\alpha n_{\max})^\beta}$, using a shared shape parameter $\beta > 0$. The resulting power-law curve is monotone decreasing and captures diminishing returns from higher evaluation fidelity without changing the underlying per-query latent quality.

Metrics. We evaluate the per-round net utility using the Bayes-consistent posterior score: $U_t = \max_{j \in \mathcal{O}_t} S_j(y_{t,j}, \alpha_{t,j}) - \sum_{i \in \mathcal{O}_t} c_i(n_{t,i})$. Regret is measured against the per-round oracle operating under the exact same observation model. The primary online-learning metrics (net utility, regret, cost, score) are averaged over 50 random seeds with standard error, using a horizon of $T = 1000$.

Baselines & ablations. All baseline methods share the identical arm set, fidelity grid, and cost functions. The ablation variants isolate specific components of our design:

- `MeanOnly`: Omits noise modeling, relying solely on a mean-only index.
- `FixFd`: Fixes $n = 16$ for all probes, disabling adaptive fidelity control.
- `RndFd`: Randomly samples fidelity for each probe instead of active selection.

Compute resources. The reported online-learning experiments are semi-empirical simulations over cached LLM traces and calibrated observation models; reproducing the routing curves from these traces does not require model training or online LLM inference. The simulation server has an Intel Core i9-10980XE processor with a 3.00 GHz base frequency, 18 physical cores (36 hardware threads), and 128 GiB of system memory. The CPU-based simulation workload is dominated by $T = 1000$ rounds, $K = 7$ arms, a 16-point fidelity grid, and GP updates over the one-dimensional fidelity space. Regenerating the AlpacaEval traces requires a one-time offline pass that generates candidate responses with the seven LLM arms and repeatedly scores them with Prometheus-7B-v2.0; we used an NVIDIA GeForce RTX 3090 GPU with 24 GiB of device memory for this model-inference step. The offline cost scales with the number of candidate-query pairs and judge repetitions, but it is separate from reproducing the reported online routing comparisons.

Existing assets and licenses. External datasets, model artifacts, and evaluators are credited in the main paper and appendix. The experiments use AlpacaEval (CC-BY-NC-4.0) and MMLU (MIT). The motivating experiment also uses SummEval, whose repository is MIT-licensed; its source CNN/DailyMail news articles remain subject to their upstream terms. Prometheus-7B-v2.0, Mistral-7B-Instruct-v0.3, Qwen2.5-1.5B, Qwen2.5-7B, and Qwen3-30B use Apache-2.0. Qwen2.5-3B uses the Qwen Research License, Gemma-2-9B uses the Gemma Terms of Use, Llama3.1-8B uses the Llama 3.1 Community License, and Phi-3.5 uses MIT. GPT-3.5 Turbo is accessed through an API governed by the applicable OpenAI service terms and usage policies. We use these assets only for research evaluation and follow their respective licenses and terms of use.

L.2 Supplementary experiment results

Cost-quality landscape. Figure 11 visualizes the offline arm cost-quality landscape. In both benchmarks, lower-cost arms are not uniformly dominated, and the most attractive arm depends on the quality-cost preference. This rules out a universally best fixed model and provides a static backdrop for adaptive routing.

Supplementary evidence for adaptive fidelity and stopping (Q2). Figure 12 shows that on MMLU, `DialBandit` still concentrates most probability mass on low fidelities, matching the adaptive-fidelity

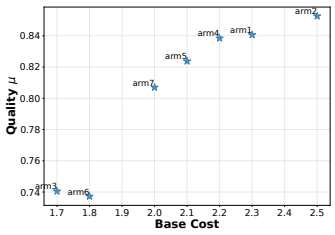
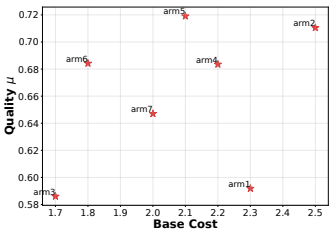


Figure 11: Arm cost-quality landscape on AlpacaEval and MMLU.



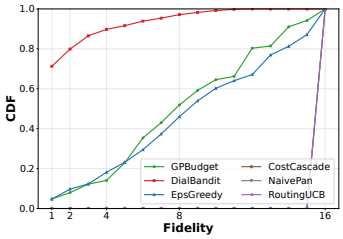


Figure 12: Adaptive fidelity on MMLU.

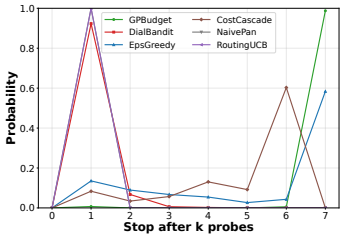
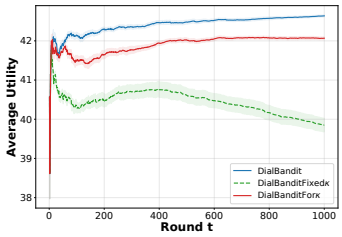


Figure 13: Stopping behavior on AlpacaEval.



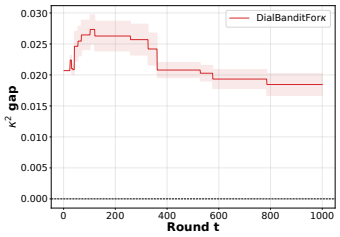


Figure 14: Online learning for κ on AlpacaEval.

trend reported in the main text. Figure 13 shows that on AlpacaEval, the policy usually stops after one or two probes, consistent with efficient early stopping in an easier environment.

Online κ learning extension. Figure 14 evaluates the fully online variant that learns cross-query heterogeneity κ_i^2 on-the-fly. The left panel shows that this variant achieves a 5.6% utility improvement over a mismatched offline profile, closing the gap to the known-κ oracle to within 1.25%. The right panel confirms that the online κ trajectory tightly tracks the true parameter despite sparse updates.

L.3 Empirical Prometheus residuals

The AlpacaEval replay samples cached Prometheus judgments of fixed query–arm responses. To inspect the empirical observation distribution, we subtract each response's empirical mean from its repeated judgments and pool the standardized within-response residuals. Table 2 shows skew and tails relative to a Gaussian reference. The observed replay thus evaluates the policy on discrete empirical judge draws rather than Gaussian samples from the analytical observation model.

Table 2: Distribution of response-centered Prometheus judgments used in the AlpacaEval replay. The Gaussian tail probability is shown for reference.

Diagnostic	Value
Fixed query–arm responses	5,635
Raw Prometheus judgments	112,700
Standardized residual skew	-0.701
Standardized residual excess kurtosis	1.183
Empirical $\Pr(\lvert z\rvert > 3)$	0.0094
Gaussian $\Pr(\lvert z\rvert > 3)$	0.0027

L.4 Sensitivity to the GP confidence scale

We vary the policy's GP confidence scale one factor at a time at $\lambda = 54$. Over $\beta \in \{0.5, 1, 2\}$, average reward changes by 0.0% to $+1.1\%$ relative to the default $\beta = 1$. This check isolates the effect of confidence scaling in the planning rule; the online-κ experiment in Figure 14 separately evaluates learning the arm heterogeneity profile.

L.5 Decision-layer computation and memory

For a discrete fidelity set $\mathcal{A}$, evaluating candidate indices across arms and fidelity levels entails $O(K|\mathcal{A}|)$ value evaluations per planning state. Exchange checks compare arm pairs over the baseline grid $\mathcal{Z}$, requiring $O(K^2|\mathcal{Z}|)$ comparisons; topological ordering and suffix planning then operate on the resulting K-arm cascade. These operations describe policy-side computation separately from model generation and evaluator calls.

Let M_i be the number of noise labels collected for arm i. Storing the raw per-arm labels requires $O(\sum_i M_i)$ memory, while cached factors for exact GP regression require $O(\sum_i M_i^2)$ memory. With the square-time replication schedule, $M_i = O(\sqrt{T})$, giving $O(K\sqrt{T})$ raw-label storage and up to $O(KT)$ exact-factor storage. These bounds describe the noise learner; they do not include the offline response and judgment bank used by replay.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main contributions: the tunable-fidelity sequential-search formulation, the oracle index policy, `DialBandit`, the regret guarantee, and semi-empirical evaluations. The scope and assumptions are further specified in the problem formulation, theoretical analysis, and discussion of limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses modeling and deployment limitations in the theoretical discussion and Appendix I.7, including sparse noise updates, replicate-based noise labels, offline calibration of heterogeneity, and extensions to non-stationary settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The assumptions are stated in the problem formulation and Section 6. Proofs of the oracle policy, concentration, planning stability, and regret bound are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Section 7 and Appendix L.1 describe the datasets, candidate models, observation models, fidelity grid, cost decomposition, baselines, metrics, horizon, and number of random seeds used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The experiments use publicly available benchmarks and open-weight models cited in the paper. The implementation code is provided in the supplementary material for reproducing the reported semi-empirical experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Section 7 gives the main experimental protocol, and Appendix L.1 provides additional details on offline calibration, arm profiles, fidelity levels, cost functions, baselines, ablations, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Appendix L.1 states that the primary online-learning metrics (net utility, regret, cost, and score) are averaged over 50 random seeds with standard error, capturing variability across random online interaction trajectories.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix L.1 describes the compute profile of the semi-empirical simulations, clarifying that reproducing the online routing curves from cached traces uses a standard CPU Python environment, while regenerating the offline LLM traces requires separate GPU inference resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work studies algorithmic routing and evaluation efficiency using public datasets and model outputs, without collecting private data or involving human subjects. We are not aware of deviations from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The broader-impact discussion in Appendix A notes potential benefits from reduced evaluation compute, latency, and energy use, as well as risks from confidently optimizing biased evaluators, and recommends calibration and bias-audit safeguards for deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release a new pre-trained model, image generator, scraped dataset, or other high-risk asset. The released material is limited to code for reproducing the reported algorithmic experiments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The paper cites the datasets, model artifacts, and evaluator used in the experiments, and Appendix L.1 lists their licenses or applicable terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: The paper does not introduce a new dataset or model, but it provides code for reproducing the algorithmic experiments in the supplementary material. The experimental protocol and semi-empirical environment are documented in Appendix L.1.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing experiments or new research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or new research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: LLMs are central to the experimental setting: candidate arms are open-weight LLMs, and AlpacaEval responses are scored with an LLM judge as described in Section 7 and Appendix L.1.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.