ONLINE LEARNING WITH IMPROVING PROXIES: A MULTI-FIDELITY BANDIT APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

In many online learning problems, reliable evaluation can be costly, motivating inexpensive proxy feedback that approximates a trusted target. While proxy quality is often treated as static, learned evaluators may improve through continued training, self-generated supervision, or computational refinement, creating a trade-off between further proxy investment and trusted feedback. Motivated by such self-evolving evaluators, including LLM-based judges, we study a multi-fidelity multi-armed bandit with a stationary high-fidelity target and a low-fidelity source whose discrepancy evolves with use. We introduce an a priori bound on the average proxy-target mismatch, derive improvement-aware confidence bounds, and propose the Threshold-Based Adaptive Continuation Companion (TACC), which determines whether to continue low-fidelity evaluation or escalate to high fidelity. We establish an instance-dependent regret bound showing that limited low-fidelity continuation eliminates the logarithmic high-fidelity term for a class of suboptimal arms, and a lower bound capturing the cost of proxy improvement. Experiments on synthetic bandits and NAS-Bench-201 evaluate when adaptive continuation improves performance.

1 INTRODUCTION

Reliable evaluation is essential for selecting models, comparing policies, and guiding sequential decisions (Sutton et al., 1998; Raschka, 2018). We refer to a trusted source that defines the evaluation target, such as expert assessment, gold-standard verification, high-resolution simulation, or fully trained model performance, as a *high-fidelity evaluator*. Because obtaining feedback from such sources can be expensive, many evaluation pipelines rely on a *low-fidelity evaluator*: a cheaper proxy whose feedback approximates, but may systematically differ from, the high-fidelity target. Examples include large language model (LLM)-based evaluators, surrogate models, approximate simulators, and partial-training measurements (Li et al., 2018; Peherstorfer et al., 2018). Importantly, this discrepancy need not remain fixed. A low-fidelity evaluator may improve through continued training, self-generated supervision, or additional computational refinement without requiring new high-fidelity feedback at each step. We refer to such a proxy as a *self-evolving evaluator*. Allocating resources to such an evaluator can therefore provide immediate evaluation feedback while also improving subsequent evaluations. This dual benefit creates a resource-allocation problem: when should one invest further in a self-evolving low-fidelity evaluator, and when should one pay for high-fidelity feedback?

LLM-based evaluation provides a concrete example. When human judgments or expert assessments serve as the high-fidelity target for comparing candidate models or policies, an LLM can act as a low-fidelity evaluator by scoring candidate outputs or comparing pairs of responses. These automated judgments are inexpensive but may systematically disagree with trusted assessments (Zheng et al., 2023; Huang et al., 2025). The LLM evaluator can become self-evolving when evaluation is coupled with a refinement process that uses *self-generated supervision*, meaning training signals produced by the evaluator or its associated generation pipeline rather than newly collected trusted labels (Yuan et al., 2024; Wang et al., 2024). For example, the pipeline can generate response pairs, produce synthetic judgments indicating which response is preferable, filter these judgments using consistency checks, and use the retained comparisons to update the evaluator. Additional resources devoted to this process can thus support both candidate evaluation and improvements in subsequent judgments. In the setting we study, low-fidelity queries are coupled with this refinement process, allowing fidelity

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

discrepancy to change with accumulated low-fidelity usage and making anticipated improvement relevant to the choice between continued low-fidelity evaluation and high-fidelity assessment.

Multi-fidelity multi-armed bandits (MF-MAB) provide a natural framework for this allocation problem. Each *arm* represents a candidate model, policy, or decision, and each query selects both an arm and an evaluation fidelity. The high-fidelity mean reward measures the quality of an arm under the trusted evaluation target, while the selected fidelity determines the cost and feedback available to the learner. The objective is to maximize cumulative reward measured against the high-fidelity target under a limited evaluation budget. Standard MF-MAB formulations characterize fidelity discrepancy using a fixed bias bound (Kandasamy et al., 2016; Wang et al., 2023). Although such a bound can conservatively describe an improving evaluator, it does not explicitly exploit the reduction in discrepancy enabled by further low-fidelity investment.

Accounting for this improvement introduces a distinct exploration challenge: additional low-fidelity queries can reduce both *statistical uncertainty* from noisy observations and fidelity discrepancy from imperfect proxy–target agreement, potentially at different rates. A static rule that escalates once statistical uncertainty falls below a threshold may therefore switch too early, when a few additional low-fidelity queries could reduce the discrepancy enough to certify suboptimality; conversely, continued proxy evaluation can waste budget when the discrepancy remains too large or improves too slowly. The learner must distinguish arms that benefit from further evaluator refinement from those that require trusted feedback.

We study this challenge through a canonical two-fidelity bandit model with stationary high-fidelity mean rewards and a low-fidelity evaluator whose discrepancy depends on the number of low-fidelity queries. To describe the information provided by an evolving evaluator, we introduce an a priori *selected-average mismatch bound*: a conservative bound on the difference between an arm's high-fidelity mean and the average conditional means of its selected low-fidelity observations. Controlling this average is essential because accumulated observations come from different stages of evaluator refinement; the accuracy of the current evaluator alone does not characterize earlier feedback. We refer to the supplied bound as a *mismatch certificate* and combine it with statistical concentration to construct confidence bounds for the high-fidelity target.

The mismatch certificate is an explicit assumption of our framework. Before deployment, historical refinement runs or separate benchmark data with trusted evaluations can be used to estimate how fidelity discrepancy changes with low-fidelity investment (Wang et al., 2024; Kalra & Tang, 2025; Dubois et al., 2023). Under suitable calibration assumptions, these estimates can support conservative certificates across refinement budgets (Sheng et al., 2025). We treat the certificate as fixed prior information during online learning, and our theoretical guarantees require it to hold for the online evaluation process. A low-fidelity query reveals only low-fidelity feedback; high-fidelity feedback is observed only when the learner explicitly selects and pays for a high-fidelity query.

Building on this formulation, we propose the *Threshold-Based Adaptive Continuation Companion* (TACC), which combines optimistic arm selection with bounded low-fidelity continuation. Once low-fidelity statistical uncertainty crosses a prescribed threshold, TACC uses the anticipated reduction in mismatch correction to decide whether additional low-fidelity queries remain worthwhile or high-fidelity evaluation is needed. This mechanism targets an intermediate regime where the current proxy is insufficient to establish suboptimality, but limited additional refinement can make it sufficient.

We analyze TACC using *cost-weighted regret*, which measures the loss relative to allocating the evaluation budget to an optimal arm, accounting for the costs of the selected fidelities. Our instance-dependent bound shows that, for arms satisfying the continuation and certification conditions, bounded low-fidelity sampling can replace a high-fidelity confirmation cost that grows logarithmically with the budget. The result identifies when anticipated evaluator improvement can produce a concrete reduction in expensive trusted evaluations.

Contributions. Our contributions are as follows.

1. **A framework for self-evolving evaluators.** We formulate a two-fidelity bandit model with usage-dependent discrepancy and an a priori selected-average mismatch certificate, yielding improvement-aware confidence bounds for the high-fidelity target (Section 3).

2. **Adaptive allocation through bounded continuation.** We propose TACC, which combines optimistic arm selection with a bounded post-threshold continuation rule to exploit anticipated evaluator improvement when choosing between low- and high-fidelity queries (Section 4).
3. **Instance-dependent regret guarantees.** We establish a regret bound that characterizes when limited additional low-fidelity evaluation can replace logarithmic high-fidelity confirmation, quantifying the benefit of exploiting an improving evaluator (Section 5).
4. **Evaluation across controlled and empirical settings.** We evaluate TACC on synthetic bandits, a controlled LLM-as-a-judge task, and NAS-Bench-201, including experiments with empirical learning trajectories and offline-calibrated mismatch certificates (Section 6).

2 Related Work

Multi-fidelity and evolving-source bandits. Our work builds on stochastic multi-armed bandits and multi-fidelity extensions (Lattimore & Szepesvári, 2020; Kandasamy et al., 2016; Poiani et al., 2022; Wang et al., 2023; Poiani et al., 2024). Standard multi-fidelity methods manage cost–accuracy trade-offs using statistical uncertainty together with a fixed or uniformly bounded fidelity bias; related ideas also appear in multi-fidelity optimization and Hyperband (Kandasamy et al., 2019; Li et al., 2018). Our setting instead allows the low-fidelity discrepancy to evolve with usage, connecting to structured non-stationary bandits with pull-dependent trajectories and improving arms (Seznec et al., 2019; Metelli et al., 2022; Mussi et al., 2023; Cella et al., 2021; Latypov et al., 2025). The key distinction is that the high-fidelity target remains stationary while only the low-fidelity proxy evolves.

Self-improving LLM evaluators and proxy–target mismatch. LLMs are increasingly used as evaluators or reward providers, with recent work showing that evaluation quality can improve through self-generated supervision or additional inference-time computation (Yuan et al., 2024; Wang et al., 2024; Kalra & Tang, 2025). Learned evaluators may nevertheless retain systematic disagreement with human preferences or trusted criteria (Zheng et al., 2023; Gao et al., 2023; Huang et al., 2025), and their reliability can be assessed using human-labeled or verifiable benchmarks and calibration procedures (Dubois et al., 2023; Lambert et al., 2025; Tan et al., 2025; Sheng et al., 2025). These works study evaluator improvement or proxy–target agreement; we study sequential fidelity allocation with an improving proxy, a costly stationary target, and an a priori mismatch certificate.

3 Problem Setup

We study a budget-constrained stochastic multi-armed bandit problem featuring a cheaper, improving low-fidelity source (L) and a target, high-fidelity source (H).

There are K arms indexed by $k \in [K]$. Arm k has high-fidelity mean $\mu_k^{(H)}$, with a unique optimal arm $k^\star \in \arg\max_{k\in[K]} \mu_k^{(H)}$, target value $\mu^\star := \mu_{k^\star}^{(H)}$, and suboptimality gap $\Delta_k := \mu^\star - \mu_k^{(H)}$ for $k \neq k^\star$. Querying fidelity $m \in \{L, H\}$ incurs cost $\lambda^{(m)}$, with $0 < \lambda^{(L)} < \lambda^{(H)}$. Given budget $\Lambda > 0$, let $T_\Lambda := \left\lceil \frac{\Lambda}{\lambda^{(L)}} \right\rceil$, which upper bounds the total number of queries.

At round t, the learner selects an arm-fidelity pair $(I_t, m_t) \in [K] \times \{L, H\}$, observes a noisy sample, and pays cost $\lambda^{(m_t)}$. Let $C_t := \sum_{s=1}^{t} \lambda^{(m_s)}$ be the cumulative cost after t rounds. We measure performance by the cost-weighted pseudo-regret

$$R(\Lambda) := \sum_{t:C_t \le \Lambda} \lambda^{(m_t)} \big(\mu^\star - \mu_{I_t}^{(H)}\big). \tag{1}$$

This follows the standard multi-fidelity bandit setup (Kandasamy et al., 2016), although Wang et al. (2023) use a different regret definition. The u-th high-fidelity query of arm k yields $Y_{k,u}^{(H)} = \mu_k^{(H)} + \eta_{k,u}^{(H)}$, whereas its τ-th low-fidelity query yields $Y_{k,\tau}^{(L)} = \mu_k^{(L)}(\tau) + \eta_{k,\tau}^{(L)}$. Conditional on the past and the selected arm–fidelity pair, each observation noise is zero-mean and 1-sub-Gaussian.

The key modeling choice is that the low-fidelity response is indexed by arm-specific usage. Analogous to rested bandit models (Cella et al., 2021), the low-fidelity state of arm k changes only when that arm is queried. Thus, a low-fidelity query may affect subsequent low-fidelity observations without

requiring an accompanying high-fidelity observation. We deliberately exclude global-time drift and cross-arm coupling to isolate how localized proxy evolution influences fidelity choice.

For $n \geq 1$, define the selected-average low-fidelity mean $\bar{\mu}_k^{(L)}(n) := \frac{1}{n}\sum_{\tau=1}^{n} \mu_k^{(L)}(\tau)$ (Latypov et al., 2025), which is the quantity estimated by the low-fidelity empirical mean. We assume access to an a priori conservative certificate on the low-to-high discrepancy, obtained before online learning and fixed throughout the bandit run.

Assumption 3.1 (Evolving low-fidelity mismatch certificate)**.** *For each arm k, there exists a known nondecreasing function $U_k(\cdot) : \mathbb{N} \to \mathbb{R}_+$ such that, for every $n \geq 1$,*

$$\sum_{\tau=1}^{n} \left| \mu_k^{(L)}(\tau) - \mu_k^{(H)} \right| \leq U_k(n). \tag{2}$$

Define the associated selected-average mismatch bound by $B_k(n) := \sup_{s \geq n} U_k(s)/s$ for $n \geq 1$, and assume $B_k(1) < \infty$. Then

$$\left| \bar{\mu}_k^{(L)}(n) - \mu_k^{(H)} \right| \leq \frac{U_k(n)}{n} \leq B_k(n). \tag{3}$$

Here, a priori upper certificate $U_k(n)$ bounds cumulative mismatch, while $B_k(n)$ gives the corresponding nonincreasing bound for the selected-average low-fidelity mean. The theory permits arm-specific certificates, with a common certificate as a special case.

Remark 3.2 (Relation to static multi-fidelity models). Standard multi-fidelity bandits assume a fixed bias bound such as $|\mu_k^{(L)} - \mu_k^{(H)}| \leq \zeta_k$ (Kandasamy et al., 2016; Wang et al., 2023). Assumption 3.1 replaces this with the count-dependent bound $B_k(n)$ on $|\bar{\mu}_k^{(L)}(n) - \mu_k^{(H)}|$, which serves as the mismatch correction for low-fidelity confidence bounds.

Given $B_k(n)$, the effective low-fidelity gap for each suboptimal arm $k \neq k^\star$ can be defined by

$$\Delta_k^{(L)}(n) := \mu^\star - \bar{\mu}_k^{(L)}(n) - B_k(n). \tag{4}$$

If $\Delta_k^{(L)}(n) > 0$, then the low-fidelity upper proxy $\bar{\mu}_k^{(L)}(n) + B_k(n)$ lies below $\mu^\star$, so arm k can in principle be ruled out using low-fidelity information.

Building on the mismatch bound $B_k(n)$, we introduce *Threshold-Based Adaptive Continuation Companion* (TACC), a confidence-based policy in the next section to decide whether low-fidelity sampling remains worthwhile before switching to high fidelity.

4 Algorithm

Our algorithm follows a confidence-based template: maintain upper bounds on the high-fidelity target, select an optimistic arm, and choose its sampling fidelity. Unlike a static threshold rule, TACC allows bounded low-fidelity continuation after the statistical uncertainty crosses the switching threshold when the anticipated decrease in the mismatch correction remains worthwhile.

4.1 Confidence bounds

Both fidelities yield confidence intervals for the high-fidelity target $\mu_k^{(H)}$. High fidelity uses a standard interval around its empirical mean, while low fidelity adds the a priori mismatch correction $B_k(n)$ to its empirical-mean interval to ensure validity for $\mu_k^{(H)}$.

For total budget Λ, define the budget-uniform confidence radius $G_\Lambda(n, \delta) := \sqrt{\frac{2\log(4KT_\Lambda/\delta)}{n}}$, $n \geq 1$. For a fixed arm, fidelity, and sample count n, the empirical mean of 1-sub-Gaussian observations satisfies

$$\Pr(|\widehat{\mu}_n - \mathbb{E}[\widehat{\mu}_n]| > G_\Lambda(n, \delta)) \leq 2\exp\left(-\frac{nG_\Lambda^2(n, \delta)}{2}\right) = \frac{\delta}{2KT_\Lambda}.$$

A union bound over the K arms, two fidelities, and all sample counts $1 \leq n \leq T_\Lambda$ therefore yields a single event of probability at least $1 - \delta$ on which all empirical-mean deviations are simultaneously

controlled. The use of the deterministic budget horizon T_Λ is convenient in our fixed-budget setting and yields a fixed low-fidelity switching count.

Let $N_{k,t}^{(L)}$ and $N_{k,t}^{(H)}$ denote the numbers of low- and high-fidelity queries of arm k before round t, with corresponding empirical means $\hat{\mu}_{k,t}^{(L)}$ and $\hat{\mu}_{k,t}^{(H)}$. The confidence radii are defined as:

$$\mathrm{rad}_k^{(L)}(n,\delta) := G_\Lambda(n,\delta) + B_k(n), \quad \mathrm{rad}^{(H)}(n,\delta) := G_\Lambda(n,\delta).$$

The corresponding confidence bounds from fidelity $m \in \{L, H\}$ for the target mean $\mu_k^{(H)}$ are[1]

$$\mathrm{UCB}_{k,t}^{(m)} = \hat{\mu}_{k,t}^{(m)} + \mathrm{rad}_k^{(m)}(N_{k,t}^{(m)},\delta), \quad \mathrm{LCB}_{k,t}^{(m)} = \hat{\mu}_{k,t}^{(m)} - \mathrm{rad}_k^{(m)}(N_{k,t}^{(m)},\delta). \tag{5}$$

Both fidelities yield valid intervals for the target quantity, with low fidelity carrying the count-dependent mismatch correction $B_k(n)$ rather than a fixed fidelity bias (Wang et al., 2023).

4.2 Arm and fidelity selection

The decision rule has two components: an arm-selection step based on the aggregate confidence score, followed by a fidelity-selection step for the chosen arm.

Arm selection. Given the two target intervals, we select the arm that remains most plausible under the high-fidelity objective. Since either fidelity may rule out an arm as suboptimal, we define the aggregate optimistic score as the smaller of the two fidelity-specific upper bounds:

$$\mathrm{UCB}_{k,t} := \min_{m\in\{L,H\}} \mathrm{UCB}_{k,t}^{(m)}, \qquad I_t \in \arg\max_{k\in[K]} \mathrm{UCB}_{k,t}. \tag{6}$$

Thus, an arm remains competitive only if neither fidelity-specific upper bound rules it out.

Fidelity selection with bounded continuation. After selecting I_t, we decide which fidelity to query next for that arm. If $G_\Lambda(N_{I_t,t}^{(L)},\delta) \geq \gamma$, the low-fidelity estimate is still statistically too coarse, so we continue querying low fidelity.

The more delicate case occurs after this radius falls below γ. A static threshold rule would now escalate to high fidelity. In contrast, we ask whether a few additional low-fidelity queries could still reduce the mismatch correction enough to be worthwhile. For the selected arm I_t, define the continuation gain

$$V_{I_t,t}(s) := B_{I_t}\left(N_{I_t,t}^{(L)}\right) - B_{I_t}\left(N_{I_t,t}^{(L)} + s\right), \qquad 1 \leq s \leq S_0. \tag{7}$$

The continuation gain is computed online from the pre-specified certificate B_k and the current low-fidelity count via equation 7. The certificates are supplied before online learning and remain fixed throughout the bandit run.

Let A_k denote the number of continuation pulls already spent on arm k after its low-fidelity radius first drops below γ. The fidelity-selection rule is

$$m_t = \begin{cases} L, & \text{if } G_\Lambda\left(N_{I_t,t}^{(L)},\delta\right) \geq \gamma \text{ or } \left[A_{I_t} < S_0,\ \exists\, 1 \leq s \leq S_0 - A_{I_t} : V_{I_t,t}(s) \geq 2\eta\gamma\right], \\ H, & \text{otherwise.} \end{cases} \tag{8}$$

Here, $\eta \in (0,1)$ controls how readily TACC continues after the threshold crossing, while $S_0\lambda^{(L)} \leq \lambda^{(H)}$ ensures that a full continuation block costs no more than one high-fidelity query. The threshold $2\eta\gamma$ is an η-fraction of the certification scale 2γ formalized in Section 5, allowing continuation only when the prescribed decrease in the mismatch correction is at least $2\eta\gamma$. Following prior multi-fidelity practice (Kandasamy et al., 2016), γ may be chosen using a cost-balanced scaling before the run.

In each subsequent round, Algorithm 1 constructs improvement-aware confidence intervals, selects an arm using the aggregate optimistic index and queries low fidelity while its statistical radius exceeds γ. After the threshold crossing, TACC continues at low fidelity only when the anticipated decrease in $B_k(n)$ justifies bounded continuation; otherwise, it escalates to high fidelity.

[1]For high fidelity, $\mathrm{rad}_k^{(H)}(n,\delta)$ is identical to $\mathrm{rad}^{(H)}$ here.

Algorithm 1 Threshold-Based Adaptive Continuation Companion (TACC)

Input: budget Λ, arm set $[K]$, costs $\lambda^{(L)}, \lambda^{(H)}$, confidence level δ, threshold γ, continuation parameter η, continuation cap S_0, mismatch bound $B_k(\cdot)$
Initialize $\hat{\mu}_k^{(L)}, \hat{\mu}_k^{(H)}, N_{k,t}^{(L)}, N_{k,t}^{(H)}, C$ and $A_k \leftarrow 0$
while $C \leq \Lambda$ **do**
    Construct $\mathrm{UCB}_{k,t}^{(L)}, \mathrm{UCB}_{k,t}^{(H)}$ using equation 5
    Compute $\mathrm{UCB}_{k,t}$ with equation 6 and select $I_t \in \arg\max_{k\in[K]} \mathrm{UCB}_{k,t}$
    if $G_\Lambda(N_{I_t,t}^{(L)}, \delta) \geq \gamma$ **then**
        Set $m_t = L$
    else if $A_{I_t} < S_0$ and there exists $1 \leq s \leq S_0 - A_{I_t}$ such that $V_{I_t,t}(s) \geq 2\eta\gamma$ **then**
        Set $m_t = L$ and update $A_{I_t} \leftarrow A_{I_t} + 1$
    else
        Set $m_t = H$
    if $C + \lambda^{(m_t)} > \Lambda$ **then**
        break
    Query arm I_t at fidelity m_t, observe reward, and update empirical means and query counts
    Update cumulative cost $C \leftarrow C + \lambda^{(m_t)}$

5 Theoretical Results

We analyze Algorithm 1 on the budget-uniform event $\mathcal{E}_\Lambda$ from Lemma B.1, which holds with probability at least $1-\delta$. Let $\ell_\Lambda(\delta) := 2\log\big(\frac{4KT_\Lambda}{\delta}\big)$, $G_\Lambda(n,\delta) := \sqrt{\frac{\ell_\Lambda(\delta)}{n}}$, and define the nominal low-fidelity switching count $N_\gamma(\delta) := \min\{n \geq 1 : G_\Lambda(n,\delta) < \gamma\} = \left\lfloor \frac{\ell_\Lambda(\delta)}{\gamma^2} \right\rfloor + 1$. Thus $N_\gamma(\delta) = O\Big(\frac{\log(4KT_\Lambda/\delta)}{\gamma^2}\Big)$. The logarithmic factor follows directly from a union bound over arms, fidelities, and sample counts up to the budget horizon.

For each suboptimal arm $k \neq k^\star$, define the low-fidelity certification time

$$\tau_k(\gamma) := \inf\Big\{n \geq 1 : \Delta_k^{(L)}(n) \geq 2\gamma\Big\}, \tag{9}$$

with $\tau_k(\gamma) = \infty$ if the set is empty. We partition suboptimal arms according to whether low-fidelity certification occurs before, within, or after the continuation window:

$$\begin{aligned} \mathcal{A}_\gamma(\delta) &:= \{k \neq k^\star : \tau_k(\gamma) \leq N_\gamma(\delta)\}, \\ \mathcal{B}_\gamma(\delta) &:= \{k \neq k^\star : \tau_k(\gamma) > N_\gamma(\delta) + S_0\}, \\ \mathcal{C}_\gamma(\delta) &:= \{k \neq k^\star : N_\gamma(\delta) < \tau_k(\gamma) \leq N_\gamma(\delta) + S_0\}. \end{aligned} \tag{10}$$

Arms in $\mathcal{A}_\gamma(\delta)$ are certifiable by the nominal switching threshold, arms in $\mathcal{B}_\gamma(\delta)$ are not certifiable within the continuation window, and $\mathcal{C}_\gamma(\delta)$ contains the intermediate arms that become certifiable within at most S_0 additional low-fidelity queries.

For arms in $\mathcal{C}_\gamma(\delta)$, low-fidelity certification is reachable within the continuation window, but TACC must also continue sampling until the certification count is reached. We therefore define

$$\mathcal{C}_\gamma^{\mathrm{d}}(\delta) := \left\{ k \in \mathcal{C}_\gamma(\delta) : \begin{array}{l} \text{for every } N_\gamma(\delta) \leq n < \tau_k(\gamma), \\ \exists\, 1 \leq s \leq N_\gamma(\delta) + S_0 - n \text{ such that } B_k(n) - B_k(n+s) \geq 2\eta\gamma \end{array} \right\}. \tag{11}$$

Thus, $\mathcal{C}_\gamma(\delta)$ captures certificate reachability, while $\mathcal{C}_\gamma^{\mathrm{d}}(\delta)$ contains arms for which TACC's continuation test remains satisfied until certification. Let

$$\mathcal{C}_\gamma^{\mathrm{u}}(\delta) := \mathcal{C}_\gamma(\delta) \setminus \mathcal{C}_\gamma^{\mathrm{d}}(\delta), \qquad \mathcal{H}_\gamma(\delta) := \mathcal{B}_\gamma(\delta) \cup \mathcal{C}_\gamma^{\mathrm{u}}(\delta),$$

where $\mathcal{H}_\gamma(\delta)$ collects the arms that may still require logarithmic high-fidelity confirmation.

The regret analysis uses one additional regularity condition, which ensures that a low-fidelity certificate, once reached within the continuation window, does not disappear before the window ends.

Assumption 5.1 (Local low-fidelity certificate stability). *Fix $\gamma > 0$ and S_0. For each suboptimal arm $k \neq k^\star$, if $\Delta_k^{(L)}(n) \geq 2\gamma$ for some $n \leq N_\gamma(\delta) + S_0$, then $\Delta_k^{(L)}(m) \geq 2\gamma$, $\forall m \in [n, N_\gamma(\delta) + S_0]$.*

Remark 5.2. Unlike global monotonicity, concavity, or Lipschitz assumptions on the low-fidelity trajectory, Assumption 5.1 is local to the continuation window: once a certificate reaches 2γ, it must remain valid until the window ends.[2] Related local regularity conditions appear in structured evolving bandits (Heidari et al., 2016; Metelli et al., 2022; Seznec et al., 2019).

Theorem 5.3 (Regret bound for TACC). *Suppose Assumption 3.1 and Assumption 5.1 hold. Then with probability at least $1 - \delta$, Algorithm 1 satisfies*

$$R(\Lambda) \leq \sum_{k \in \mathcal{A}_\gamma(\delta)} \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + 1) + \lambda^{(H)}\right] + \sum_{k \in \mathcal{C}_\gamma^{\mathrm{d}}(\delta)} \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)}\right] + \sum_{k \in \mathcal{H}_\gamma(\delta)} \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)} \left(1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil\right)\right]. \tag{12}$$

The following corollary records the leading-order dependence on the budget and confidence level.

Corollary 5.4 (Order-wise regret bound). *Under the conditions of Theorem 5.3, on $\mathcal{E}_\Lambda$ and as $\Lambda \to \infty$ with the instance, γ, and S_0 fixed,*

$$R(\Lambda) = O\left(\lambda^{(L)} \sum_{k \neq k^\star} \frac{\Delta_k}{\gamma^2} \log \frac{2KT_\Lambda}{\delta} + \lambda^{(H)} \sum_{k \in \mathcal{H}_\gamma(\delta)} \frac{\log(2KT_\Lambda/\delta)}{\Delta_k}\right), \tag{13}$$

The low-fidelity term is paid by all suboptimal arms, whereas the logarithmic high-fidelity term is paid only by $\mathcal{H}_\gamma(\delta)$. Thus, bounded continuation removes logarithmic high-fidelity confirmation for arms in $\mathcal{C}_\gamma^{\mathrm{d}}(\delta)$. A smaller δ increases the nominal low-fidelity switching count, yielding more conservative confidence bounds but potentially certifying more arms before high-fidelity escalation. Full proof is given in Appendix B.2.

Remark 5.5 (Comparison with existing multi-fidelity rules). A static MF-UCB-style rule (Kandasamy et al., 2016) queries high fidelity if an arm is selected after $G_\Lambda(N_{k,t}^{(L)}, \delta) < \gamma$. Its upper bound retains a logarithmic high-fidelity term for $\mathcal{C}_\gamma(\delta)$; TACC's bound removes this term for $\mathcal{C}_\gamma^{\mathrm{d}}(\delta)$. This does not imply that the static rule must make such queries on every instance. If $B_k(n) \equiv b_k$, then $V_{k,t}(s) = 0$ and TACC reduces to its static-threshold counterpart. Proposition B.8 compares upper bounds, not realized regrets or all retuned static policies.

Lower bound. We complement the upper bound with a finite-budget lower bound that captures the cost of proxy improvement. For a suboptimal arm k, define

$$M_k := \sup\{m \in \mathbb{N}_0 : \Delta_k s \leq U_k(s) \text{ for all } 1 \leq s \leq m\}, \tag{14}$$

which is the largest number of low-fidelity queries for which a target shift of size Δ_k can remain consistent with the same mismatch certificate.

Theorem 5.6 (Finite-budget lower bound). *Fix $K \geq 2$ and gaps $0 < \Delta_k \leq 1/4$ for $k = 2, \ldots, K$, and let $\Delta_{\min} := \min_{k \geq 2} \Delta_k$. There exists a family of independent unit-variance Gaussian instances $\mathcal{V} = \{\nu_0, \nu_2, \ldots, \nu_K\}$ sharing the same certificate U such that the first M_k low-fidelity observations of arm k have the same distribution under ν_0 and ν_k.*

Let $z_\Lambda := \Lambda\Delta_{\min}/\left(\lambda^{(H)} \sum_{k=2}^K \Delta_k^{-1}\right)$. If $z_\Lambda \geq 16$, then any budget-feasible policy whose total query cost is at least $\Lambda - \lambda^{(H)}$ almost surely under every $\nu \in \mathcal{V}$ satisfies

$$\max_{\nu \in \mathcal{V}} \mathbb{E}_\nu[R(\Lambda)] \geq \frac{1}{16} \sum_{k=2}^K \Delta_k \min\left\{\lambda^{(L)} \max\left\{M_k + 1, \frac{\log z_\Lambda}{\Delta_k^2}\right\}, \lambda^{(H)} \frac{\log z_\Lambda}{\Delta_k^2}\right\}. \tag{15}$$

The bound captures both improvement (M_k) and statistical ($\log z_\Lambda/\Delta_k^2$) requirements. For $U_k(n) = Cn^{1-r}$, $0 < r \leq 1$, $M_k = \lfloor (C/\Delta_k)^{1/r} \rfloor$, making the dependence on r explicit. Full proof is given in Appendix B.3.

[2] Appendix C gives a phase-based reference scheme that avoids this continuation-specific condition.

6 EXPERIMENTS

We evaluate TACC and its empirical parameterizations in three complementary settings. Synthetic bandits instantiate the improving-proxy model. A NLI task studies fidelity allocation in policy evaluation. NAS-Bench-201 uses empirical learning curves and an offline-calibrated mismatch certificate. These comparisons explore when bounded continuation can reduce high-fidelity sampling of suboptimal arms. Code is available at `https://anonymous.4open.science/r/tacc-anonymous-release-CF04/`.

6.1 SYNTHETIC BANDIT EXPERIMENTS

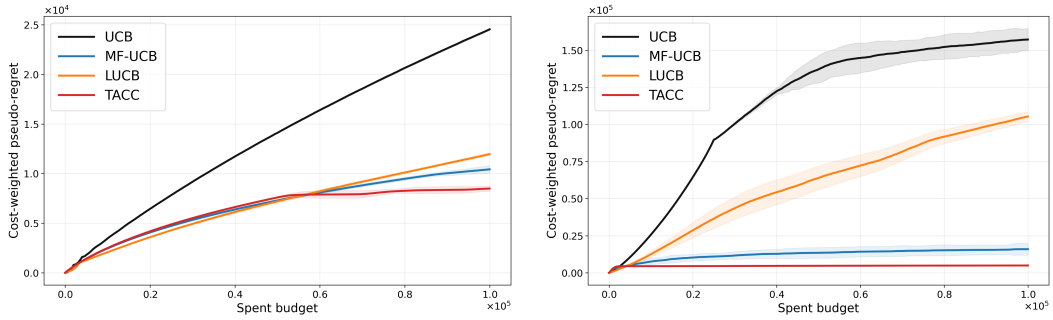
Synthetic model and setting. We first evaluate TACC in controlled synthetic bandits that directly instantiate the improving-proxy model. For each arm k, the low-fidelity mean evolves as

$$\mu_k^{(L)}(n) = \mu_k^{(H)} + a_k(n+n_0)^{-r}, \qquad a_k = s_k\zeta m_k(n_0+1)^r, \tag{16}$$

where $s_k \in \{-1,+1\}$ determines the mismatch direction and $m_k \sim \mathcal{U}(0.4,1)$ is its arm-specific magnitude. Thus, $\mu_k^{(L)}(1) - \mu_k^{(H)} = s_k\zeta m_k$, with ζ controlling the initial mismatch scale and r its decay with low-fidelity usage. The selected-average bound $B_k(n)$ is computed from this trajectory.

We consider two settings. **Set A** has $K = 200$ arms, $\mu_k^{(H)} \sim \mathcal{U}(0.1, 0.9)$, $\zeta = 0.2$, and costs $(\lambda^{(L)}, \lambda^{(H)}) = (1, 10)$. **Set B** has $K = 500$ arms, $\mu_k^{(H)} \sim \mathcal{N}(0,1)$, $\zeta = 1$, and costs $(1, 50)$. Both use $r = 0.5$; full parameter settings and sensitivity analyses are given in Appendix D.1 and D.2.

Baselines and results. We compare TACC with high-fidelity UCB (Auer et al., 2002), static MF-UCB (Kandasamy et al., 2016), and an LUCB-style two-fidelity rule (Wang et al., 2023). Figure 1 reports mean cost-weighted pseudo-regret over 10 independent trials. TACC achieves lower average regret in both settings, with a larger advantage in Set B, where high-fidelity confirmation is more expensive. This supports the intended mechanism: bounded low-fidelity continuation is most valuable when it can replace costly high-fidelity queries.



(a) Bandit Set A, costs $(1, 10)$ (b) Bandit Set B, costs $(1, 50)$

Figure 1: Synthetic simulation results for the reported configurations with $r = 0.5$.

6.2 NLI POLICY EVALUATION

We next study a Natural Language Inference (NLI) task motivated by LLM-as-a-judge applications. Each arm is a fixed rule-based policy mapping a premise–hypothesis pair to a label in $\{\text{ENTAILMENT}, \text{NEUTRAL}, \text{CONTRADICTION}\}$. High-fidelity feedback is sampled from a Bernoulli distribution with the policy's verifier-accuracy mean; low-fidelity feedback follows a controlled improving-proxy model. We use four policy arms to focus on fidelity allocation rather than live LLM generation or judge refinement.

Proxy model. To isolate fidelity allocation, we use the same count-dependent proxy construction as in the synthetic experiments with an additional residual mismatch b_k. All mismatch-aware methods receive the same valid selected-average bound $B_k(n)$ without observing the latent low-fidelity trajectory. Thus, the experiment evaluates TACC under controlled LLM-judge mismatch. Persistent judge–target disagreement motivates this model (Zheng et al., 2023; Huang et al., 2025). Further protocol details and separate empirical weak-judge analyses are provided in Appendix D.7–D.10.

Baselines and protocol. We compare TACC with DNC, MF-UCB, and high-fidelity UCB. DNC uses the same aggregate optimistic score, count-dependent bound $B_k(n)$, and threshold γ as TACC, but queries high fidelity if an arm is selected after its low-fidelity bonus falls below γ. Thus, it serves as an ablation under a matched nominal threshold. Additional diagnostic baselines are reported in Appendix D.

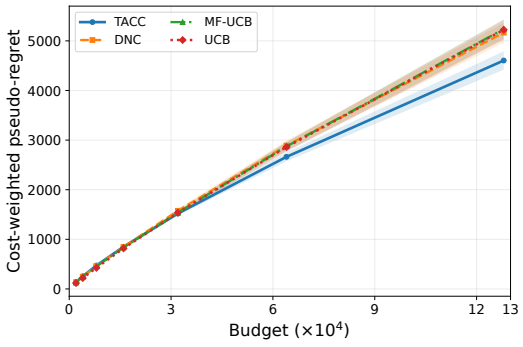


Figure 2: Exploratory NLI experiment. Curves show mean cost-weighted pseudo-regret over 200 seeds.

Exploratory results. We use $\lambda^{(L)} = 1$, $\lambda^{(H)} = 500$, $r = 0.75$, $\zeta = 0.4$, $b_k \equiv 0.05$, $\gamma = 0.025$, $\eta = 10^{-4}$, and $S_0 = 128$, with budgets up to $\Lambda = 128000$. Figure 2 reports results over 200 independent seeds. TACC achieves the lowest final mean regret. In particular, its improvement over DNC shows that bounded continuation provides additional benefit even when both methods use the same count-dependent mismatch information; comparisons with MF-UCB and UCB show the benefit over static two-fidelity switching and high-fidelity-only evaluation.

6.3 NAS-Bench-201 Architecture Evaluation

The preceding experiments use prescribed proxy-improvement trajectories. We next evaluate TACC with empirical learning curves and an offline-calibrated mismatch certificate using NAS-Bench-201 (Dong & Yang, 2020). Each arm is an architecture, partial-training performance serves as the improving low-fidelity proxy, and fully trained validation performance defines the high-fidelity target.

We construct the mismatch certificate before bandit execution using paired partial- and full-training trajectories from a separate calibration set of architectures. Architectures used for proxy training, certificate calibration, hyperparameter tuning, and final evaluation are disjoint. The resulting certificate achieves approximately 96–98% simultaneous coverage on held-out architectures and remains fixed throughout the online bandit run. Thus, no high-fidelity observations from the evaluation budget are used to construct or update the certificate.

We evaluate ImageNet16-120 (Chrabaszcz et al., 2017), CIFAR-10-valid, and CIFAR-100 (Krizhevsky et al., 2009) with $K = 20$, $\lambda^{(L)} = 1$, $\lambda^{(H)} = 40$, $\gamma = 0.015$, $\zeta = 1$ and budget $\Lambda = 2 \times 10^5$. Results are averaged over three independent architecture splits. As shown in Table 1, TACC achieves the lowest mean cost-weighted regret on all three datasets, indicating that bounded continuation remains effective when both the low-fidelity improvement trajectory and the mismatch certificate are obtained from empirical learning curves rather than a manually specified model.

Table 1: Cost-weighted regret on NAS-Bench-201, reported as mean $\pm$ standard error over 300 runs.

Dataset	TACC	DNC	MF-UCB	UCB
ImageNet16-120	$\mathbf{123.31 \pm 2.07}$	150.51 ± 2.54	173.39 ± 2.62	186.55 ± 2.64
CIFAR-10-valid	$\mathbf{107.90 \pm 1.61}$	116.64 ± 2.91	131.63 ± 2.93	143.77 ± 2.84
CIFAR-100	$\mathbf{159.90 \pm 2.31}$	167.69 ± 2.75	218.48 ± 2.74	214.92 ± 2.69

7 Conclusion

Motivated by self-evolving evaluators whose fidelity improves with additional investment, we study multi-fidelity bandits with usage-dependent proxy discrepancy. Using an a priori selected-average mismatch bound, we derive improvement-aware confidence bounds and propose TACC, which combines adaptive arm selection with certificate-informed bounded continuation. Our analysis identifies conditions under which logarithmic high-fidelity confirmation of suboptimal arms can be avoided, supported by experiments on synthetic bandits, a NLI policy evaluation, and NAS-Bench-201. The current framework assumes two fidelities, arm-local proxy evolution, and a prespecified mismatch certificate. Extending TACC to multiple or jointly evolving proxies and jointly learning the mismatch certificate and fidelity-allocation policy are promising directions.

AI USE STATEMENT

In this work, we used Generative AI tools to improve the clarity, readability, and language of the manuscript. We have not used generative AI tools for developing the theoretical framework, formulating mathematical claims or proofs, designing the methodology or experiments, implementing the proposed methods, analyzing data, or interpreting experimental results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Leonardo Cella, Massimiliano Pontil, and Claudio Gentile. Best model identification: A rested bandit formulation. In *International Conference on Machine Learning*, pp. 1362–1372. PMLR, 2021.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Xuanyi Dong and Yi Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. *arXiv preprint arXiv:2001.00326*, 2020.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, 2018.

Hoda Heidari, Michael Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1562–1570, 2016.

Hui Huang, Yingqi Qu, Xingyuan Bu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics*, 2025.

Nimit Kalra and Leonard Tang. Verdict: A library for scaling judge-time compute. *arXiv preprint arXiv:2502.18018*, 2025.

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. The multi-fidelity multi-armed bandit. *Advances in Neural Information Processing Systems*, 29, 2016.

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.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Ilgam Latypov, Alexandra Suvorikova, Alexey Kroshnin, Alexander Gasnikov, and Yuriy Dorn. Ucb-type algorithm for budget-constrained expert learning. *arXiv preprint arXiv:2510.22654*, 2025.

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.

Alberto Maria Metelli, Francesco Trovò, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15421–15457. PMLR, 2022.

Marco Mussi, Alessandro Montenegro, Francesco Trovo, Marcello Restelli, and Alberto Maria Metelli. Best arm identification for stochastic rising bandits. *arXiv preprint arXiv:2302.07510*, 2023.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Siam Review*, 60(3):550–591, 2018.

Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 2019.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of llm-as-a-judge: Interval evaluations with conformal prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11297–11339, 2025.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pp. 63277–63303, 2025.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*, 2024.

Xuchuang Wang, Qingyun Wu, Wei Chen, and John C. S. Lui. Multi-fidelity multi-armed bandits revisited. *Advances in Neural Information Processing Systems*, 36, 2023.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

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.

A NOTATION

K, $[K]$	Number of arms; arm index set.
L, H	Low fidelity and high fidelity.
$k^\star$, $\mu^\star$	Optimal arm and its high-fidelity mean.
$\mu_k^{(H)}$	High-fidelity mean of arm k.
Δ_k	High-fidelity suboptimality gap of arm k.
$\lambda^{(L)}$, $\lambda^{(H)}$	Query costs of low and high fidelity.
Λ, T_Λ	Total budget; query upper bound induced by the budget.
C_t	Cumulative cost up to round t.
$R(\Lambda)$	Cost-weighted pseudo-regret under budget Λ.
(I_t, m_t)	Arm and fidelity selected at round t.
$Y_{k,u}^{(H)}$, $Y_{k,\tau}^{(L)}$	High- and low-fidelity observations.
$\mu_k^{(L)}(\tau)$	Low-fidelity mean at the τ-th low-fidelity query of arm k.
$\bar{\mu}_k^{(L)}(n)$	Selected-average low-fidelity mean after n low-fidelity queries.
$U_k(n)$	A priori cumulative mismatch certificate.
$B_k(n)$	Selected-average mismatch bound induced by $U_k(n)$.
$\Delta_k^{(L)}(n)$	Effective low-fidelity gap relative to the high-fidelity optimum.
$\ell_\Lambda(\delta)$, $G_\Lambda(n,\delta)$	Budget-uniform log factor and statistical confidence radius.
$\mathrm{rad}_k^{(L)}(n,\delta)$, $\mathrm{rad}^{(H)}(n,\delta)$	Low- and high-fidelity confidence radii.
$N_{k,t}^{(L)}$, $N_{k,t}^{(H)}$	Low- and high-fidelity query counts before round t.
$\hat{\mu}_{k,t}^{(L)}$, $\hat{\mu}_{k,t}^{(H)}$	Empirical means before round t.
$\mathrm{UCB}_{k,t}^{(m)}$, $\mathrm{LCB}_{k,t}^{(m)}$	Fidelity-m confidence bounds for $\mu_k^{(H)}$.
$\mathrm{UCB}_{k,t}$	Aggregate optimistic score used for arm selection.
γ	Low-fidelity statistical uncertainty threshold.
η	Continuation threshold fraction.
S_0	Maximum number of post-threshold continuation pulls.
A_k	Continuation pulls already spent on arm k.
$V_{k,t}(s)$	Anticipated decrease in B_k after s additional low-fidelity queries.
$N_\gamma(\delta)$	Nominal low-fidelity switching count.
$\tau_k(\gamma)$	Low-fidelity certification time at level 2γ.
$\mathcal{A}_\gamma(\delta)$, $\mathcal{B}_\gamma(\delta)$, $\mathcal{C}_\gamma(\delta)$	Easy, hard, and intermediate arm classes.
$\mathcal{C}_\gamma^{\mathrm{d}}(\delta)$, $\mathcal{C}_\gamma^{\mathrm{u}}(\delta)$	Detectable and undetectable intermediate subclasses.
$\mathcal{H}_\gamma(\delta)$	Arms that may require logarithmic high-fidelity confirmation.
$\mathcal{E}_\Lambda$	Budget-uniform confidence event.

B ADDITIONAL PROOF DETAILS AND ALGORITHMIC VARIANTS

This appendix contains the proofs of Lemma B.1, Propositions B.5, B.6, and B.7, Theorem 5.3, Corollary 5.4, Theorem 5.6 and Proposition B.8.

Lemma B.1 (Valid confidence intervals)**.** *Suppose Assumption 3.1 holds. Then there exists an event $\mathcal{E}_\Lambda$ with probability at least $1-\delta$ such that, on $\mathcal{E}_\Lambda$, for all arms $k \in [K]$ and all rounds $t \le T_\Lambda$,*

$$\mu_k^{(H)} \in \left[\mathrm{LCB}_{k,t}^{(L)}, \mathrm{UCB}_{k,t}^{(L)}\right] \quad \textit{and} \quad \mu_k^{(H)} \in \left[\mathrm{LCB}_{k,t}^{(H)}, \mathrm{UCB}_{k,t}^{(H)}\right]. \tag{17}$$

Throughout the appendix, we work on the event $\mathcal{E}_\Lambda$ from Lemma B.1. Recall that

$$\ell_\Lambda(\delta) := 2\log\left(\frac{4KT_\Lambda}{\delta}\right),\ G_\Lambda(n,\delta) := \sqrt{\frac{\ell_\Lambda(\delta)}{n}},\ N_\gamma(\delta) := \min\{n \ge 1 : G_\Lambda(n,\delta) < \gamma\}.$$

For any fixed arm, fidelity, and sample size n, the empirical mean of 1-sub-Gaussian observations satisfies

$$\Pr(|\widehat{\mu}_n - \mathbb{E}[\widehat{\mu}_n]| > G_\Lambda(n,\delta)) \le 2\exp\left(-\frac{nG_\Lambda^2(n,\delta)}{2}\right) = \frac{\delta}{2KT_\Lambda}.$$

A union bound over the K arms, two fidelities, and all $1 \le n \le T_\Lambda$ gives the desired budget-uniform event.

Proof of Lemma B.1. Unqueried fidelities have interval $(-\infty, +\infty)$, so it suffices to consider positive sample counts. For high fidelity, the claim is standard: for each fixed arm k and sample size n, the empirical mean of the first n high-fidelity samples is 1-sub-Gaussian around $\mu_k^{(H)}$. By a uniform concentration bound over all $k \in [K]$ and all $1 \le n \le T_\Lambda$, with probability at least $1 - \delta/2$,

$$\left|\hat{\mu}_{k,t}^{(H)} - \mu_k^{(H)}\right| \le G_\Lambda\big(N_{k,t}^{(H)}, \delta\big) \tag{18}$$

for all arms k and all rounds $t \le T_\Lambda$.

For low fidelity, recall the selected-average low-fidelity mean

$$\bar{\mu}_k^{(L)}(n) := \frac{1}{n}\sum_{\tau=1}^{n} \mu_k^{(L)}(\tau).$$

The low-fidelity empirical mean concentrates around $\bar{\mu}_k^{(L)}(n)$, not around the instantaneous mean $\mu_k^{(L)}(n)$. Using

$$\hat{\mu}_{k,t}^{(L)} - \mu_k^{(H)} = \big(\hat{\mu}_{k,t}^{(L)} - \bar{\mu}_k^{(L)}(N_{k,t}^{(L)})\big) + \big(\bar{\mu}_k^{(L)}(N_{k,t}^{(L)}) - \mu_k^{(H)}\big), \tag{19}$$

the first term is controlled by the same sub-Gaussian concentration argument, which yields, with probability at least $1 - \delta/2$,

$$\left|\hat{\mu}_{k,t}^{(L)} - \bar{\mu}_k^{(L)}(N_{k,t}^{(L)})\right| \le G_\Lambda\big(N_{k,t}^{(L)}, \delta\big) \tag{20}$$

for all arms k and all rounds $t \le T_\Lambda$.

For the second term, Assumption 3.1 gives

$$\left|\bar{\mu}_k^{(L)}(n) - \mu_k^{(H)}\right| \le \frac{1}{n}\sum_{\tau=1}^{n}\left|\mu_k^{(L)}(\tau) - \mu_k^{(H)}\right| \le \frac{U_k(n)}{n} \le B_k(n). \tag{21}$$

Combining the stochastic and deterministic bounds gives

$$\left|\hat{\mu}_{k,t}^{(L)} - \mu_k^{(H)}\right| \le G_\Lambda\big(N_{k,t}^{(L)}, \delta\big) + B_k\big(N_{k,t}^{(L)}\big). \tag{22}$$

This is exactly the low-fidelity confidence interval. Taking a union bound over the high-fidelity and low-fidelity events proves that, with probability at least $1 - \delta$, both intervals cover $\mu_k^{(H)}$ uniformly over all arms and all rounds within the budget horizon.

Remark B.2 (Budget-uniform confidence radius). Because the budget Λ and fidelity costs are known in advance, T_Λ is a deterministic upper bound on the total number of queries. The radius $G_\Lambda(n, \delta)$ therefore defines a single confidence event that is uniform over arms, fidelities, and sample counts up to the budget horizon. An anytime confidence radius could be used instead, but the budget-uniform form yields a deterministic switching count $N_\gamma(\delta)$ and simplifies the continuation analysis.

B.1 Robustness to an approximate mismatch certificate

The algorithm does not require the supplied mismatch bound to coincide with the realized low–high discrepancy. Suppose instead that it uses an approximate certificate $\widetilde{B}_k(n)$.

Proposition B.3 (Conservative certificate robustness)**.** *If*

$$\widetilde{B}_k(n) \ge \left|\bar{\mu}_k^{(L)}(n) - \mu_k^{(H)}\right|$$

for every relevant arm k and sample count n, then the low-fidelity confidence intervals remain valid after replacing B_k by $\widetilde{B}_k$, and the regret analysis applies with the corresponding effective gaps, certification times, and arm classes.

Proof. The proof of Lemma B.1 uses $B_k(n)$ only through

$$\left|\bar{\mu}_k^{(L)}(n) - \mu_k^{(H)}\right| \le B_k(n).$$

Replacing $B_k(n)$ by any valid upper bound $\widetilde{B}_k(n)$ therefore preserves the low-fidelity confidence event. The remaining analysis follows after redefining the quantities that depend on the mismatch correction. □

In particular, if

$$B_k(n) \le \widetilde{B}_k(n) + \varepsilon$$

uniformly over the relevant arms and sample counts, then

$$B_k^{\mathrm{r}}(n) := \widetilde{B}_k(n) + \varepsilon$$

is a valid conservative certificate. Such inflation can delay low-fidelity certification and move arms toward classes requiring high-fidelity confirmation, but does not invalidate the confidence guarantee. By contrast, uncontrolled underestimation of the mismatch may invalidate the low-fidelity confidence interval and hence the original regret guarantee. A constant additive correction cancels from the continuation gain $B_k(n) - B_k(n+s)$, whereas errors in the shape of the supplied certificate can alter continuation decisions.

Remark B.4 (Online certificate estimation). The present framework treats the mismatch certificate as prior information fixed before the bandit run. Estimating $B_k(n)$ online would require high-fidelity observations to jointly learn the target mean and calibrate the low–high discrepancy. Without such high-fidelity anchors or additional structure on the mismatch trajectory, the certificate is not generally identifiable from low-fidelity observations alone. We therefore leave joint online certificate estimation and fidelity allocation to future work.

B.2 Upper-bound construction and proofs

We next record the two selection conditions used in the regret proof. Algorithm 1 selects arms by the aggregate score $\mathrm{UCB}_{k,t} = \min\{\mathrm{UCB}_{k,t}^{(L)}, \mathrm{UCB}_{k,t}^{(H)}\}$. Thus a suboptimal arm can be selected only if neither fidelity-specific upper bound has already certified it as suboptimal.

Proposition B.5 (Low-fidelity selection condition)**.** *Suppose $\mathcal{E}_\Lambda$ holds. If a suboptimal arm $k \neq k^\star$ is selected at round t with $N_{k,t}^{(L)} \ge 1$, then $2G_\Lambda\Big(N_{k,t}^{(L)}, \delta\Big) \ge \Delta_k^{(L)}\Big(N_{k,t}^{(L)}\Big)$. Consequently, if $G_\Lambda\Big(N_{k,t}^{(L)}, \delta\Big) < \gamma$, and $\Delta_k^{(L)}\Big(N_{k,t}^{(L)}\Big) \ge 2\gamma$, then arm k is certified as suboptimal by low fidelity and will not be selected again.*

Proposition B.6 (High-fidelity selection condition)**.** *Suppose $\mathcal{E}_\Lambda$ holds. If a suboptimal arm $k \neq k^\star$ is selected and queried at high fidelity at round t, then $2G_\Lambda\Big(N_{k,t}^{(H)}, \delta\Big) \ge \Delta_k$. Consequently,*

$$N_{k,T_\Lambda}^{(H)} \le 1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil. \tag{23}$$

Proof of Proposition B.5. Fix a suboptimal arm $k \neq k^\star$ and suppose it is selected at round t. Since the algorithm selects

$$I_t \in \arg\max_{j\in[K]} UCB_{j,t}, \qquad UCB_{j,t} := \min_{m\in\{L,H\}} UCB_{j,t}^{(m)},$$

we have

$$UCB_{k,t} \ge UCB_{k^\star,t}. \tag{24}$$

On $\mathcal{E}_\Lambda$, both fidelity-specific confidence intervals for the optimal arm cover $\mu_{k^\star}^{(H)} = \mu^\star$. Hence

$$UCB_{k^\star,t}^{(L)} \ge \mu^\star, \qquad UCB_{k^\star,t}^{(H)} \ge \mu^\star,$$

and therefore

$$UCB_{k^\star,t} = \min\{UCB_{k^\star,t}^{(L)}, UCB_{k^\star,t}^{(H)}\} \ge \mu^\star. \tag{25}$$

Thus

$$UCB_{k,t} \ge \mu^\star. \tag{26}$$

Since $UCB_{k,t}$ is the minimum of the two fidelity-specific upper bounds, this implies in particular that

$$UCB_{k,t}^{(L)} \ge \mu^\star. \tag{27}$$

Let $n = N_{k,t}^{(L)}$. By the definition of the low-fidelity upper bound,

$$UCB_{k,t}^{(L)} = \hat{\mu}_{k,t}^{(L)} + G_\Lambda(n,\delta) + B_k(n). \tag{28}$$

On $\mathcal{E}_\Lambda$, the empirical low-fidelity mean satisfies

$$\hat{\mu}_{k,t}^{(L)} \le \bar{\mu}_k^{(L)}(n) + G_\Lambda(n,\delta).$$

Therefore,

$$\mu^\star \le UCB_{k,t}^{(L)} \le \bar{\mu}_k^{(L)}(n) + 2G_\Lambda(n,\delta) + B_k(n). \tag{29}$$

Rearranging gives

$$2G_\Lambda(n,\delta) \ge \mu^\star - \bar{\mu}_k^{(L)}(n) - B_k(n) = \Delta_k^{(L)}(n). \tag{30}$$

For the consequence, suppose that

$$G_\Lambda\big(N_{k,t}^{(L)},\delta\big) < \gamma \qquad \text{and} \qquad \Delta_k^{(L)}\big(N_{k,t}^{(L)}\big) \ge 2\gamma.$$

Then

$$2G_\Lambda\big(N_{k,t}^{(L)},\delta\big) < 2\gamma \le \Delta_k^{(L)}\big(N_{k,t}^{(L)}\big),$$

which contradicts the necessary selection condition just proved. Hence arm k cannot be selected under these two inequalities. Since the arm is not selected, its low-fidelity count does not increase, and the same certificate continues to rule out future selections. Thus the arm is certified as suboptimal by low fidelity.

Proof of Proposition B.6. Fix a suboptimal arm $k \ne k^\star$. Suppose arm k is selected and queried at high fidelity at round t. Since k is selected, the same argument used in the proof of Proposition B.5 gives

$$UCB_{k,t} \ge \mu^\star. \tag{31}$$

Because

$$UCB_{k,t} = \min\{UCB_{k,t}^{(L)}, UCB_{k,t}^{(H)}\},$$

we must have

$$UCB_{k,t}^{(H)} \ge \mu^\star. \tag{32}$$

Let $n = N_{k,t}^{(H)}$. By definition,

$$UCB_{k,t}^{(H)} = \hat{\mu}_{k,t}^{(H)} + G_\Lambda(n,\delta).$$

On $\mathcal{E}_\Lambda$,

$$\hat{\mu}_{k,t}^{(H)} \le \mu_k^{(H)} + G_\Lambda(n,\delta).$$

Therefore,

$$\mu^\star \le UCB_{k,t}^{(H)} \le \mu_k^{(H)} + 2G_\Lambda(n,\delta). \tag{33}$$

Rearranging gives

$$2G_\Lambda(n,\delta) \ge \mu^\star - \mu_k^{(H)} = \Delta_k. \tag{34}$$

Since

$$G_\Lambda(n,\delta) = \sqrt{\frac{\ell_\Lambda(\delta)}{n}},$$

the condition $2G_\Lambda(n,\delta) \ge \Delta_k$ implies

$$n \le \frac{4\ell_\Lambda(\delta)}{\Delta_k^2}. \tag{35}$$

At each high-fidelity query, $N_{k,t}^{(H)}$ denotes the number of previous high-fidelity queries before the current one. Hence the total number of high-fidelity queries of arm k is bounded by

$$N_{k,T_\Lambda}^{(H)} \le 1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil, \tag{36}$$

where the additive constant accounts for initialization and the boundary query.

We next bound the number of low-fidelity queries. The bound follows from the threshold rule before $N_\gamma(\delta)$ and from the continuation cap after the threshold crossing. For arms in $\mathcal{C}_\gamma(\delta)$, low-fidelity

certification is reachable within the continuation window, but TACC must also continue sampling until the certification count is reached. We therefore define

$$\mathcal{C}_\gamma^{\mathrm{d}}(\delta) := \left\{ k \in \mathcal{C}_\gamma(\delta) : \begin{array}{l} \text{for every } N_\gamma(\delta) \le n < \tau_k(\gamma), \\ \exists\, 1 \le s \le N_\gamma(\delta) + S_0 - n \text{ such that } B_k(n) - B_k(n+s) \ge 2\eta\gamma \end{array} \right\}. \tag{37}$$

Finally, let $\mathcal{C}_\gamma^{\mathrm{u}}(\delta) := \mathcal{C}_\gamma(\delta) \setminus \mathcal{C}_\gamma^{\mathrm{d}}(\delta)$.

Proposition B.7 (Low-fidelity query bound)**.** *Suppose $\mathcal{E}_\Lambda$ holds. Under Algorithm 1, every suboptimal arm $k \neq k^\star$ satisfies*

$$N_{k,T_\Lambda}^{(L)} \le N_\gamma(\delta) + S_0 + 1, \tag{38}$$

where the additive constant accounts for initialization and boundary effects. Moreover:

(i) if $k \in \mathcal{A}_\gamma(\delta)$, then k is certified by low fidelity at the nominal switching threshold;

(ii) if $k \in \mathcal{C}_\gamma^{\mathrm{d}}(\delta)$, then TACC continues querying low fidelity whenever the arm is selected before $\tau_k(\gamma)$, and the arm is certified by low fidelity once this count is reached.

Proof. Fix a suboptimal arm $k \neq k^\star$. Low fidelity can be queried before the threshold crossing only while $G_\Lambda(N_{k,t}^{(L)}, \delta) \ge \gamma$. By the definition of $N_\gamma(\delta)$, this implies $N_{k,t}^{(L)} < N_\gamma(\delta)$. Hence the pre-threshold phase contributes at most $N_\gamma(\delta) + 1$ low-fidelity pulls, including the boundary effect from pre-round counts. After the threshold crossing, low fidelity can be queried only through the continuation rule, whose counter is capped by S_0. Therefore

$$N_{k,T_\Lambda}^{(L)} \le N_\gamma(\delta) + S_0 + 1.$$

If $k \in \mathcal{A}_\gamma(\delta)$, then $\tau_k(\gamma) \le N_\gamma(\delta)$. By Assumption 5.1, the inequality $\Delta_k^{(L)}(n) \ge 2\gamma$ remains valid through the relevant continuation window once it first holds. In particular, when the low-fidelity radius first falls below γ, we have both $G_\Lambda(N_{k,t}^{(L)}, \delta) < \gamma$ and $\Delta_k^{(L)}(N_{k,t}^{(L)}) \ge 2\gamma$. Proposition B.5 then rules out further selection of arm k.

If $k \in \mathcal{C}_\gamma^{\mathrm{d}}(\delta)$, then for every count $N_\gamma(\delta) \le n < \tau_k(\gamma)$, the definition of $\mathcal{C}_\gamma^{\mathrm{d}}(\delta)$ guarantees an admissible s such that

$$B_k(n) - B_k(n+s) \ge 2\eta\gamma.$$

Hence, whenever arm k is selected before reaching $\tau_k(\gamma)$, the continuation condition is satisfied and TACC queries low fidelity. Since $\tau_k(\gamma) \le N_\gamma(\delta) + S_0$ for $k \in \mathcal{C}_\gamma(\delta)$, the certification count is reached within the continuation window whenever the arm continues to be selected. At that count, $\Delta_k^{(L)}(\tau_k(\gamma)) \ge 2\gamma$; Assumption 5.1 keeps the certificate valid within the window, and Proposition B.5 rules out further selection. □

Proof of Theorem 5.3. On $\mathcal{E}_\Lambda$, Lemma B.1 ensures that all confidence intervals are valid uniformly over the budget horizon. The regret of querying a suboptimal arm k at fidelity $m \in \{L, H\}$ is $\lambda^{(m)}\Delta_k$. Therefore,

$$R(\Lambda) = \sum_{k \neq k^\star} \Delta_k \left[\lambda^{(L)} N_{k,T_\Lambda}^{(L)} + \lambda^{(H)} N_{k,T_\Lambda}^{(H)} \right], \tag{39}$$

where the optimal arm contributes zero pseudo-regret.

For $k \in \mathcal{A}_\gamma(\delta)$, low fidelity certifies suboptimality by the nominal switching threshold because $\tau_k(\gamma) \le N_\gamma(\delta)$. By Proposition B.7, such an arm uses at most $N_\gamma(\delta)+1$ low-fidelity pulls, including the boundary effect, and does not incur the logarithmic high-fidelity confirmation term. Keeping one high-fidelity initialization or reference query gives

$$R_k(\Lambda) \le \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + 1) + \lambda^{(H)} \right]. \tag{40}$$

For $k \in \mathcal{C}_\gamma^{\mathrm{d}}(\delta)$, the arm is not certified at the static threshold, but the continuation condition remains satisfied whenever it is selected before $\tau_k(\gamma)$. Since $\tau_k(\gamma) \le N_\gamma(\delta) + S_0$, Proposition B.7 gives at most $N_\gamma(\delta) + S_0 + 1$ low-fidelity pulls and no logarithmic high-fidelity confirmation term. Hence

$$R_k(\Lambda) \le \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)} \right]. \tag{41}$$

For $k \in \mathcal{H}_\gamma(\delta) = \mathcal{B}_\gamma(\delta) \cup \mathcal{C}^{\mathrm{u}}_\gamma(\delta)$, the analysis does not credit the arm with realized low-fidelity continuation certification. The low-fidelity count is still bounded by Proposition B.7, while the number of high-fidelity queries is bounded by Proposition B.6:

$$N^{(L)}_{k,T_\Lambda} \le N_\gamma(\delta) + S_0 + 1, \qquad N^{(H)}_{k,T_\Lambda} \le 1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil.$$

Therefore

$$R_k(\Lambda) \le \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)} \left(1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil \right) \right]. \tag{42}$$

Summing these three bounds over the corresponding arm classes yields

$$\begin{aligned} R(\Lambda) \le & \sum_{k \in \mathcal{A}_\gamma(\delta)} \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + 1) + \lambda^{(H)} \right] \\ & + \sum_{k \in \mathcal{C}^{\mathrm{d}}_\gamma} \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)} \right] \\ & + \sum_{k \in \mathcal{B}_\gamma(\delta) \cup \mathcal{C}^{\mathrm{u}}_\gamma} \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)} \left(1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil \right) \right]. \end{aligned} \tag{43}$$

This is the claimed regret bound.

Proof of Corollary 5.4. The result follows by simplifying the finite-sample bound in Theorem 5.3. Since $N_\gamma(\delta) \le 1 + \ell_\Lambda(\delta)/\gamma^2$, the low-fidelity contribution of each suboptimal arm is

$$O\left(\lambda^{(L)} \Delta_k \left(\frac{\ell_\Lambda(\delta)}{\gamma^2} + S_0 + 1 \right) \right).$$

For arms in $\mathcal{A}_\gamma(\delta) \cup \mathcal{C}^{\mathrm{d}}_\gamma(\delta)$, the high-fidelity contribution is only the initialization or reference cost. The logarithmic high-fidelity confirmation term appears only for arms in $\mathcal{H}_\gamma(\delta)$. For each such arm,

$$\Delta_k \lambda^{(H)} \left(1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil \right) = O\left(\lambda^{(H)} \frac{\ell_\Lambda(\delta)}{\Delta_k} \right) + O(\lambda^{(H)} \Delta_k).$$

Using

$$\ell_\Lambda(\delta) = 2 \log \left(\frac{4KT_\Lambda}{\delta} \right)$$

gives equation 13.

Proposition B.8 (Arm-level upper-bound comparison with the static threshold rule)**.** *For any arm $k \in \mathcal{C}^{\mathrm{d}}_\gamma(\delta)$, define the static and adaptive arm-level upper bounds*

$$\overline{R}_{k,\mathrm{static}}(\Lambda) := \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + 1) + \lambda^{(H)} \left(1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil \right) \right], \tag{44}$$

$$\overline{R}_{k,\mathrm{adapt}}(\Lambda) := \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)} \right]. \tag{45}$$

Then

$$R_{k,\mathrm{static}}(\Lambda) \le \overline{R}_{k,\mathrm{static}}(\Lambda), \qquad R_{k,\mathrm{adapt}}(\Lambda) \le \overline{R}_{k,\mathrm{adapt}}(\Lambda),$$

and

$$\overline{R}_{k,\mathrm{static}}(\Lambda) - \overline{R}_{k,\mathrm{adapt}}(\Lambda) = \Delta_k \left[\lambda^{(H)} \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil - \lambda^{(L)} S_0 \right]. \tag{46}$$

In particular, if $S_0 \lambda^{(L)} \le \lambda^{(H)}$ and $\lceil 4\ell_\Lambda(\delta)/\Delta_k^2 \rceil \ge 2$, then the adaptive arm-level upper bound is strictly smaller.

Proof. Fix $k \in \mathcal{C}^{\mathrm{d}}_\gamma(\delta)$. Under the static threshold rule, low fidelity is queried only until $G_\Lambda(N^{(L)}_{k,t}, \delta) < \gamma$. Since $k \in \mathcal{C}_\gamma(\delta)$, we have $\tau_k(\gamma) > N_\gamma(\delta)$, so the sufficient 2γ certificate

has not been reached at the nominal threshold. The static rule queries high fidelity only if this arm is selected again; such queries are not necessary on every instance. Using the same threshold-count argument as in Proposition B.7, together with Proposition B.6, gives

$$R_{k,\mathrm{static}}(\Lambda) \le \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta)+1) + \lambda^{(H)}\left(1 + \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil\right)\right] = \overline{R}_{k,\mathrm{static}}(\Lambda). \tag{47}$$

For the adaptive rule, $k \in \mathcal{C}_\gamma^{\mathrm{d}}(\delta)$ implies that, whenever the arm is selected before $\tau_k(\gamma)$, the continuation condition remains satisfied and TACC queries low fidelity. Since $\tau_k(\gamma) \le N_\gamma(\delta) + S_0$, Proposition B.7 gives at most $N_\gamma(\delta) + S_0 + 1$ low-fidelity queries and no logarithmic high-fidelity confirmation term. Hence

$$R_{k,\mathrm{adapt}}(\Lambda) \le \Delta_k \left[\lambda^{(L)}(N_\gamma(\delta) + S_0 + 1) + \lambda^{(H)}\right] = \overline{R}_{k,\mathrm{adapt}}(\Lambda). \tag{48}$$

Subtracting the two upper-bound expressions gives

$$\overline{R}_{k,\mathrm{static}}(\Lambda) - \overline{R}_{k,\mathrm{adapt}}(\Lambda) = \Delta_k \left[\lambda^{(H)} \left\lceil \frac{4\ell_\Lambda(\delta)}{\Delta_k^2} \right\rceil - \lambda^{(L)} S_0\right]. \tag{49}$$

If $S_0\lambda^{(L)} \le \lambda^{(H)}$ and $\lceil 4\ell_\Lambda(\delta)/\Delta_k^2 \rceil \ge 2$, this difference is strictly positive. □

B.3 Lower-bound construction and proofs

For the lower bound, let

$$W_k(\Lambda) := \lambda^{(L)} N_{k,T_\Lambda}^{(L)} + \lambda^{(H)} N_{k,T_\Lambda}^{(H)}$$

denote the total cost spent on arm k.

Let $\mathcal{M}(U)$ denote the class of two-fidelity Gaussian bandit instances satisfying Assumption 3.1 with the prespecified certificate $U = (U_k)_{k\in[K]}$. For each suboptimal arm $k \ge 2$ and $\ell > 0$, define

$$\Psi_k(\ell) := \min\left\{\lambda^{(L)} \max\left\{M_k + 1, \frac{\ell}{\Delta_k^2}\right\},\ \lambda^{(H)} \frac{\ell}{\Delta_k^2}\right\}. \tag{50}$$

Also write

$$\Xi_\Delta := \lambda^{(H)} \sum_{k=2}^{K} \frac{1}{\Delta_k}, \tag{51}$$

so that

$$z_\Lambda = \frac{\Lambda \Delta_{\min}}{\Xi_\Delta}.$$

Common-certificate hard family. Fix gaps $0 < \Delta_k \le 1/4$, $k = 2, \dots, K$, and write

$$M_k := M_k(\Delta_k).$$

In the base instance ν_0,

$$\mu_1^{(H)} = \frac{1}{2}, \qquad \mu_k^{(H)} = \frac{1}{2} - \Delta_k \quad (k \ge 2),$$

and arm 1 is exact at low fidelity. For every $k \ge 2$, define

$$\mu_{k,0}^{(L)}(\tau) = \begin{cases} \frac{1}{2}, & \tau \le M_k, \\ \frac{1}{2} - \Delta_k, & \tau > M_k. \end{cases}$$

For each $k \ge 2$, define an alternative instance ν_k by changing only arm k:

$$\mu_{k,\nu_k}^{(H)} = \frac{1}{2} + \Delta_k, \qquad \mu_{k,\nu_k}^{(L)}(\tau) = \begin{cases} \frac{1}{2}, & \tau \le M_k, \\ \frac{1}{2} + \Delta_k, & \tau > M_k. \end{cases}$$

If $M_k = \infty$, the second branch is never reached.

Lemma B.9 (Admissibility and transcript information). *The family $\mathcal{V} = \{\nu_0, \nu_2, \ldots, \nu_K\}$ belongs to the same certificate class $\mathcal{M}(U)$. Moreover, for every adaptive policy stopped within budget,*

$$D_{\mathrm{KL}}\big(P_{\nu_0}^{\Lambda} \| P_{\nu_k}^{\Lambda}\big) = 2\Delta_k^2 \mathbb{E}_{\nu_0}\left[N_{k,T_\Lambda}^{(H)} + \big(N_{k,T_\Lambda}^{(L)} - M_k\big)_+\right]. \tag{52}$$

Proof. For arm k, the cumulative mismatch in either instance is

$$\Delta_k \min\{n, M_k\}.$$

For $n \le M_k$, this is at most $U_k(n)$ by the definition of $M_k(\Delta_k)$. For $n > M_k$, monotonicity of U_k gives

$$\Delta_k M_k \le U_k(M_k) \le U_k(n).$$

If $M_k = 0$, the mismatch is identically zero. If $M_k = \infty$, then the defining condition for M_k gives

$$\Delta_k n \le U_k(n)$$

for every finite n.

Hence every instance satisfies Assumption 3.1 with the same certificate.

The first M_k low-fidelity observations of arm k have identical distributions under ν_0 and ν_k. Every later low-fidelity observation, as well as every high-fidelity observation of arm k, differs in mean by $2\Delta_k$. Since the noise is unit-variance Gaussian, each such observation contributes KL divergence $2\Delta_k^2$. Summing the conditional KL contributions along the adaptive transcript gives equation 52. □

The construction is genuinely improving: the instantaneous low-to-high mismatch decreases from Δ_k to zero after the masked prefix.

Lemma B.10 (Cost of revealing an improving proxy). *Let A be an event measurable with respect to the final transcript such that*

$$P_{\nu_0}(A) \ge 1 - \varepsilon, \qquad P_{\nu_k}(A) \le \varepsilon, \qquad 0 < \varepsilon \le \frac{1}{16}.$$

Then

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \ge \frac{1}{8}\Psi_k\left(\log\frac{1}{\varepsilon}\right). \tag{53}$$

Proof. Write $\Delta = \Delta_k$, $M = M_k$, and

$$q := \frac{\log(1/\varepsilon)}{\Delta^2}.$$

Data processing and $\mathrm{kl}(p, q') \ge p\log(1/q') - \log 2$ give

$$D_{\mathrm{KL}}\left(P_{\nu_0}^{\Lambda} \| P_{\nu_k}^{\Lambda}\right) \ge \frac{1}{2}\log\frac{1}{\varepsilon}.$$

Together with Lemma B.9, this implies

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \ge \frac{\lambda^{(L)} q}{4}. \tag{54}$$

If $M = \infty$, all low-fidelity observations of arm k have identical distributions under ν_0 and ν_k. Lemma B.9 therefore gives

$$2\Delta^2 \mathbb{E}_{\nu_0}\left[N_{k,T_\Lambda}^{(H)}\right] \ge \frac{1}{2}\log\frac{1}{\varepsilon},$$

and hence

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \ge \frac{\lambda^{(H)} q}{4} \ge \frac{1}{8}\Psi_k\left(\log\frac{1}{\varepsilon}\right).$$

Thus suppose $M < \infty$.

If

$$\mathbb{E}_{\nu_0}\left[N_{k,T_\Lambda}^{(L)}\right] \ge \frac{M+1}{4},$$

then

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \geq \frac{\lambda^{(L)}(M+1)}{4}.$$

Combining this with equation 54 gives

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \geq \frac{\lambda^{(L)}}{4}\max\{M+1, q\} \geq \frac{1}{8}\Psi_k\left(\log\frac{1}{\varepsilon}\right).$$

Otherwise,

$$\mathbb{E}_{\nu_0}\left[N_{k,T_\Lambda}^{(L)}\right] < \frac{M+1}{4}.$$

Let σ denote the transcript censored immediately before the $(M+1)$-st low-fidelity observation of arm k, or at the terminal time if that observation is never reached, and define

$$B := \left\{N_{k,T_\Lambda}^{(L)} \leq M\right\}, \qquad A' := A \cap B.$$

By Markov's inequality,

$$P_{\nu_0}(B^c) \leq \frac{\mathbb{E}_{\nu_0}[N_{k,T_\Lambda}^{(L)}]}{M+1} < \frac{1}{4}.$$

Hence

$$P_{\nu_0}(A') \geq 1 - \varepsilon - \frac{1}{4} \geq \frac{11}{16}, \qquad P_{\nu_k}(A') \leq \varepsilon.$$

The event A' is measurable with respect to the censored transcript. Before σ, all low-fidelity observations of arm k have identical distributions under ν_0 and ν_k, so only high-fidelity observations of arm k contribute to the transcript KL divergence. Therefore,

$$D_{\mathrm{KL}}\left(P_{\nu_0}^{\sigma} \| P_{\nu_k}^{\sigma}\right) = 2\Delta^2\mathbb{E}_{\nu_0}\left[N_{k,\sigma}^{(H)}\right].$$

Applying data processing to A' gives

$$2\Delta^2\mathbb{E}_{\nu_0}\left[N_{k,\sigma}^{(H)}\right] \geq \frac{11}{16}\log\frac{1}{\varepsilon} - \log 2 \geq \frac{1}{4}\log\frac{1}{\varepsilon},$$

where the last inequality uses $\varepsilon \leq 1/16$. Thus

$$\mathbb{E}_{\nu_0}\left[N_{k,T_\Lambda}^{(H)}\right] \geq \mathbb{E}_{\nu_0}\left[N_{k,\sigma}^{(H)}\right] \geq \frac{q}{8}.$$

Consequently,

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \geq \frac{\lambda^{(H)}q}{8} \geq \frac{1}{8}\Psi_k\left(\log\frac{1}{\varepsilon}\right),$$

which proves the claim equation 53. □

Proof of Theorem 5.6. Let

$$r_\Lambda := \max_{\nu\in\mathcal{V}} \mathbb{E}_\nu[R(\Lambda)]$$

and define

$$\Phi(\ell) := \sum_{k=2}^{K} \Delta_k \Psi_k(\ell).$$

Suppose for contradiction that

$$r_\Lambda < \frac{1}{16}\Phi(\log z_\Lambda).$$

By the definition of Ψ_k,

$$\Psi_k(\ell) \leq \lambda^{(H)}\frac{\ell}{\Delta_k^2},$$

and therefore

$$\Phi(\ell) \leq \lambda^{(H)}\ell\sum_{k=2}^{K}\frac{1}{\Delta_k} = \Xi_\Delta\ell.$$

Since

$$z_\Lambda = \frac{\Lambda \Delta_{\min}}{\Xi_\Delta},$$

it follows that

$$\varepsilon := \frac{4r_\Lambda}{\Lambda \Delta_{\min}} < \frac{\log z_\Lambda}{4z_\Lambda} \le \frac{1}{16},$$

where the last inequality holds for $z_\Lambda \ge 16$.

For each $k \ge 2$, let

$$A_k := \left\{ W_k(\Lambda) \le \frac{\Lambda}{2} \right\}.$$

Under ν_0,

$$R(\Lambda) = \sum_{j=2}^{K} \Delta_j W_j(\Lambda),$$

so

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \le \frac{r_\Lambda}{\Delta_k}.$$

Markov's inequality therefore gives

$$P_{\nu_0}(A_k^c) \le \frac{2r_\Lambda}{\Lambda \Delta_k} \le \varepsilon.$$

Under ν_k, every arm other than k has high-fidelity gap at least Δ_k. By the assumed total-cost condition, on A_k the cost spent on arms other than k is at least $\Lambda/2 - \lambda^{(H)}$.

The condition $z_\Lambda \ge 16$ and $\Delta_{\min} \le 1/4$ imply $\lambda^{(H)} \le \Lambda/4$, so

$$P_{\nu_k}(A_k) \le \frac{4r_\Lambda}{\Lambda \Delta_k} \le \varepsilon.$$

Applying Lemma B.10 to A_k for every $k \ge 2$, yielding

$$\mathbb{E}_{\nu_0}[W_k(\Lambda)] \ge \frac{1}{8} \Psi_k \left(\log \frac{1}{\varepsilon} \right).$$

Summing over suboptimal arms gives

$$r_\Lambda \ge \mathbb{E}_{\nu_0}[R(\Lambda)] \ge \frac{1}{8} \Phi \left(\log \frac{1}{\varepsilon} \right).$$

Since $\varepsilon < \frac{\log z_\Lambda}{4z_\Lambda}$, we have

$$\log \frac{1}{\varepsilon} > \log \left(\frac{4z_\Lambda}{\log z_\Lambda} \right) \ge \frac{1}{2} \log z_\Lambda,$$

where the last inequality holds for $z_\Lambda \ge 16$.

Finally, for any $0 < s \le 1$,

$$\max \left\{ M_k + 1, \frac{s\ell}{\Delta_k^2} \right\} \ge s \max \left\{ M_k + 1, \frac{\ell}{\Delta_k^2} \right\},$$

and hence

$$\Psi_k(s\ell) \ge s \Psi_k(\ell).$$

Therefore,

$$\Phi \left(\frac{1}{2} \log z_\Lambda \right) \ge \frac{1}{2} \Phi(\log z_\Lambda),$$

which gives

$$r_\Lambda \ge \frac{1}{16} \Phi(\log z_\Lambda),$$

contradicting the assumption. This proves the result. □

Corollary B.11 (Power-law mismatch certificate)**.** *Under the conditions of Theorem 5.6, suppose*

$$U_k(n) = Cn^{1-r}, \qquad 0 < r \le 1,$$

for every $k \ge 2$. Then

$$M_k = \left\lfloor \left(\frac{C}{\Delta_k} \right)^{1/r} \right\rfloor,$$

and

$$\max_{\nu \in \mathcal{V}} \mathbb{E}_\nu[R(\Lambda)] \ge \frac{1}{16} \sum_{k=2}^{K} \Delta_k \min \left\{ \lambda^{(L)} \max \left\{ \left\lfloor \left(\frac{C}{\Delta_k} \right)^{1/r} \right\rfloor + 1, \frac{\log z_\Lambda}{\Delta_k^2} \right\}, \lambda^{(H)} \frac{\log z_\Lambda}{\Delta_k^2} \right\}. \quad (55)$$

Proof. For $U_k(n) = Cn^{1-r}$,

$$\Delta_k s \le C s^{1-r} \quad \Longleftrightarrow \quad s \le \left(\frac{C}{\Delta_k} \right)^{1/r},$$

so

$$M_k = \left\lfloor \left(\frac{C}{\Delta_k} \right)^{1/r} \right\rfloor.$$

Substituting this expression into Theorem 5.6 gives equation 55. □

We call a policy uniformly efficient on $\mathcal{M}(U)$ if, for every instance $\nu \in \mathcal{M}(U)$ and every $a > 0$,

$$\mathbb{E}_\nu[R(\Lambda)] = o(\Lambda^a) \qquad \text{as } \Lambda \to \infty.$$

Proposition B.12 (Residual-floor asymptotic lower bound)**.** *Let*

$$b_k := \inf_{n \ge 1} \frac{U_k(n)}{n}.$$

Consider a budget-exhausting policy that is uniformly efficient on $\mathcal{M}(U)$, and an instance with a unique optimal arm in which each suboptimal arm has a stationary low-fidelity offset

$$\mu_k^{(L)}(\tau) = \mu_k^{(H)} + \beta_k, \qquad |\beta_k| \le b_k.$$

Then

$$\liminf_{\Lambda \to \infty} \frac{\mathbb{E}_\nu[R(\Lambda)]}{\log \Lambda} \ge \sum_{k \ne k^\star} \Delta_k \min \left\{ \frac{2\lambda^{(L)}}{(\Delta_k - \beta_k - b_k)_+^2}, \frac{2\lambda^{(H)}}{\Delta_k^2} \right\}. \quad (56)$$

Here, $x_+ := \max\{x, 0\}$, with the convention that $1/0^2 = +\infty$.

Proof. Fix $k \ne k^\star$ and $\epsilon > 0$, and let $d = \Delta_k + \epsilon$. Construct an alternative instance by raising arm k's high-fidelity mean by d and its low-fidelity mean by

$$h := (d - \beta_k - b_k)_+.$$

The resulting low-to-high offset remains in $[-b_k, b_k]$, so both instances satisfy the same certificate because

$$n b_k \le U_k(n), \quad \forall n \ge 1.$$

The adaptive Gaussian chain rule gives

$$D_{\mathrm{KL}}\big(P_\nu^\Lambda \| P_{\nu'}^\Lambda\big) = \frac{h^2}{2} \mathbb{E}_\nu N_{k,T_\Lambda}^{(L)} + \frac{d^2}{2} \mathbb{E}_\nu N_{k,T_\Lambda}^{(H)}.$$

Using the event

$$\left\{ W_k(\Lambda) \le \frac{\Lambda}{2} \right\},$$

uniform efficiency and budget exhaustion imply the standard Lai–Robbins change-of-measure constraint

$$\liminf_{\Lambda\to\infty} \frac{h^2\mathbb{E}_\nu N_{k,T_\Lambda}^{(L)} + d^2\mathbb{E}_\nu N_{k,T_\Lambda}^{(H)}}{2\log\Lambda} \geq 1.$$

Minimizing the cost required to satisfy this information constraint gives

$$\liminf_{\Lambda\to\infty} \frac{\mathbb{E}_\nu[W_k(\Lambda)]}{\log\Lambda} \geq \min\left\{\frac{2\lambda^{(L)}}{h^2}, \frac{2\lambda^{(H)}}{d^2}\right\},$$

where the first term is interpreted as $+\infty$ when $h = 0$. Letting $\epsilon \downarrow 0$, multiplying by Δ_k, and summing over suboptimal arms proves the claim. □

The exact-proxy result previously stated in the main text is the special case $\beta_k = 0$. The optimistic extreme $\beta_k = b_k$ yields the stronger denominator $(\Delta_k - 2b_k)_+^2$, showing that persistent residual mismatch can force logarithmic high-fidelity information even when the proxy is stationary.

Corollary B.13 (A continuation-active matching regime)**.** *Fix $\Lambda > 1$ and $0 < \Delta \leq 1/4$. For $K = 2$ and $\lambda^{(L)} = 1$, set*

$$\delta = T_\Lambda^{-2}, \qquad \gamma = \Delta/4, \qquad \eta = 1/8, \qquad M = N_\gamma(\delta), \qquad S_0 = \lambda^{(H)} = 5M.$$

Consider the hard family $\mathcal{V} = \{\nu_0, \nu_2\}$ from Lemma B.9, with gap Δ and certificates $U_1(n) = 0$ and $U_2(n) = \Delta\min\{n, M\}$. If $z_\Lambda = \Lambda\Delta^2/(5M) \geq 16$, then

$$\max_{\nu\in\mathcal{V}} \mathbb{E}_\nu[R_{\mathrm{TACC}}(\Lambda)] \leq 176 \inf_\pi \max_{\nu\in\mathcal{V}} \mathbb{E}_\nu[R_\pi(\Lambda)].$$

The infimum ranges over budget-feasible policies whose total query cost is at least $\Lambda - \lambda^{(H)}$ almost surely under both instances. Moreover, arm 2 belongs to $\mathcal{C}_\gamma^{\mathrm{d}}(\delta)$ in ν_0.

Proof. The certificates give $B_1(n) = 0$, $B_2(n) = \Delta\min\{1, M/n\}$, and masking time $M_2 = M$. In ν_0, the effective low-fidelity gap is $-\Delta$ for $n \leq M$ and $\Delta - 2\Delta M/n$ thereafter. It is nondecreasing, with $\tau_2(\gamma) = 4M$. For every $M \leq n < 4M$, the admissible lookahead $s = 6M - n$ gives

$$B_2(n) - B_2(n+s) > \frac{\Delta}{4} - \frac{\Delta}{6} = \frac{\Delta}{12} > \frac{\Delta}{16} = 2\eta\gamma.$$

Thus arm 2 is in $\mathcal{C}_\gamma^{\mathrm{d}}(\delta)$. In ν_2, the only suboptimal arm is arm 1, whose effective gap is constantly Δ, so it belongs to $\mathcal{A}_\gamma(\delta)$. Certificate stability holds in both worlds, and $S_0\lambda^{(L)} = \lambda^{(H)}$. Theorem 5.3, together with $R(\Lambda) \leq \Lambda\Delta$ and $\delta\Lambda \leq 1$, yields

$$\max_{\nu\in\mathcal{V}} \mathbb{E}_\nu[R_{\mathrm{TACC}}(\Lambda)] \leq \Delta(11M + 1) + \delta\Lambda\Delta \leq 11\Delta(M + 1).$$

The parameter choices imply $1 \leq \log z_\Lambda/\Delta^2 < M$, hence $\Psi_2(\log z_\Lambda) = M + 1$. Theorem 5.6 therefore gives

$$\inf_\pi \max_{\nu\in\mathcal{V}} \mathbb{E}_\nu[R_\pi(\Lambda)] \geq \frac{\Delta(M+1)}{16}.$$

TACC also satisfies the stated cost condition: it stops only when the selected query is unaffordable. Combining the two bounds proves the claim. □

The certificate and costs are fixed before each run but may vary with Λ; this match is restricted to the stated finite-budget family.

C A COMPLEMENTARY ORACLE-BOUND REFERENCE SCHEME WITHOUT CONTINUATION ASSUMPTIONS

The main-text analysis of TACC is continuation specific: it uses additional structural conditions to certify when a short post-threshold low-fidelity continuation block is provably aligned with the elimination of suboptimal arms. To separate this layer from the underlying complexity of fidelity

Algorithm 2 Resolution-Dependent Fidelity Elimination (RDFE)

Require: number of arms K, costs $(\lambda^{(L)}, \lambda^{(H)})$, confidence level δ, budget Λ
query each arm once at low fidelity and once at high fidelity
initialize empirical means, counts, and $\Lambda_{\mathrm{used}} \leftarrow K(\lambda^{(L)} + \lambda^{(H)})$
set $r \leftarrow 1$, $\mathcal{S}_1 \leftarrow [K]$, and let $\widehat{k}_{\mathrm{out}}$ be the warm-start LCB leader
while $|\mathcal{S}_r| > 1$ **do**
  $\varepsilon_r \leftarrow 2^{-r}$
  for each arm $k \in \mathcal{S}_r$ **do**
    choose $m_{k,r} \in \arg\min\{\mathsf{c}_k^{(L)}(\varepsilon_r; \Lambda), \mathsf{c}^{(H)}(\varepsilon_r; \Lambda)\}$
    if $m_{k,r} = L$ **then**
      query arm k at low fidelity until $G_\Lambda(N_k^{(L)}, \delta) + B_k(N_k^{(L)}) \le \varepsilon_r/8$, or until the remaining budget cannot pay for the next low-fidelity query
      if the stopping condition is not reached **then**
        return $\widehat{k}_{\mathrm{out}}$
      set $\mathrm{UCB}_{k,r} \leftarrow \hat{\mu}_k^{(L)} + G_\Lambda(N_k^{(L)}, \delta) + B_k(N_k^{(L)})$ and $\mathrm{LCB}_{k,r} \leftarrow \hat{\mu}_k^{(L)} - G_\Lambda(N_k^{(L)}, \delta) - B_k(N_k^{(L)})$
    else
      query arm k at high fidelity until $G_\Lambda(N_k^{(H)}, \delta) \le \varepsilon_r/8$, or until the remaining budget cannot pay for the next high-fidelity query
      if the stopping condition is not reached **then**
        return $\widehat{k}_{\mathrm{out}}$
      set $\mathrm{UCB}_{k,r} \leftarrow \hat{\mu}_k^{(H)} + G_\Lambda(N_k^{(H)}, \delta)$ and $\mathrm{LCB}_{k,r} \leftarrow \hat{\mu}_k^{(H)} - G_\Lambda(N_k^{(H)}, \delta)$
  choose $\hat{k}_r \in \arg\max_{k \in \mathcal{S}_r} \mathrm{LCB}_{k,r}$
  update $\mathcal{S}_{r+1} \leftarrow \{k \in \mathcal{S}_r : \mathrm{UCB}_{k,r} \ge \mathrm{LCB}_{\hat{k}_r,r} - \varepsilon_r/4\}$
  set $\widehat{k}_{\mathrm{out}} \in \arg\max_{k \in \mathcal{S}_{r+1}} \mathrm{LCB}_{k,r}$
  $r \leftarrow r + 1$
return the unique arm in $\mathcal{S}_r$ if $|\mathcal{S}_r| = 1$; otherwise return $\widehat{k}_{\mathrm{out}}$

choice itself, we record here a complementary oracle-bound reference scheme based on phase-wise resolution targeting.

This reference scheme is not meant to replace the main-text policy. Its role is to isolate the resolution-dependent certification principle induced by the selected-average bound model without invoking the continuation-specific assumptions used in the TACC analysis. In this sense, it serves as a structural reference point for the self-improving two-fidelity problem, while TACC remains the primary online policy.

For a target resolution $\varepsilon > 0$, define the low-fidelity certification cost for arm k by

$$\mathsf{c}_k^{(L)}(\varepsilon; \Lambda) := \lambda^{(L)} \min\Big\{n \ge 1 : G_\Lambda(n, \delta) + B_k(n) \le \varepsilon/8\Big\}, \tag{57}$$

and the high-fidelity certification cost by

$$\mathsf{c}^{(H)}(\varepsilon; \Lambda) := \lambda^{(H)} \min\Big\{n \ge 1 : G_\Lambda(n, \delta) \le \varepsilon/8\Big\}. \tag{58}$$

With the convention that the minimum of an empty set is $+\infty$, define the oracle cost complexity

$$\mathsf{c}_k^\star(\varepsilon; \Lambda) := \min\big\{\mathsf{c}_k^{(L)}(\varepsilon; \Lambda),\, \mathsf{c}^{(H)}(\varepsilon; \Lambda)\big\}. \tag{59}$$

The reference scheme uses a dyadic resolution schedule

$$\varepsilon_r := 2^{-r}, \qquad r = 1, 2, \ldots, \tag{60}$$

and maintains an active set $\mathcal{S}_r$ of arms not yet certified as suboptimal.

The scheme is deliberately phase based. Its role is not to provide the most competitive finite-budget online policy, but to isolate the resolution-dependent complexity object induced by the shrinking selected-average mismatch bound.

The analysis reuses the confidence-validity event $\mathcal{E}_\Lambda$ from Lemma B.1. Unlike the main-text TACC analysis, it does not invoke continuation-specific assumptions. The proof has three steps: once a phase is completed, every active arm has confidence width at most $\varepsilon_r/4$; any suboptimal arm with gap at least ε_r is eliminated at that phase; and the additional phase cost is charged to the corresponding oracle certification cost $\mathsf{c}_k^\star(\varepsilon_r;\Lambda)$.

For bookkeeping, call a phase r *completed* if the algorithm finishes processing every arm in $\mathcal{S}_r$ and performs the elimination update that defines $\mathcal{S}_{r+1}$. Let R_Λ denote the number of completed phases under budget Λ. For each arm k and completed phase $r \le R_\Lambda$, let

$$\Delta\Lambda_{k,r} \tag{61}$$

denote the additional cost spent on arm k during phase r, with $\Delta\Lambda_{k,r} := 0$ if $k \notin \mathcal{S}_r$.

Lemma C.1 (Phase-end confidence width for RDFE)**.** *Suppose phase r is completed and the event $\mathcal{E}_\Lambda$ holds. Then for every arm $k \in \mathcal{S}_r$,*

$$\mu_k^{(H)} \in [\mathrm{LCB}_{k,r}, \mathrm{UCB}_{k,r}] \qquad \textit{and} \qquad \mathrm{UCB}_{k,r} - \mathrm{LCB}_{k,r} \le \varepsilon_r/4. \tag{62}$$

Consequently,

$$\mathrm{LCB}_{k,r} \ge \mu_k^{(H)} - \varepsilon_r/4, \qquad \mathrm{UCB}_{k,r} \le \mu_k^{(H)} + \varepsilon_r/4. \tag{63}$$

Proof. The claim follows directly from the stopping rule of Algorithm 2 and Lemma B.1. If low fidelity is chosen, then

$$G_\Lambda(N_k^{(L)}, \delta) + B_k(N_k^{(L)}) \le \varepsilon_r/8,$$

so the resulting interval width is at most $\varepsilon_r/4$. If high fidelity is chosen, then

$$G_\Lambda(N_k^{(H)}, \delta) \le \varepsilon_r/8,$$

and the same conclusion follows. Coverage is inherited from Lemma B.1. □

Lemma C.2 (Phase elimination for RDFE)**.** *Suppose phase r is completed and the event $\mathcal{E}_\Lambda$ holds. If $\Delta_k \ge \varepsilon_r$, then arm k is eliminated by the end of phase r.*

Proof. By Lemma C.1,

$$\mathrm{UCB}_{k,r} \le \mu_k^{(H)} + \varepsilon_r/4.$$

For the optimal arm $k^\star$,

$$\mathrm{LCB}_{k^\star,r} \ge \mu^\star - \varepsilon_r/4.$$

Since $\hat{k}_r$ maximizes the lower confidence bound over $\mathcal{S}_r$,

$$\mathrm{LCB}_{\hat{k}_r,r} \ge \mathrm{LCB}_{k^\star,r} \ge \mu^\star - \varepsilon_r/4,$$

and hence

$$\mathrm{LCB}_{\hat{k}_r,r} - \varepsilon_r/4 \ge \mu^\star - \varepsilon_r/2.$$

If $\Delta_k = \mu^\star - \mu_k^{(H)} \ge \varepsilon_r$, then

$$\mathrm{UCB}_{k,r} \le \mu_k^{(H)} + \varepsilon_r/4 = \mu^\star - \Delta_k + \varepsilon_r/4 \le \mu^\star - \frac{3\varepsilon_r}{4} < \mu^\star - \frac{\varepsilon_r}{2} \le \mathrm{LCB}_{\hat{k}_r,r} - \varepsilon_r/4.$$

Thus arm k fails the survival test and is eliminated. □

Lemma C.3 (Single-phase additional cost upper bound for RDFE)**.** *Fix a completed phase $r \le R_\Lambda$ and an arm $k \in \mathcal{S}_r$. Then*

$$\Delta\Lambda_{k,r} \le \mathsf{c}_k^\star(\varepsilon_r;\Lambda). \tag{64}$$

Proof. If phase r chooses low fidelity, the algorithm continues querying until

$$G_\Lambda(N_k^{(L)}, \delta) + B_k(N_k^{(L)}) \le \varepsilon_r/8.$$

Hence the additional low-fidelity cost incurred during phase r is at most the from-scratch low-fidelity certification cost $\mathsf{c}_k^{(L)}(\varepsilon_r;\Lambda)$. If phase r chooses high fidelity, the same argument gives an upper bound by $\mathsf{c}^{(H)}(\varepsilon_r;\Lambda)$. Since the phase chooses the cheaper of the two fidelities, the additional phase cost is at most $\mathsf{c}_k^\star(\varepsilon_r;\Lambda)$. □

Assumption C.4 (Dyadic regularity of oracle costs for the reference scheme)**.** *For each suboptimal arm $k \neq k^\star$, there exists a constant $C_{\mathrm{dyad}} > 0$ such that*

$$\sum_{r \le r_k} \mathsf{c}_k^\star(\varepsilon_r; \Lambda) \le C_{\mathrm{dyad}}\, \mathsf{c}_k^\star(\varepsilon_{r_k}; \Lambda), \qquad r_k := \min\{r : \varepsilon_r \le \Delta_k/2\}. \tag{65}$$

Remark C.5. Assumption C.4 is a dyadic summability condition on the oracle certification cost. It holds whenever $\mathsf{c}_k^\star(\varepsilon; \Lambda)$ grows regularly as ε decreases, for example under the polynomial-type oracle costs induced by the high-fidelity term $\mathsf{c}^{(H)}(\varepsilon; \Lambda) \asymp \lambda^{(H)} \log(T_\Lambda)/\varepsilon^2$ and by the common power-mismatch bound regime $B_k(n) \lesssim \Gamma_k n^{-\alpha}$. Its role is to exclude highly oscillatory oracle-cost profiles across dyadic resolutions, so that the cumulative phase cost is controlled by the terminal certification scale.

Theorem C.6 (Regret guarantee for RDFE)**.** *Under Assumption 3.1 and Assumption C.4, the reference scheme Algorithm 2 satisfies, with probability at least $1 - \delta$,*

$$R_{\mathrm{RDFE}}(\Lambda) \le C_1 \sum_{k \neq k^\star} \Delta_k\, \mathsf{c}_k^\star(\Delta_k/4; \Lambda) + C_2 K(\lambda^{(L)} + \lambda^{(H)}), \tag{66}$$

for some absolute constants $C_1, C_2 > 0$.

Proof. Fix a suboptimal arm $k \neq k^\star$, and let

$$r_k := \min\{r : \varepsilon_r \le \Delta_k/2\}.$$

By Lemma C.2, arm k cannot survive beyond phase r_k. Therefore its total additional phase-based exploration cost is bounded by

$$\sum_{r \le r_k} \Delta\Lambda_{k,r}.$$

Applying Lemma C.3 phase by phase gives

$$\sum_{r \le r_k} \Delta\Lambda_{k,r} \le \sum_{r \le r_k} \mathsf{c}_k^\star(\varepsilon_r; \Lambda).$$

Assumption C.4 then implies

$$\sum_{r \le r_k} \mathsf{c}_k^\star(\varepsilon_r; \Lambda) \le C\, \mathsf{c}_k^\star(\Delta_k/4; \Lambda).$$

Since every unit of cost spent on arm k contributes at most Δ_k to the cost-weighted pseudo-regret,

$$R_k(\Lambda) \le C\, \Delta_k\, \mathsf{c}_k^\star(\Delta_k/4; \Lambda).$$

Summing over all suboptimal arms yields the main term. The warm start contributes at most one low-fidelity and one high-fidelity pull per arm, which gives the additive lower-order term. □

D Additional Experimental Details

This appendix gives the experimental details behind Section 6. Appendix D.1 gives the synthetic data-generation procedure, and Appendix D.2 studies how the decay rate of the low-fidelity mismatch bound affects TACC. Appendix D.3 gives additional details on the NAS-Bench-201 experiment and offline certificate calibration. Appendices D.4–D.6 report hyperparameter sensitivity, mismatch-certificate misspecification, and an exploratory arm-specific switching threshold. Appendix D.7 specifies the LLM-as-a-judge task, prompt-policy arms, weak-judge model, and baseline conventions. Appendix D.8 reports the residual-mismatch regime for multiple high-fidelity costs. Appendix D.9 studies a vanishing-mismatch regime in which the low–high gap eventually disappears. Appendix D.10 reports empirical weak-judge accuracy and low–high gap measurements from trained checkpoints, and Appendix D.11 uses those checkpoints in a bandit experiment.

Table 2: Synthetic low-fidelity generation parameters. The initial mismatch magnitude is $|\mu_k^{(L)}(1) - \mu_k^{(H)}|$.

Setting	ζ	Initial mismatch magnitude	n_0	Distribution of a_k
Set A	0.2	$[0.08, 0.20]$	8	$[-0.60, -0.24] \cup [0.24, 0.60]$
Set B	1	$[0.4, 1]$	8	$[-3.00, -1.20] \cup [1.20, 3.00]$

D.1 Synthetic Bandit Generation Details

For each arm k, the synthetic low-fidelity trajectory is generated as

$$\mu_k^{(L)}(n) = \mu_k^{(H)} + a_k(n + n_0)^{-r}, \qquad a_k = s_k \zeta m_k (n_0 + 1)^r, \tag{67}$$

where $s_k \in \{-1, +1\}$ and $m_k \sim \mathcal{U}(0.4, 1)$. This normalization gives

$$\mu_k^{(L)}(1) - \mu_k^{(H)} = s_k \zeta m_k, \tag{68}$$

so ζ controls the initial mismatch scale. The cumulative mismatch certificate $U_k(n)$ and the selected-average bound $B_k(n)$ are computed from the resulting trajectory.

Main experimental settings. For Set A, we use $K = 200$, $\mu_k^{(H)} \sim \mathcal{U}(0.1, 0.9)$, $(\lambda^{(L)}, \lambda^{(H)}) = (1, 10)$, $\gamma = 0.063$, and $S_0 = 10$. For Set B, we use $K = 500$, $\mu_k^{(H)} \sim \mathcal{N}(0, 1)$, $(\lambda^{(L)}, \lambda^{(H)}) = (1, 50)$, $\gamma = 0.141$, and $S_0 = 50$. Both settings use $n_0 = 8$, $r = 0.5$, and $\eta = 10^{-4}$.

Table 2 summarizes the mismatch parameters. With $n_0 = 8$ and $r = 0.5$, $(n_0 + 1)^r = 3$, so a_k is arm dependent rather than fixed across arms.

D.2 Sensitivity to the Low-Fidelity Improvement Rate

Using Bandit Set A, we vary the convergence rate r in the mismatch bound $B_k(n) = O(n^{-r})$. A larger r means that the low-fidelity source improves faster. Figure 3 shows that TACC benefits from faster decay: the low-fidelity interval tightens more quickly, so more suboptimal arms can be certified through cheap continuation rather than high-fidelity escalation. This behavior is consistent with the mechanism analyzed in Section 5.

D.3 NAS-Bench-201 Protocol and Offline Certificate Calibration

In the NAS-Bench-201 experiment (Dong & Yang, 2020), each arm is a neural architecture. Its partial-training trajectory serves as the improving low-fidelity source, while its fully trained validation performance defines the high-fidelity target. We use disjoint architecture sets for proxy training, certificate calibration, hyperparameter tuning, and final evaluation. The main experiments use $K = 20$, $\lambda^{(L)} = 1$, $\lambda^{(H)} = 40$, and final budget $\Lambda = 6000$, with results averaged over three independent architecture splits and 9,600 runs per method.

The mismatch certificate is constructed before bandit execution from paired partial- and full-training trajectories on the calibration architectures. On held-out architectures, the resulting certificate achieves approximately 96–98% simultaneous coverage. The certificate remains fixed throughout the bandit run, so no high-fidelity observations from the evaluation budget are used to construct or update it. This experiment is therefore an offline replay study of real improvement trajectories rather than online certificate recalibration.

D.4 Hyperparameter Sensitivity

We evaluate sensitivity to η, S_0, δ, and γ on ImageNet16-120. The default setting is $\gamma = 0.0277$, $\eta = 0.0025$, $S_0 = 30$, and $\delta = 0.05$. Each entry in Table 3 reports the mean cost-weighted regret over 200 runs while varying one parameter at a time.

Smaller η allows more continuation, while increasing S_0 helps until the benefit saturates around $S_0 = 30$. The effect of δ is mild. Increasing γ does not consistently improve performance, so $\gamma \to \infty$

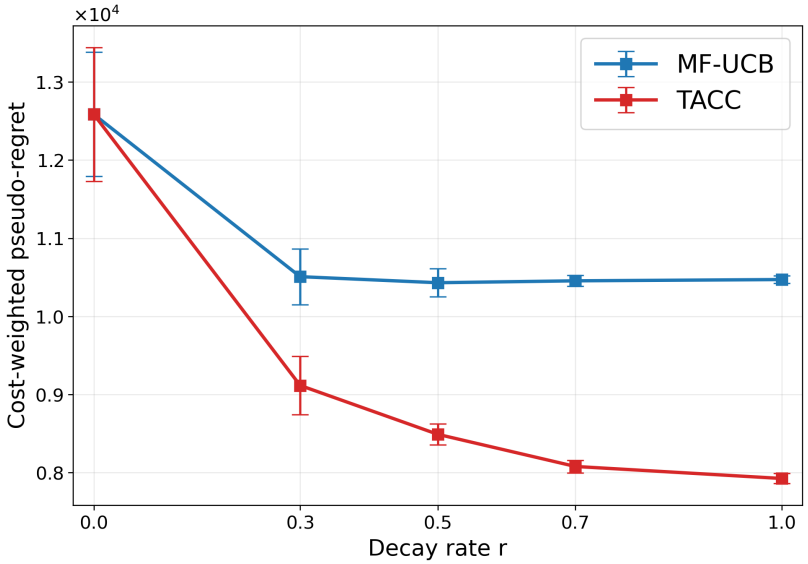


Figure 3: Sensitivity of TACC to the decay rate r in the low-fidelity mismatch bound $B_k(n) = O(n^{-r})$. Larger r corresponds to faster low-fidelity improvement.

Table 3: Hyperparameter sensitivity on ImageNet16-120. Bold values denote the default setting.

Parameter	Tested values	Mean regret
η	**0.0025**, 0.05, 0.10	**120.58**, 125.58, 138.13
S_0	0, 16, **30**, 39	138.13, 124.16, **120.58**, 120.58
δ	0.01, **0.05**, 0.10	121.94, **120.58**, 120.03
γ	**0.0277**, 0.0554, 0.1071	**120.58**, 122.52, 133.39

is not optimal. In practice, we recommend $S_0 \leq \lfloor \lambda^{(H)}/\lambda^{(L)} \rfloor$ and selecting η and γ using held-out calibration trajectories.

D.5 Robustness to Mismatch-Certificate Misspecification

We first perturb the calibrated certificate as

$$\widetilde{B}_k(n) = \alpha B_k(n), \tag{69}$$

and evaluate TACC on ImageNet16-120 using 600 matched runs for each value of α. The reference certificate $B_k(n)$ is pre-calibrated using 25% of the architectures. A violation occurs when the supplied certificate is too small to cover the true low–high mismatch.

Underestimating the certificate ($\alpha < 1$) reduces apparent regret but produces substantially more certificate violations. Overestimating it ($\alpha > 1$) is safer but requires more low-fidelity queries and increases regret. Thus, conservative misspecification primarily affects efficiency, whereas aggressive underestimation can compromise certificate validity.

Decay-rate misspecification. We also fix the true decay rate at $r_{\text{true}} = 0.5$ and vary the rate r_{alg} used by TACC to construct the supplied mismatch bound and continuation gain. Table 5 reports the resulting mean regret.

The matched model performs best. Assuming slower improvement is conservative but delays certification, with $r_{\text{alg}} = 0$ causing the largest degradation. Moderate rate misspecification produces a smaller deterioration.

Table 4: Sensitivity to multiplicative mismatch-certificate misspecification on ImageNet16-120. Values are reported over 600 matched runs per value of α.

α	Cost-weighted regret	Low-fidelity queries	Violation (%)
0.50	106.74 ± 20.23	222.70 ± 56.35	25.38 ± 13.25
0.75	113.58 ± 20.57	256.95 ± 60.07	5.36 ± 6.58
1.00	121.93 ± 22.21	294.15 ± 63.02	0.73 ± 1.78
1.50	141.68 ± 21.95	364.27 ± 62.62	0.037 ± 0.111
2.00	161.59 ± 23.63	426.62 ± 62.85	0.002 ± 0.024

Table 5: Sensitivity to misspecification of the assumed improvement rate with $r_{\mathrm{true}} = 0.5$.

r_{alg}	r_{true}	Mean regret
0.0	0.5	10913.91
0.3	0.5	7785.99
0.5	0.5	7028.87
0.7	0.5	7276.16
1.0	0.5	7465.63

D.6 Exploratory Arm-Specific Switching Thresholds

As an exploratory extension, we set

$$\gamma_k = B_k(1)\sqrt{\frac{\lambda^{(L)}}{\lambda^{(H)}}}. \tag{70}$$

A larger initial mismatch therefore moves an arm into the switching stage earlier. For the uniform-threshold comparator, all arms use

$$\gamma = \max_j B_j(1)\sqrt{\frac{\lambda^{(L)}}{\lambda^{(H)}}}. \tag{71}$$

We evaluate the two choices over 450 matched runs from three independent architecture splits on CIFAR-10-valid.

The arm-specific threshold lowers cost-weighted regret by 13.76, with a clustered 95% confidence interval of $[7.43, 21.14]$. It uses approximately 120 additional low-fidelity evaluations while avoiding 2.95 high-fidelity evaluations, with simple regret nearly unchanged. This experiment is intended as a preliminary empirical study of arm-adaptive switching rather than a theoretically optimal threshold construction.

D.7 LLM-as-a-Judge Experimental Protocol

The LLM-as-a-judge experiment treats a bandit arm as a prompt-policy rather than as an individual NLI example. A prompt-policy maps an NLI input to a candidate label. The high-fidelity target is the expected verifier correctness of that policy, while the low-fidelity source is a cheaper weak-judge feedback process associated with the same policy.

Task and verifier. Each evaluation item consists of a premise p, a hypothesis h, and a verifier label $y \in \{\text{ENTAILMENT}, \text{NEUTRAL}, \text{CONTRADICTION}\}$. A prompt-policy arm k takes (p, h) as input and returns a predicted label $\hat{y}_k(p, h)$. The high-fidelity verifier reward is

$$Y_k^{(H)} = \mathbf{1}\{\hat{y}_k(p, h) = y\}, \qquad \mu_k^{(H)} = \mathbb{E}\left[Y_k^{(H)}\right], \tag{72}$$

where the expectation is over the shared evaluation pool. In the bandit simulation, a high-fidelity query samples an evaluation item and observes verifier correctness for the selected prompt-policy. Regret is always computed with respect to the high-fidelity means $\mu_k^{(H)}$.

Table 6: Uniform and arm-specific switching thresholds on CIFAR-10-valid over 450 matched runs.

Uniform γ	Per-arm γ_k
135.58 ± 44.86	121.82 ± 35.99

Table 7: Controlled prompt-policy arms used in the main LLM-as-a-Judge experiment. All policies receive the same NLI item and return one label in {ENTAILMENT, NEUTRAL, CONTRADICTION}.

Arm	Prompt-policy description
`lexical_overlap`	Uses lexical overlap and containment between the premise and hypothesis; predicts entailment under high overlap or containment, contradiction under explicit negation cues, and neutral otherwise.
`hypothesis_only`	Ignores the premise and predicts from the hypothesis alone; this represents a spurious-prior policy induced by hypothesis-side artifacts.
`neutral_prior`	Favors the neutral label; this is a conservative weak distractor arm.
`contrastive_check`	Compares contradiction evidence and entailment evidence, returns the label with stronger support, and defaults to neutral when neither side is decisive.

Prompt-policy arms. The prompt-policy pool includes `direct_label`, `cot_reasoning`, `role_nli_expert`, `self_check`, `contrastive_check`, `lexical_overlap`, `hypothesis_only`, and `neutral_prior`. The main residual-mismatch experiment uses the four-arm subset `lexical_overlap`, `hypothesis_only`, `neutral_prior`, and `contrastive_check`; the checkpoint-based bandit experiment uses the five-arm subset specified in Appendix D.11. Table 7 summarizes the main four arms.

Remark D.1. We use a four-arm benchmark in the main text to keep the main effect identifiable: the arms are fixed across all methods, so differences in regret come from fidelity allocation rather than from scaling to a large arm set. The inclusion of `hypothesis_only` and lexical or prior-based policies is natural for NLI, where hypothesis-only baselines and annotation artifacts are known to create spurious predictive signals (Poliak et al., 2018; Gururangan et al., 2018).

Other prompt-policies are used in the checkpoint-based diagnostic. `direct_label` uses number mismatch, negation conflict, overlap, and containment; `cot_reasoning` checks contradiction cues before entailment cues and then applies overlap thresholds; `role_nli_expert` is a more conservative expert-style rule; and `self_check` takes a majority vote over several rule-based policies, preferring neutral on ties.

Controlled weak-judge model. For arm k, after its n-th low-fidelity query,

$$\mu_k^{(L)}(n) = \mu_k^{(H)} + b_k + a_k(n + n_0)^{-r}. \tag{73}$$

Here n is the arm-specific low-fidelity count. The transient term $a_k(n + n_0)^{-r}$ models improvement with repeated low-fidelity use. The residual term b_k models persistent weak-judge bias relative to the verifier. In each seed, the low-fidelity process is fixed before running any bandit algorithm and is shared by all methods. The scale parameter ζ controls the initial low–high mismatch; in the residual-mismatch experiments, we set $b_k \equiv 0.05$ to enforce a persistent residual gap. When bounded observations are required, the resulting means are clipped to $[0, 1]$ before sampling rewards.

The residual term b_k is included because weak judges need not converge exactly to the target evaluator. MT-Bench and Chatbot Arena document judge biases such as position, verbosity, self-enhancement, and limited reasoning ability (Zheng et al., 2023). Follow-up empirical studies also find that fine-tuned judges may lag stronger judges in generalizability, fairness, and adaptability (Huang et al., 2025). These findings motivate a model in which calibration improves the weak judge but need not eliminate the low–high gap.

The power-law transient in equation 73 is a modeling choice rather than an empirical claim about all judges. It gives a simple way to interpolate between an initially biased weak judge and a better but still imperfect judge.

Table 8: Residual-mismatch experiment at final budget $\Lambda = 128000$. Values are mean cost-weighted pseudo-regret $\pm$ standard error over 200 seeds. Lower is better; the best method for each high-fidelity cost is bolded.

$\lambda^{(H)}$	TACC	DNC	MF-UCB	UCB	Weak-Fixed	Low-Improving
200	$\mathbf{3277.7 \pm 243.2}$	3894.0 ± 266.6	4317.0 ± 276.8	4264.1 ± 289.1	7331.9 ± 119.3	6827.5 ± 85.7
500	$\mathbf{4023.4 \pm 247.3}$	5201.0 ± 289.7	5359.1 ± 281.8	5083.2 ± 286.8	7501.6 ± 127.7	6808.4 ± 83.4
1000	$\mathbf{4688.6 \pm 253.6}$	5532.7 ± 277.0	5334.4 ± 251.1	5190.7 ± 253.9	7058.4 ± 101.2	6668.6 ± 50.9

Table 9: Paired final-budget comparisons for the residual-mismatch experiment with $\lambda^{(H)} = 500$. The reported value is TACC minus baseline; negative values favor TACC. Confidence intervals are paired 95% intervals over 200 common seeds.

Baseline	Mean diff.	95% CI	Favors TACC?
DNC	-1177.6	$[-1916.6, -438.6]$	Yes
MF-UCB	-1335.7	$[-2107.3, -564.2]$	Yes
UCB	-1059.9	$[-1826.3, -293.4]$	Yes
Weak-Fixed	-3478.2	$[-4043.0, -2913.5]$	Yes

Checkpoint weak judges. The empirical-gap experiments replace equation 73 with trained weak-judge checkpoints. The weak judge is `cross-encoder/nli-deberta-v3-small`. We use cumulative continued fine-tuning at checkpoint sizes $q \in \{32, 64, 128, 256, 512, 1024, 2048, 4096, 8192\}$, with maximum length 512, 2 epochs per stage, learning rate 2×10^{-6}, batch size 8, evaluation batch size 32, and seed 0. Policy names are excluded from the weak-judge input to avoid policy-name leakage. The weak judge sees the task, premise, hypothesis, and candidate answer, and predicts whether the candidate answer matches the gold NLI relation.

For each checkpoint size q and arm k, we estimate an empirical low-fidelity mean $\widehat{\mu}_{k,q}^{(L)}$ from weak-judge scores on the shared evaluation pool, and compare it with the verifier mean $\mu_k^{(H)}$. This gives a finite checkpoint-based low–high gap profile.

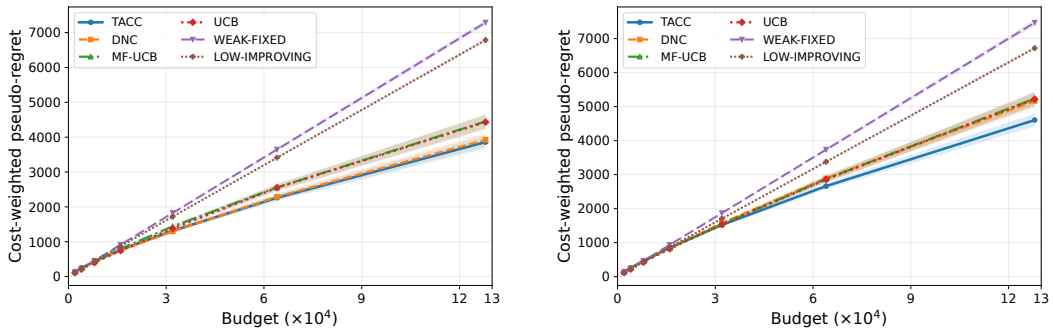
Baselines. All algorithms are evaluated on the same high-fidelity means, low-fidelity process, budget sequence, and random seeds when paired comparisons are reported. UCB uses only verifier feedback. MF-UCB uses a fixed multi-fidelity switching rule. DNC is the direct no-continuation ablation of TACC: it uses the same confidence bounds based on the dynamic mismatch bound and the same threshold γ, but after the low-fidelity radius crosses γ, it switches to high fidelity without checking whether additional low-fidelity samples would further reduce $B_k(n)$. WEAK-FIXED uses only the initial weak judge. LOW-IMPROVING always uses the improving low-fidelity source and never falls back to the verifier; it is included as a diagnostic baseline.

D.8 RESIDUAL-MISMATCH REGIME

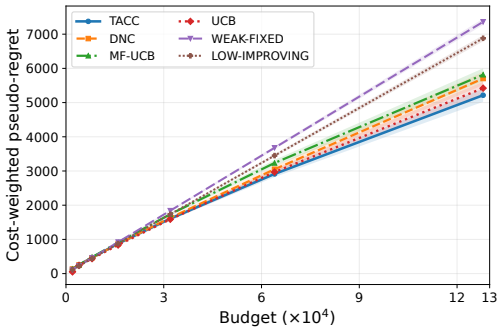
The main text reports the representative residual-mismatch experiment with $\lambda^{(H)} = 500$. Here we give the full results for $\lambda^{(H)} \in \{200, 500, 1000\}$. We use the four prompt-policy arms from Table 7, $\lambda^{(L)} = 1$, $r = 0.75$, $\zeta = 0.4$, $b_k \equiv 0.05$, $\gamma = 0.025$, $\eta = 10^{-4}$, and $S_0 = 128$. Budgets range up to $\Lambda = 128000$, and results are averaged over 200 seeds.

Table 8 reports final-budget regret for all methods and all three high-fidelity costs. Figure 4 shows the corresponding six-method regret curves. TACC remains competitive across all three high-fidelity costs. The residual mismatch makes LOW-IMPROVING unreliable, while the improvement of the weak judge still leaves room for bounded continuation to reduce costly high-fidelity confirmations.

For the representative cost $\lambda^{(H)} = 500$, Table 9 gives paired final-budget comparisons. The TACC-minus-baseline intervals are below zero for all primary baselines, including DNC; this isolates the contribution of the continuation block in the residual-mismatch setting.



(a) $\lambda^{(H)} = 200$ (b) $\lambda^{(H)} = 500$



(c) $\lambda^{(H)} = 1000$

Figure 4: Residual-mismatch regret curves for $\lambda^{(H)} \in \{200, 500, 1000\}$. The weak judge improves with use but retains persistent mismatch from the verifier.

Figure 4 gives the full regret curves for all six methods and all three high-fidelity costs.

Figure 5 reports the continuation counts for the representative $\lambda^{(H)} = 500$ setting. The continuation budget remains bounded, consistent with the design of TACC as a post-threshold correction rather than a replacement for high-fidelity verification.

D.9 Vanishing-Mismatch Regime: $B_k(n) \to 0$

We next remove the residual term and consider

$$\mu_k^{(L)}(n) = \mu_k^{(H)} + a_k(n + n_0)^{-r}. \tag{74}$$

This is the regime where the low–high mismatch shrinks toward zero. In this case, low fidelity eventually becomes aligned with the high-fidelity target, so Low-Improving can become competitive at large budgets.

Table 10 reports final-budget regret for the main methods and Weak-Fixed, and Figure 6 shows the corresponding six-method regret curves. TACC has small final regret across all high-fidelity costs. We interpret this experiment as a theory-aligned diagnostic rather than the hardest practical LLM-judge regime: once the cheap source effectively converges to the target, a low-only improving strategy can also become competitive.

D.10 Empirical Weak-Judge Gap

The residual-mismatch and vanishing-mismatch experiments control the low-fidelity process analytically. We also evaluate a checkpoint-based construction. Rather than defining low fidelity by algebraically perturbing the high-fidelity mean, we train weak judges at increasing data scales and estimate the low–high gap from their policy-level behavior.

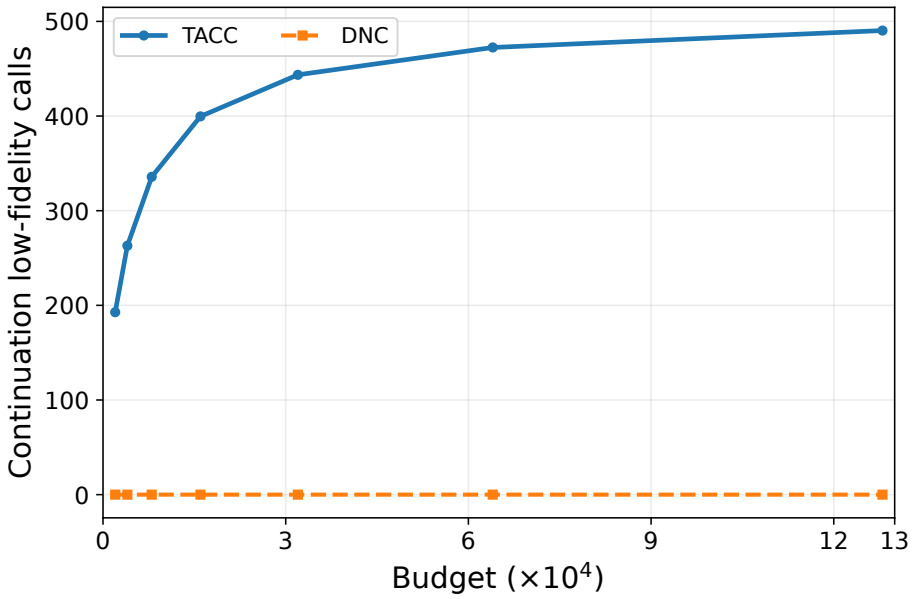


Figure 5: Continuation calls in the residual-mismatch experiment with $\lambda^{(H)} = 500$.

Table 10: Vanishing-mismatch experiment at final budget $\Lambda = 128000$. Values are mean cost-weighted pseudo-regret $\pm$ standard error. Lower is better; the best reported method for each high-fidelity cost is bolded.

$\lambda^{(H)}$	TACC	DNC	MF-UCB	UCB	Weak-Fixed
200	$\mathbf{75.7 \pm 0.5}$	677.5 ± 28.6	1503.1 ± 37.1	1506.1 ± 39.4	6503.0 ± 0.0
500	$\mathbf{143.0 \pm 0.3}$	657.7 ± 113.4	2687.8 ± 74.0	2711.1 ± 52.0	6503.0 ± 0.0
1000	$\mathbf{261.8 \pm 0.4}$	598.7 ± 187.9	3719.0 ± 94.4	3790.0 ± 64.1	6503.0 ± 0.0

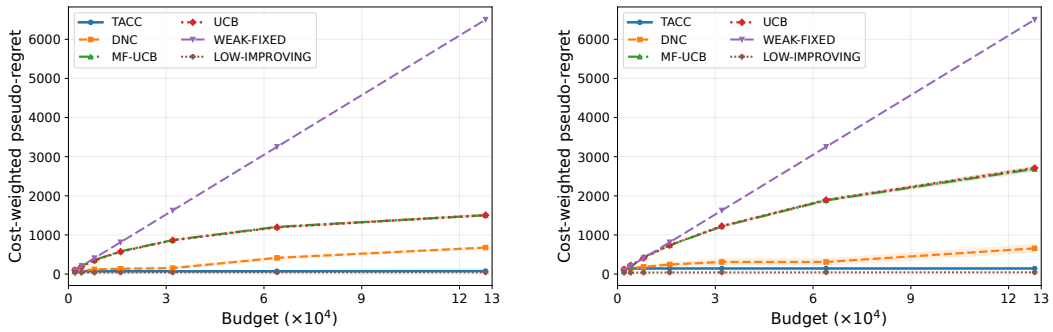
Figure 7 and Table 11 report weak-judge prediction accuracy across checkpoints. Accuracy increases from 0.4881 at $q = 32$ to 0.5669 at $q = 512$, 0.6344 at $q = 1024$, 0.7040 at $q = 2048$, and 0.7619 at $q = 8192$. The mean loss also decreases from 0.7892 at $q = 32$ to 0.6922 at $q = 512$, 0.6096 at $q = 2048$, and 0.5291 at $q = 8192$.

For each prompt-policy arm and weak-judge scale q, we compute an empirical low-fidelity policy mean and compare it with the verifier mean. Figure 8 and Table 12 report the resulting policy-level low–high gap. The mean absolute gap decreases from 0.1755 at $q = 32$ to 0.0452 at $q = 512$ and 0.0318 at $q = 1024$, with later checkpoints fluctuating at a lower level. These measurements support the modeling premise that cheap feedback can be initially unreliable but becomes more aligned with the high-fidelity verifier after additional calibration.

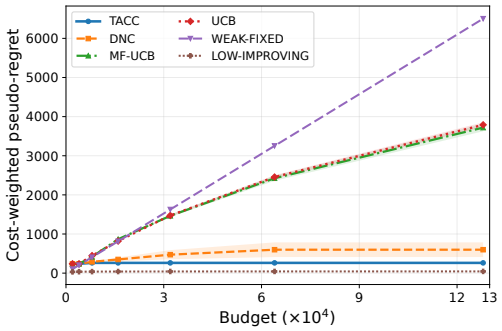
D.11 Checkpoint-Based Bandit Experiment

Finally, we use the empirical weak-judge checkpoints from Appendix D.10 inside the bandit simulation. This experiment is checkpoint-based rather than rate-aligned: the weak-judge process is finite, capped at $q_{\max} = 2048$, and not exactly monotone. The regret curve should therefore be read as a finite-budget diagnostic, not as an asymptotic-rate demonstration.

We report a five-arm setting with `lexical_overlap`, `direct_label`, `self_check`, `cot_reasoning`, and `hypothesis_only`. The high-fidelity means are 0.5568, 0.5376, 0.5376, 0.5364, and 0.5060, respectively. We use $\lambda^{(H)} = 500$, $\lambda^{(L)} = 1$, $\gamma = 0.03$, $\eta = 10^{-4}$, $S_0 = 128$, $q_{\max} = 2048$, budgets up to $\Lambda = 256000$, and 120 seeds.



(a) $\lambda^{(H)} = 200$ (b) $\lambda^{(H)} = 500$



(c) $\lambda^{(H)} = 1000$

Figure 6: Vanishing-mismatch regret curves for $\lambda^{(H)} \in \{200, 500, 1000\}$. The low-fidelity source eventually aligns with the high-fidelity target.

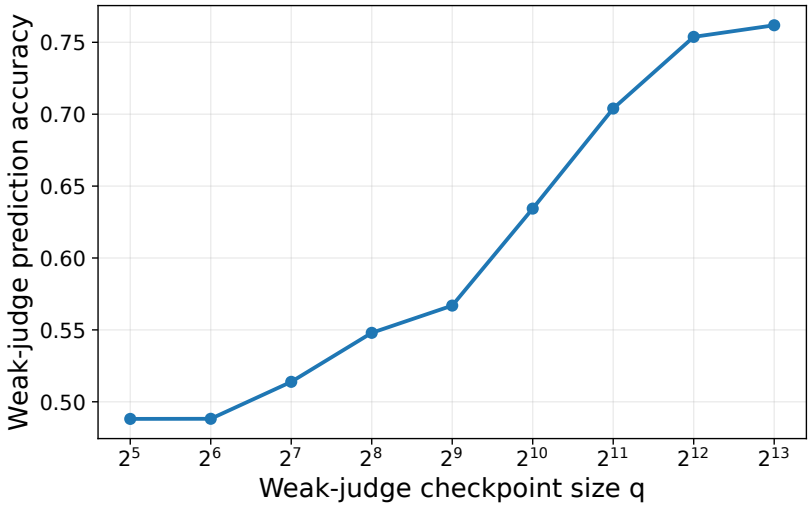


Figure 7: Weak-judge prediction accuracy as a function of checkpoint size q.

Table 13 gives paired final-budget comparisons, and Figure 9 shows the corresponding regret curves. TACC significantly improves over UCB and WEAK-FIXED. It also has lower mean regret than MF-UCB and DNC, although the paired confidence intervals overlap zero. LOW-IMPROVING is better at the final budget, which is expected when later weak-judge checkpoints are substantially more aligned with the verifier. We therefore interpret this experiment conservatively: it supports the

Table 11: Weak-judge prediction accuracy across checkpoint sizes.

q	Prediction accuracy
32	0.4881
64	0.4882
128	0.5139
256	0.5480
512	0.5669
1024	0.6344
2048	0.7040
4096	0.7538
8192	0.7619

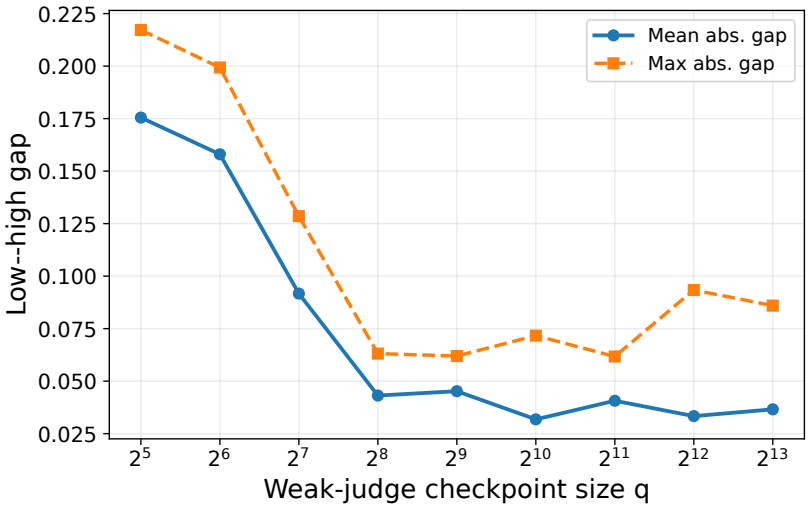


Figure 8: Policy-level low–high gap as a function of weak-judge checkpoint size q.

practical relevance of improving weak-judge feedback, while the synthetic and residual-mismatch experiments provide cleaner evidence for the continuation mechanism.

Figure 10 separates the aggregate regret curve into normalized regret, high-fidelity calls, and continuation calls. These diagnostics help identify whether differences come from verifier usage, continuation behavior, or the cumulative scale of the budget.

The LLM-as-a-Judge experiments should not be read as visual estimates of an asymptotic regret exponent. The synthetic and vanishing-mismatch experiments test the shrinking-mismatch bound mechanism most directly. The residual-mismatch and checkpoint-based LLM experiments instead test finite-budget robustness under weak-judge mismatch, checkpoint caps, close top arms, and nonmonotone trained checkpoints.

Table 12: Empirical weak-judge gap measurements on the evaluation split.

q	Mean abs. gap	Max abs. gap
32	0.1755	0.2172
64	0.1580	0.1993
128	0.0917	0.1286
256	0.0431	0.0631
512	0.0452	0.0620
1024	0.0318	0.0717
2048	0.0407	0.0617
4096	0.0333	0.0934
8192	0.0366	0.0860

Table 13: Checkpoint-based bandit experiment at final budget $\Lambda = 256000$. The reported value is TACC minus baseline; negative values favor TACC. Confidence intervals are paired 95% intervals over 120 common seeds.

Baseline	Mean diff.	95% CI	Favors TACC?
DNC	-737.5	$[-1606.2, 131.1]$	No
MF-UCB	-427.7	$[-1125.4, 270.1]$	No
UCB	-2356.3	$[-2831.0, -1881.5]$	Yes
Weak-Fixed	-2803.1	$[-3273.9, -2332.3]$	Yes
Low-Improving	1989.8	$[1521.2, 2458.3]$	No

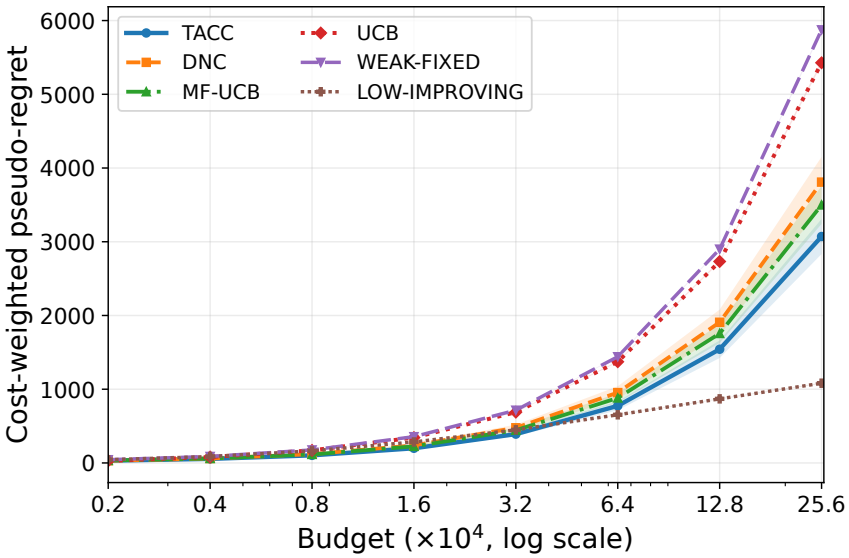
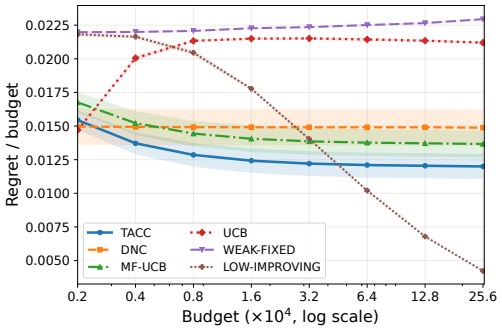
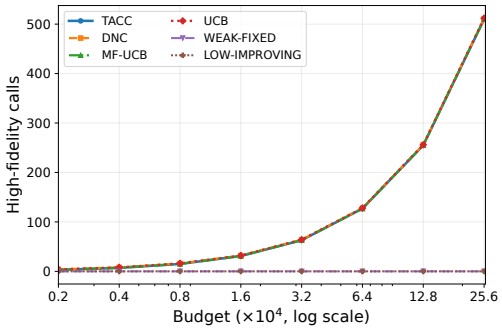


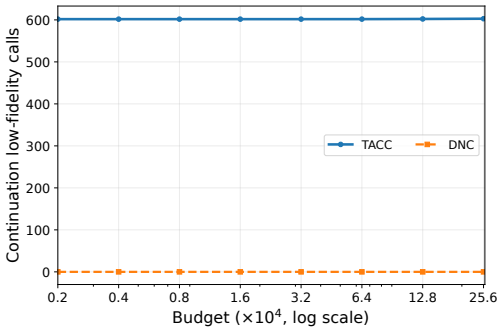
Figure 9: Checkpoint-based bandit experiment with $\lambda^{(H)} = 500$ and $q_{\max} = 2048$. The low-fidelity process is estimated from trained weak-judge checkpoints rather than imposed algebraically from the high-fidelity means.



(a) Normalized regret



(b) High-fidelity calls



(c) Continuation calls

Figure 10: Additional diagnostics for the checkpoint-based bandit experiment.