COST-AWARE MULTI-OBJECTIVE BANDITS: THEORY AND APPLICATION TO BUDGETED LLM CONFIGURATION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating large language model (LLM) configurations is challenging because evaluation budgets are limited, evaluation costs differ across configurations, and several objectives must be traded off simultaneously. We formulate LLM configuration evaluation as a cost-aware multi-objective bandit problem, in which each evaluation incurs a configuration-dependent cost and yields a noisy vector-valued outcome. Within this framework, we study two fundamental problems: online configuration selection and Pareto configuration identification. For online selection, we propose a hypervolume-based UCB algorithm and establish a regret bound of order $O\big(\sum_{i\neq i^\star} \frac{\log B}{\Delta_i}\big)$, where B is the evaluation budget, $i^\star$ is the configuration with the highest hypervolume per unit cost, and Δ_i is the efficiency gap of configuration i. This bound retains the logarithmic budget dependence of classical single-objective budgeted bandits. For fixed-budget Pareto identification, we develop a cost-aware empirical gap elimination algorithm and prove that its error probability is of order $O\big(\exp(-\frac{B}{H_{\boldsymbol{\mu},c}})\big)$, where $H_{\boldsymbol{\mu},c}$ is a cost-aware complexity that depends on the configuration costs and the Pareto classification gaps. This rate decays exponentially in the budget and recovers the standard Pareto set identification guarantee when all costs are identical. Experiments on real LLM configuration data and synthetic instances demonstrate the benefit of cost-aware allocation.

1 INTRODUCTION

In real-world applications of large language models (LLMs), practitioners must choose among configurations, such as base models, prompt templates, decoding parameters, and retrieval settings, under limited evaluation budgets. Evaluating a configuration may involve model inference, long-form generation, LLM-as-a-judge scoring, or human assessment, so the cost of an evaluation can differ substantially across configurations (Zheng et al., 2023; Chang et al., 2024; Li et al., 2025). For example, larger models and chain-of-thought prompting may improve task performance but typically consume more tokens (Wei et al., 2022; Shi et al., 2024b; Zhou et al., 2024). Moreover, LLM evaluation is inherently multi-objective: beyond task accuracy, practitioners also care about latency and efficiency (Liang et al., 2023; Chang et al., 2024). Reliable configuration evaluation therefore requires allocating a limited budget adaptively across candidate configurations while accounting for both heterogeneous evaluation costs and multiple quality objectives.

A straightforward approach is to evaluate every candidate on the entire validation set and select the one with the best empirical performance. This protocol mirrors standard benchmark-based evaluation, in which models are assessed on a predefined collection of tasks, examples, and metrics (Liang et al., 2023; Chang et al., 2024). Although simple and statistically reliable, exhaustive evaluation is increasingly impractical for modern LLM applications: assessing many configurations requires a large number of inference queries and incurs considerable monetary cost, token consumption, and computational overhead (Zheng et al., 2023; Li et al., 2025; Feng et al., 2025).

Recent studies have therefore explored adaptive evaluation strategies. Zhou et al. (2025) formulate language model evaluation as a bandit-based adaptive sampling problem, and Shi et al. (2024a) apply best-arm identification to select high-performing prompts under a limited evaluation budget. These studies show that allocating more evaluations to promising candidates, rather than evaluat-

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.

ing all candidates uniformly on all validation instances, can substantially reduce evaluation cost. Nevertheless, existing adaptive evaluation methods have three limitations. **(a)** They often ignore configuration-dependent evaluation costs, even though differences in model scale, reasoning strategy, and retrieval pipeline lead to substantial variation in token consumption and computational overhead (Chen et al., 2024; Ong et al., 2025). **(b)** Most formulations optimize a single scalar performance metric or ranking criterion, whereas practical LLM evaluation must balance multiple objectives, such as task effectiveness, robustness, and latency (Liang et al., 2023; Chang et al., 2024). **(c)** Bandit-based LLM evaluation methods typically adopt a pure-exploration formulation that assesses only the finally selected configuration (Polo et al., 2024; Zhou et al., 2025). They therefore overlook the quality of intermediate decisions, which directly determines the cumulative utility of continuously operating evaluation systems (Lattimore & Szepesvári, 2020; Ong et al., 2025; Ashizawa et al., 2025).

Motivated by these limitations, we formulate LLM configuration evaluation as a cost-aware multi-objective bandit problem with a finite evaluation budget. Each configuration is modeled as an arm, and each evaluation consumes a configuration-dependent cost while producing stochastic vector-valued feedback. This formulation captures two central characteristics of practical LLM evaluation: a limited budget must be allocated across configurations with different costs, and multiple performance criteria must be optimized simultaneously. To the best of our knowledge, this is the first study of cost-aware multi-objective bandits. We consider two complementary learning goals, namely online configuration selection via regret minimization and Pareto-optimal configuration identification under a fixed budget. Our main contributions are as follows.

- We propose CoHV-UCB, a hypervolume-based UCB algorithm for cost-aware online configuration selection that optimizes a hypervolume-per-cost index to balance multi-objective performance against evaluation cost. We prove that its regret is bounded by $O\big(\sum_{i\neq i^\star} \frac{\log B}{\Delta_i}\big)$, where $i^\star$ is the configuration with the highest hypervolume per unit cost and Δ_i is the efficiency gap of configuration i. This bound matches the single-objective budgeted bandit result in both budget and gap dependence (Ding et al., 2013).
- We develop CoPSI, a cost-aware Pareto configuration identification algorithm that allocates samples according to evaluation costs and progressively eliminates configurations using empirical Pareto gaps. We prove that its error probability is bounded by $O\big(\exp(-\frac{B}{H_{\boldsymbol{\mu},c}})\big)$, where the complexity $H_{\boldsymbol{\mu},c}$ depends on both the configuration costs and the Pareto classification gaps. This bound recovers the standard rate for fixed-budget Pareto set identification when all configurations have identical costs (Kone et al., 2024).
- We evaluate both algorithms on synthetic instances and on LLM configuration evaluation tasks that span different models, prompts, and decoding parameters. Compared with cost-insensitive, single-objective, and non-adaptive baselines, our methods demonstrate the benefit of jointly accounting for evaluation cost and multiple performance objectives.

2 Related Work

Budgeted bandits. Classical bandit research considers two learning objectives: regret minimization, which evaluates the quality of decisions made throughout the learning process (Lai & Robbins, 1985; Auer, 2002), and best-arm identification, which evaluates only the final recommendation (Audibert & Bubeck, 2010; Gabillon et al., 2012; Jamieson et al., 2014). Ding et al. (2013) formalized the budgeted multi-armed bandit with stochastic costs, which captures the reward–cost trade-off through the expected reward-per-cost criterion, and established logarithmic regret guarantees. Subsequent work has extended this line in several directions. Xia et al. (2015) developed Thompson sampling algorithms for budgeted bandits with random costs; Xia et al. (2016) studied the multiple-play setting, in which several arms are selected in each round under a shared budget; and Li & Xia (2017) considered budgeted bandits with infinitely many arms, highlighting the additional exploration cost of searching a large action space. Bandits with knapsacks further generalize budgeted bandits to multiple resource constraints (Badanidiyuru et al., 2018; 2014; Immorlica et al., 2022; Sankararaman & Slivkins, 2018; Liu et al., 2022). These studies provide a principled foundation for decision-making under resource constraints, but they focus on scalar rewards.

Multi-objective bandits. Multi-objective bandits address sequential decision-making with vector-valued rewards. Existing studies have considered several performance criteria, including scalarized

regret (Drugan & Nowe, 2013), Pareto regret (Lu et al., 2019), hypervolume regret (Zhang, 2024), and Pareto set identification (PSI) (Auer et al., 2016). One line of work optimizes the hypervolume indicator: Zhang & Golovin (2020) introduced random hypervolume scalarizations for multi-objective black-box optimization, and Zhang (2024) later established optimal sublinear hypervolume regret for this approach. Another closely related line studies PSI, whose goal is to identify all non-dominated arms. Auer et al. (2016) studied PSI under stochastic bandit feedback, Kone et al. (2023) developed adaptive algorithms for relaxed fixed-confidence PSI, and Kone et al. (2024) proposed successive-rejects-based algorithms for the fixed-budget setting; Crepon et al. (2024) further examined the sequential learning of Pareto fronts. Beyond bandits, Qiu et al. (2024) extend multi-objective learning to Markov decision processes and study provably efficient multi-objective reinforcement learning. Although these studies lay important foundations for multi-objective exploration, they constrain the total number of samples and do not account for arm-dependent evaluation costs.

Bandits for LLMs. Bandit learning has recently been applied to several LLM-related problems. For prompt optimization, ProTeGi (Pryzant et al., 2023) uses bandit selection within beam search, Lin et al. (2024) and Wu et al. (2024) employ neural bandits for instruction optimization and exemplar selection, respectively, and MASPOB (Hong et al., 2026) extends UCB-style exploration to multi-agent systems. For selection and evaluation, Shi et al. (2024a) cast prompt selection as fixed-budget best-arm identification, Poon et al. (2025) study online LLM selection with cost–quality trade-offs, and Zhou et al. (2025) combine bandits with low-rank matrix factorization to evaluate configurations from fewer observations. However, most of these methods optimize a single objective or focus on offline best-arm identification. In contrast, our work jointly models configuration-dependent costs and multiple objectives, and provides guarantees for both online selection and fixed-budget identification.

3 Preliminaries

We consider LLM configuration evaluation with multiple competing objectives under a total evaluation budget $B > 0$. At each round $t = 1, 2, \ldots$, the learner selects a configuration $i_t \in [K] := \{1, \ldots, K\}$, evaluates it on a validation sample (e.g., a question from GSM8K (Cobbe et al., 2021) or PIQA (Bisk et al., 2020)), and observes a reward vector $\boldsymbol{r}_{i_t,t} = (r^{(1)}_{i_t,t}, \ldots, r^{(D)}_{i_t,t}) \in [0,1]^D$. Each coordinate corresponds to a normalized objective, such as task accuracy or inference efficiency, and is oriented so that larger values are preferred. The reward vectors of configuration i are i.i.d. with unknown mean $\boldsymbol{\mu}_i = (\mu_i^{(1)}, \ldots, \mu_i^{(D)}) = \mathrm{E}[\boldsymbol{r}_{i,t}]$. Throughout, $\mathbb{I}\{\cdot\}$ denotes the indicator function, and $x^+ = \max\{x, 0\}$ and $a \wedge b = \min\{a, b\}$ for real numbers x, a, b.

Each evaluation also consumes part of the budget, which may represent API expenditure, token consumption, or human-annotation effort. Although such quantities could also be included in the reward vector, it is the evaluation cost that determines how many evaluations can be performed. We adopt a different cost model for each of the two learning settings. In online selection, the cost may be random: evaluating configuration i at round t incurs a cost $c_{i,t} \in [\lambda, 1]$ with mean $\mu_i^c = \mathrm{E}[c_{i,t}]$, where $0 < \lambda \leq 1$. In fixed-budget Pareto identification, we consider deterministic and known costs $c_i \in [\lambda, 1]$, which is appropriate when the cost of an evaluation can be estimated in advance, for example under a fixed inference protocol or a standardized human-evaluation procedure. We summarize the feedback model as follows.

Assumption 1 (Bounded stochastic feedback) *For each configuration $i \in [K]$, the reward vectors $\boldsymbol{r}_{i,t} \in [0,1]^D$ obtained from successive evaluations of i are i.i.d. with mean $\boldsymbol{\mu}_i$. In online selection, the costs $c_{i,t} \in [\lambda, 1]$ of successive evaluations of i are i.i.d. with mean μ_i^c for a known $\lambda \in (0, 1]$; in Pareto identification, every evaluation of i incurs the known cost $c_i \in [\lambda, 1]$.*

3.1 Budgeted Online Configuration Selection

Let $C_t = \sum_{s=1}^{t} c_{i_s,s}$ denote the cumulative cost after t evaluations, so that the number of evaluations completed within budget B is $T(B) = \max\{t \geq 0 : C_t \leq B\}$. We measure the multi-objective quality of a configuration by the hypervolume indicator with the origin as the reference point. For a single point $\boldsymbol{x} \in [0,1]^D$, the hypervolume is the volume of the box $[\mathbf{0}, \boldsymbol{x}]$, i.e., $H(\boldsymbol{x}) = \mathrm{HV}(\boldsymbol{x}) = \prod_{d=1}^{D} x^{(d)}$, and we write $H_i = H(\boldsymbol{\mu}_i)$ for configuration i. Because H is strictly increasing on $(0,1]^D$, it is Pareto-compliant: if configuration j strictly dominates configuration i, then $H_j > H_i$. Unlike

a linear scalarization, it requires no preference weights and favors balanced improvements across objectives. To account for heterogeneous costs, we define the *hypervolume efficiency* $\nu_i = H_i/\mu_i^c$, the optimal configuration $i^\star = \operatorname{argmax}_{i\in[K]} \nu_i$, and the efficiency gap $\Delta_i = \nu_{i^\star} - \nu_i$. We consider two regret notions, the hypervolume-efficiency regret and the standard hypervolume regret:

$$R_B^{\mathrm{HV}} = \sum_{t=1}^{T(B)} \Delta_{i_t}, \quad R_B^{\mathrm{std}} = R_B^\star - \sum_{t=1}^{T(B)} H_{i_t},$$

where $R_B^\star$ denotes the maximum expected cumulative hypervolume attainable under budget B. Both notions measure cumulative performance under the budget constraint, whereas R_B^{HV} further emphasizes the hypervolume gained per unit cost.

3.2 Budgeted Pareto Configuration Identification

In the fixed-budget identification setting, the learner aims to recover the set of Pareto-optimal configurations. Following Kone et al. (2024), for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$ we write $\boldsymbol{u} \preceq \boldsymbol{v}$ if $u^{(j)} \le v^{(j)}$ for all $j \in [D]$, and $\boldsymbol{u} \prec \boldsymbol{v}$ if $u^{(j)} < v^{(j)}$ for all $j \in [D]$. Configuration j strictly dominates configuration i if $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j$, and the Pareto set is $\mathcal{P}^\star = \{i \in [K] : \nexists\, j \in [K] \text{ such that } \boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j\}$. The learner evaluates configurations with deterministic costs $\{c_i\}_{i=1}^K$. Once the budget is exhausted, outputs an estimate $\widehat{\mathcal{P}}_B$. Its performance is measured by the misidentification probability $\Pr(\widehat{\mathcal{P}}_B \neq \mathcal{P}^\star)$.

To characterize the difficulty of Pareto classification, following the gap construction of Empirical Gap Elimination (Kone et al., 2024), let $m(i,j) = \min_{d\in[D]}(\mu_j^{(d)} - \mu_i^{(d)})$ and $M(i,j) = \max_{d\in[D]}(\mu_i^{(d)} - \mu_j^{(d)})$ for any $i, j \in [K]$. Then, we define

$$\Delta_i^- = \max_{j\neq i} m(i,j), \quad \Delta_i^+ = \min_{j\neq i}\Big[M(i,j) \wedge \big(M(j,i)^+ + (\Delta_j^-)^+\big)\Big], \quad \gamma_i = \begin{cases}\Delta_i^-, & i \notin \mathcal{P}^\star,\\ \Delta_i^+, & i \in \mathcal{P}^\star.\end{cases}$$

Intuitively, γ_i measures how far configuration i is from a change in its Pareto status. If i is dominated, $\gamma_i = \Delta_i^-$ is the largest margin by which another configuration exceeds i simultaneously in all objectives. If i is Pareto-optimal, $M(i,j)$ is the margin by which i escapes domination by j, and $M(j,i)^+ + (\Delta_j^-)^+$ is the margin by which j escapes domination by i, enlarged by j's own domination margin; $\gamma_i = \Delta_i^+$ is the smallest of these margins over all competitors. We assume that $\gamma_i > 0$ for all i, since exact identification is not statistically well posed otherwise. Let $\gamma_{(1)} \le \cdots \le \gamma_{(K)}$ and $C(\mathcal{A}) = \sum_{i\in\mathcal{A}} c_i$ be the aggregate cost of $\mathcal{A} \subseteq [K]$. The problem complexity is defined as

$$H_{\boldsymbol{\mu},c} = \max_{\emptyset\neq\mathcal{A}\subseteq[K]} \frac{C(\mathcal{A})}{\gamma^2_{(|\mathcal{A}|)}},$$

which jointly captures the statistical difficulty of separating Pareto-optimal from dominated configurations and the cost of evaluating the configurations that remain unresolved.

4 Algorithms

In this section, we present two cost-aware algorithms for multi-objective bandits, one for online configuration selection and one for Pareto set identification.

4.1 Cost-Aware Hypervolume UCB

CoHV-UCB (Algorithm 1) follows the principle of optimism in the face of uncertainty. It constructs an optimistic reward vector for each configuration, scores this vector by the hypervolume indicator, and divides the score by a lower confidence bound on the expected cost, which yields an optimistic hypervolume-per-cost index. After pulling each arm once, the algorithm maintains, at each round t and for each arm i and objective $d \in [D]$, the estimates

$$\bar{r}_{i,t}^{(d)} = \frac{1}{n_{i,t}} \sum_{s<t:\, i_s=i} r_{i,s}^{(d)}, \qquad \bar{c}_{i,t} = \frac{1}{n_{i,t}} \sum_{s<t:\, i_s=i} c_{i,s}, \qquad \beta_{i,t} = \sqrt{\frac{\alpha \log T_B}{n_{i,t}}},$$

Algorithm 1 CoHV-UCB: Cost-aware HyperVolume UCB

Input: Budget B, number of arms K, cost lower bound λ, parameter $\alpha \geq 2$

Pull each arm $i \in [K]$ once and set $t = K$
while $\sum_{s=1}^{t} c_{i_s,s} \leq B$ **do**
  $t \leftarrow t + 1$
  for each arm $i \in [K]$ **do**
    $U_{i,t}^{(d)} = \min\{1, \bar{r}_{i,t}^{(d)} + \beta_{i,t}\}$ for $d \in [D]$, $\underline{c}_{i,t} = \max\{\lambda/2, \bar{c}_{i,t} - \beta_{i,t}\}$
    Compute the cost-aware hypervolume index $I_{i,t} = \mathrm{HV}(\mathbf{U}_{i,t})/\underline{c}_{i,t}$
  end for
  Pull $i_t = \mathrm{argmax}_{i \in [K]} I_{i,t}$, observe $(\boldsymbol{r}_{i_t,t}, c_{i_t,t})$, and update $\bar{r}_{i_t,t}$, $\bar{c}_{i_t,t}$, $n_{i_t,t}$
end while

where $n_{i,t} = \sum_{s<t} \mathbb{I}\{i_s = i\}$ is the number of previous evaluations of arm i, $\alpha \geq 2$, and $T_B = \lceil B/\lambda \rceil + 1$ is a deterministic upper bound on the number of evaluations, since each evaluation costs at least λ. The optimistic objective vector and the truncated lower confidence bound on the cost are

$$U_{i,t}^{(d)} = \min\big\{1,\ \bar{r}_{i,t}^{(d)} + \beta_{i,t}\big\},\ d \in [D], \qquad \underline{c}_{i,t} = \max\big\{\lambda/2,\ \bar{c}_{i,t} - \beta_{i,t}\big\}.$$

Clipping keeps $\mathbf{U}_{i,t}$ within $[0,1]^D$, and truncation at $\lambda/2$ prevents an underestimated cost. The algorithm then selects $i_t = \mathrm{argmax}_{i \in [K]} I_{i,t}$ with $I_{i,t} = \mathrm{HV}(\mathbf{U}_{i,t})/\underline{c}_{i,t}$, observes the reward vector and the cost, and updates its estimates until the budget is exhausted. Since the numerator of $I_{i,t}$ is an optimistic measure of multi-objective quality and the denominator is a conservative estimate of the cost, $I_{i,t}$ is an optimistic estimate of the hypervolume efficiency ν_i. Each round requires only $O(KD)$ operations, since the single-point hypervolume is a product of D coordinates.

The following regret guarantee retains the logarithmic budget dependence and the inverse-gap dependence of single-objective budgeted bandits (Ding et al., 2013), up to additional factors arising from the hypervolume objective.

Theorem 1 (Regret bound of CoHV-UCB) *Let $\alpha \geq 2$ and $C_{\lambda,H} = \frac{4D}{\lambda} + \frac{4}{\lambda^2}$. Then the expected hypervolume-efficiency regret of* CoHV-UCB *satisfies*

$$\mathbb{E}[R_B^{\mathrm{HV}}] = O\left(\sum_{i \neq i^\star} \frac{C_{\lambda,H}^2 \log B}{\Delta_i}\right).$$

Proof sketch and discussion. The analysis conditions on a good event on which all empirical reward and cost means lie within their confidence radii; by Hoeffding's inequality and a union bound over at most T_B sample sizes, this event fails with probability at most $2K(D+1)T_B^{1-2\alpha}$. On this event, the monotonicity of the hypervolume and the validity of the lower cost bound imply that the index of $i^\star$ is optimistic, i.e., $I_{i^\star,t} \geq \nu_{i^\star}$. For a suboptimal arm, the hypervolume of $[0,1]^D$-valued vectors is D-Lipschitz in the ℓ_∞ norm, and truncating the cost at $\lambda/2$ keeps the denominator bounded away from zero; together these yield $I_{i,t} \leq \nu_i + C_{\lambda,H}\beta_{i,t}$. Arm i can therefore be selected only while $C_{\lambda,H}\beta_{i,t} \geq \Delta_i$, which limits its number of pulls to $O(C_{\lambda,H}^2 \log T_B/\Delta_i^2)$, and $T_B = O(B/\lambda)$ gives the stated bound. The constant $C_{\lambda,H}$ quantifies the price of estimating a product of D objectives and of dividing by an estimated cost: the first term grows linearly with the number of objectives, and the second reflects the sensitivity of a ratio index to cost errors when costs can be as small as λ. For $D = 1$, $I_{i,t}$ reduces to an optimistic reward-to-cost index of the type analyzed by Ding et al. (2013), and the bound retains the same $\log B/\Delta_i$ dependence.

Theorem 1 bounds the online loss in hypervolume per unit cost. The following corollary translates this result into a bound on the standard budgeted hypervolume regret, which compares the cumulative hypervolume with that of the optimal budgeted policy.

Corollary 1 (Standard regret bound of CoHV-UCB) *Under the conditions of Theorem 1, the standard regret of* CoHV-UCB *satisfies* $\mathbb{E}[R_B^{\mathrm{std}}] = O\big(\sum_{i \neq i^\star} C_{\lambda,H}^2 \log B/\Delta_i\big)$.

Corollary 1 relies on the fact that no budgeted policy can collect more than $\nu_{i^\star}(B+1)$ expected hypervolume, since every unit of cost yields at most $\nu_{i^\star}$ hypervolume in expectation. The standard

Algorithm 2 CoPSI: Cost-aware Pareto Set Identification

Input: Budget B, arm costs $\{c_i\}_{i=1}^K$

Initialize $\mathcal{A}_1 = [K]$, $\widehat{\mathcal{P}} = \widehat{\mathcal{R}} = \emptyset$, $N_i = 0$ for all $i \in [K]$, and $L_{K,\lambda} = 1 + \sum_{\ell=2}^{K} \frac{1}{1+(\ell-1)\lambda}$
for $r = 1, \ldots, K-1$ **do**
  Set $C_r = \sum_{i \in \mathcal{A}_r} c_i$ and the cumulative sampling target $n_r = \lfloor B/(L_{K,\lambda} C_r) \rfloor$
  for $i \in \mathcal{A}_r$ **do**
    Pull arm i exactly $n_r - N_i$ additional times, set $N_i \leftarrow n_r$, and update $\widehat{\boldsymbol{\mu}}_{i,r}$
  end for
  Compute the empirical Pareto set $\widehat{\mathcal{P}}_r^{\text{emp}}$ and the gaps $\widehat{\gamma}_{i,r}$, $i \in \mathcal{A}_r$, by (1)
  Choose $e_r \in \operatorname{argmax}_{i \in \mathcal{A}_r} \widehat{\gamma}_{i,r}$, preferring $\mathcal{A}_r \setminus \widehat{\mathcal{P}}_r^{\text{emp}}$ among ties
  Accept e_r ($\widehat{\mathcal{P}} \leftarrow \widehat{\mathcal{P}} \cup \{e_r\}$) if $e_r \in \widehat{\mathcal{P}}_r^{\text{emp}}$; otherwise reject it ($\widehat{\mathcal{R}} \leftarrow \widehat{\mathcal{R}} \cup \{e_r\}$)
  $\mathcal{A}_{r+1} = \mathcal{A}_r \setminus \{e_r\}$
end for
return $\widehat{\mathcal{P}}_B = \widehat{\mathcal{P}} \cup \mathcal{A}_K$

regret therefore decomposes into a constant boundary term and a sum over suboptimal arms of $\mu_i^c \Delta_i \, \mathbb{E}[n_i(B)]$, so each suboptimal pull costs at most its efficiency gap, and the pull-count bound established for Theorem 1 applies directly. In particular, minimizing the hypervolume-efficiency regret also maximizes the cumulative hypervolume collected under the budget.

4.2 Cost-Aware Pareto Set Identification

CoPSI (Algorithm 2) allocates samples according to costs and progressively classifies configurations as Pareto-optimal or dominated. It maintains an active set $\mathcal{A}_r$ of unresolved configurations, initialized as $\mathcal{A}_1 = [K]$, together with an accepted set $\widehat{\mathcal{P}}$ and a rejected set $\widehat{\mathcal{R}}$, both of which are initially empty.

At the beginning of phase r, CoPSI computes the total cost of the active set, $C_r = C(\mathcal{A}_r)$, and sets the cumulative sampling target

$$n_r = \left\lfloor \frac{B}{L_{K,\lambda} C_r} \right\rfloor, \qquad L_{K,\lambda} = 1 + \sum_{\ell=2}^{K} \frac{1}{1+(\ell-1)\lambda}.$$

Each active configuration i is sampled $n_r - N_i$ additional times, where N_i is its number of previous evaluations, and its empirical mean $\widehat{\boldsymbol{\mu}}_{i,r} = \frac{1}{N_i} \sum_{s=1}^{N_i} \boldsymbol{r}_{i,s}$ is updated. Because n_r is normalized by the aggregate cost of the active set, the number of samples grows as the active set shrinks.

From the empirical means, CoPSI forms the empirical Pareto set of the active configurations,

$$\widehat{\mathcal{P}}_r^{\text{emp}} = \left\{ i \in \mathcal{A}_r : \nexists \, j \in \mathcal{A}_r \setminus \{i\} \text{ such that } \widehat{\boldsymbol{\mu}}_{i,r} \prec \widehat{\boldsymbol{\mu}}_{j,r} \right\}.$$

It then quantifies how confidently each active configuration can be classified. Let $\widehat{m}_r(i,j)$, $\widehat{M}_r(i,j)$, $\widehat{\Delta}_{i,r}^-$, and $\widehat{\Delta}_{i,r}^+$ denote the plug-in counterparts of $m(i,j)$, $M(i,j)$, Δ_i^-, and Δ_i^+ in Section 3.2, obtained by replacing $\boldsymbol{\mu}_i$ with $\widehat{\boldsymbol{\mu}}_{i,r}$ and restricting j to $\mathcal{A}_r \setminus \{i\}$. Here, $\widehat{\Delta}_{i,r}^-$ measures the strongest empirical evidence that i is dominated, whereas $\widehat{\Delta}_{i,r}^+$ measures the evidence that i is Pareto-optimal while taking into account whether its competitors are themselves empirically dominated. The empirical classification gap is

$$\widehat{\gamma}_{i,r} = \widehat{\Delta}_{i,r}^- \text{ if } i \notin \widehat{\mathcal{P}}_r^{\text{emp}}, \qquad \widehat{\gamma}_{i,r} = \widehat{\Delta}_{i,r}^+ \text{ if } i \in \widehat{\mathcal{P}}_r^{\text{emp}}. \tag{1}$$

A larger $\widehat{\gamma}_{i,r}$ indicates that the status of configuration i is easier to determine. CoPSI therefore removes $e_r \in \operatorname{argmax}_{i \in \mathcal{A}_r} \widehat{\gamma}_{i,r}$, breaking ties in favor of $\mathcal{A}_r \setminus \widehat{\mathcal{P}}_r^{\text{emp}}$. It accepts e_r into $\widehat{\mathcal{P}}$ if $e_r \in \widehat{\mathcal{P}}_r^{\text{emp}}$, rejects it into $\widehat{\mathcal{R}}$ otherwise, and sets $\mathcal{A}_{r+1} = \mathcal{A}_r \setminus \{e_r\}$. After $K-1$ phases, the single remaining configuration is included in the output, i.e., $\widehat{\mathcal{P}}_B = \widehat{\mathcal{P}} \cup \mathcal{A}_K$. Computing all empirical gaps in phase r requires $O(k_r^2 D)$ operations, where $k_r = |\mathcal{A}_r|$, so the computational overhead is negligible compared with the cost of LLM evaluations.

The procedure is cost-aware because each sampling target depends on the total cost C_r of the active set, and it is Pareto-aware because configurations are accepted or rejected according to empirical

dominance relations. These properties make CoPSI well suited to fixed-budget LLM configuration evaluation, where the goal is to identify a diverse set of non-dominated configurations that captures different trade-offs among accuracy, latency, and efficiency. The following guarantee shows that the misidentification probability of CoPSI decays exponentially in the budget.

Theorem 2 (Fixed-budget error probability of CoPSI) *Let $K \geq 2$ and let $L_{K,\lambda}$ be defined as above. If $B \geq 2L_{K,\lambda}H_{\boldsymbol{\mu},c}$, then the Pareto set estimate returned by* CoPSI *satisfies*

$$\Pr\big(\widehat{\mathcal{P}}_B \neq \mathcal{P}^\star\big) \leq 2K^2 D \exp\left(-\frac{B}{256\, L_{K,\lambda} H_{\boldsymbol{\mu},c}}\right).$$

Proof sketch and discussion. The proof separates a deterministic classification argument from a probabilistic sampling argument. First, if in every phase r all empirical pairwise margins $\widehat{m}_r(i,j)$ and $\widehat{M}_r(i,j)$ are within $\gamma_{(k_r)}/8$, where $k_r = |\mathcal{A}_r|$, then every removed configuration is classified correctly and the final survivor is Pareto-optimal; the key step is an invariant showing that the elimination rule never removes the Pareto-optimal dominator of an active dominated configuration. Second, the normalization by $L_{K,\lambda}$ guarantees that the phase-wise targets never exceed the budget, because each removal reduces the active cost by at most a fraction $1/(1+(k_r-1)\lambda)$. Third, the definition of $H_{\boldsymbol{\mu},c}$ implies $n_r\gamma^2_{(k_r)} \geq B/(2L_{K,\lambda}H_{\boldsymbol{\mu},c})$ in every phase, so a time-uniform Hoeffding bound together with a union bound over phases, configurations, and objectives yields the stated rate.

Theorem 2 characterizes the difficulty of fixed-budget Pareto configuration identification through two cost-dependent quantities. The complexity $H_{\boldsymbol{\mu},c}$ weighs the squared classification gaps by the cost of the configurations that must still be evaluated, and the factor $L_{K,\lambda}$ accounts for the budget consumed across elimination phases. When all configurations have unit cost, $C(\mathcal{A}) = |\mathcal{A}|$ and $\lambda = 1$, so $H_{\boldsymbol{\mu},c} = \max_{k\in[K]} k/\gamma^2_{(k)}$ and $L_{K,1} = \sum_{\ell=1}^{K} 1/\ell$ coincide, up to an additive constant in the latter, with the complexity and logarithmic factor of successive-rejects algorithms (Audibert & Bubeck, 2010; Kone et al., 2024). The bound thus recovers the standard exponential rate for fixed-budget Pareto set identification, while measuring both the budget and the problem difficulty in units of evaluation cost in the general case.

5 Experiments

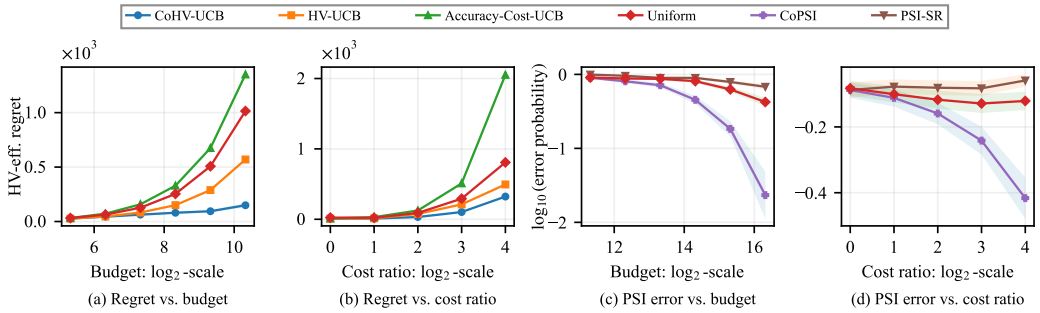
In this section, we conduct experiments to address two questions. **(Q1)** How does cost heterogeneity affect cost-aware allocation relative to cost-insensitive and non-adaptive alternatives? **(Q2)** Do the resulting differences persist on real LLM configuration data, where rewards and costs arise jointly from the evaluated configurations and cannot be varied independently? We answer Q1 with synthetic instances (Section 5.1) and Q2 with GSM8K and PIQA (Section 5.2). Experimental details and additional results are provided in Appendices A and B, respectively.

Each baseline removes one component of our methods. For online selection, HV-UCB drops the cost denominator (testing cost awareness), Accuracy-Cost-UCB optimizes accuracy per unit cost (testing multi-objective modeling), and uniform round-robin allocation serves as a non-adaptive reference; all UCB-based methods share the same confidence-radius calibration. For fixed-budget identification, PSI-SR (Kone et al., 2024) performs the same elimination as CoPSI but ignores costs, and uniform allocation followed by empirical Pareto estimation serves as a non-adaptive reference.

5.1 Synthetic Experiments

To isolate the effect of cost-aware allocation, we construct two-objective Bernoulli instances with fixed reward distributions and varying arm costs. In the online instance (eight arms), the highest-hypervolume arm is the most expensive, while a cheaper balanced arm maximizes the hypervolume per cost. In the identification instance (twelve arms), the Pareto-optimal arms and the dominated arms closest to them are cheap, whereas the clearly dominated arms are expensive. For each task, we vary either the budget or the expensive-to-cheap cost ratio $\kappa \in \{1, 2, 4, 8, 16\}$ under a fixed budget, where $\kappa = 1$ corresponds to equal costs. Full specifications are given in Appendix A.2.

Figure 1(a) shows that the advantage of CoHV-UCB widens as the budget grows: at the largest budget, its regret is 149.45, compared with 569.91 for HV-UCB, a 73.8% reduction. The cost-ratio



(a) Regret vs. budget (b) Regret vs. cost ratio (c) PSI error vs. budget (d) PSI error vs. cost ratio

Figure 1: Synthetic experiments. (a) and (c) vary the budget under heterogeneous costs; (b) and (d) vary the cost ratio κ at fixed budgets $B = 64$ and $B = 1500$, respectively. (a)–(b) report the hypervolume-efficiency regret over 200 paired replays, and (c)–(d) report $\log_{10} \Pr(\widehat{\mathcal{P}}_B \neq \mathcal{P}^\star)$ over 300 paired replays. Shaded regions are 95% confidence intervals.

sweep in Figure 1(b) separates cost awareness from adaptivity and multi-objective modeling. At $\kappa = 1$, all costs are equal and CoHV-UCB coincides with HV-UCB, as expected. For every $\kappa > 1$, CoHV-UCB attains the lowest regret; at $\kappa = 16$, the regrets of CoHV-UCB, HV-UCB, Accuracy-Cost-UCB, and uniform allocation are 322.60, 494.28, 2051.75, and 808.04, respectively, so CoHV-UCB reduces the regret of HV-UCB by 34.7%. The absolute regrets increase with κ because cheaper arms permit more decisions under the fixed budget and enlarge the hypervolume-efficiency gaps.

In the budget sweep of Figure 1(c), CoPSI attains an error probability of 0.023 at the largest budget, compared with 0.680 for PSI-SR and 0.423 for uniform allocation, i.e., relative reductions of 96.6% and 94.5%. In the cost-ratio sweep of Figure 1(d), CoPSI and PSI-SR coincide at $\kappa = 1$. As κ increases from 1 to 16, the error of CoPSI decreases from 0.817 to 0.383, whereas the errors of PSI-SR and uniform allocation are 0.873 and 0.757 at $\kappa = 16$; the error of CoPSI is thus 56.1% lower than that of PSI-SR and 49.3% lower than that of uniform allocation. Because CoPSI normalizes its sampling targets by the cost of the active set, eliminating the expensive and clearly dominated arms frees a large share of the budget for the cheap arms near the Pareto boundary, which explains why its advantage grows with the cost ratio.

5.2 Experiments on LLM Configuration Data

We use GSM8K (Cobbe et al., 2021) and PIQA (Bisk et al., 2020), containing 108 configurations of nine models and 132 configurations of 11 models, respectively. Each configuration combines a model, a temperature, a generation limit, and a prompting strategy, and is evaluated on the same 1,000 validation instances. A pull returns binary accuracy and normalized inverse latency as rewards, and its token cost is charged to the budget. Dataset statistics (Table 2), configuration landscapes (Figure 3), and preprocessing details are given in Appendix A.3.

Online selection uses all configurations of each dataset. Because the complete sets contain many near-tied means, which make exact Pareto identification uninformative, we use a pre-specified 14-arm subset (seven common models at temperatures 0.5 and 1.0) with two Pareto-optimal arms on GSM8K and three on PIQA. Approximate recovery on the complete sets is reported in Appendix B.3.

Figure 2(a)–(b) shows that CoHV-UCB consistently outperforms uniform allocation and Accuracy-Cost-UCB. At the largest token budget, it reduces the regret of uniform allocation by 64.7% on GSM8K (from 169.45 to 59.88) and by 83.2% on PIQA (from 6538.22 to 1097.75). Its improvement over HV-UCB is small on GSM8K but pronounced on PIQA, which agrees with the synthetic findings: GSM8K costs vary only mildly (118.37–147.21 tokens), and its hypervolume-maximizing configuration also maximizes the hypervolume per cost, whereas PIQA has larger cost dispersion (95.01–152.16 tokens) and the two configurations differ. Moreover, since the horizontal axis is $\log_2 B$, the approximately linear curves are consistent with the logarithmic regret in Theorem 1. CoHV-UCB also attains the lowest standard hypervolume regret at every budget (Appendix B.1), and its advantage emerges early in the decision trajectories (Appendix B.2).

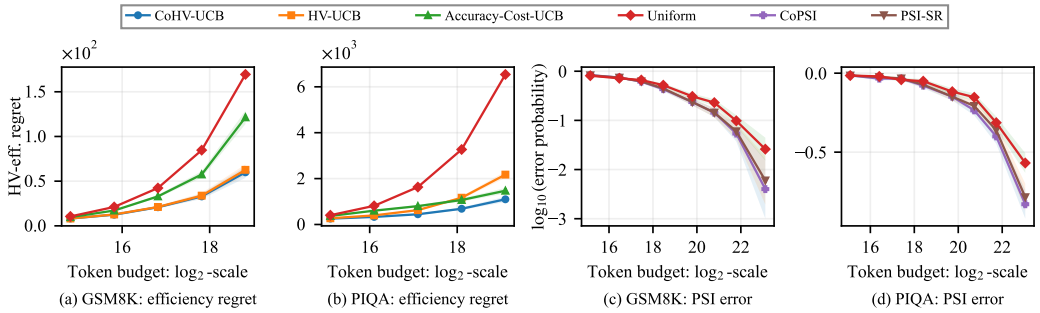


Figure 2: LLM configuration experiments. (a)–(b) Hypervolume-efficiency regret over 100 paired replays. (c)–(d) $\log_{10} \Pr(\widehat{\mathcal{P}}_B \neq \mathcal{P}^\star)$ on the 14-arm subsets over 500 paired replays. Shaded regions are 95% confidence intervals.

Figure 2(c)–(d) shows that the advantage of successive elimination grows with the budget. On GSM8K, CoPSI reduces the error probability from 0.614 to 0.004 over the displayed budget range, compared with 0.662 to 0.026 for uniform allocation, an 84.6% relative reduction at the largest budget. On PIQA, CoPSI attains an error probability of 0.148 at the largest budget, compared with 0.270 for uniform allocation, a 45.2% reduction. Uniform allocation is competitive at the smallest budgets because it avoids committing to noisy early classifications; with more observations, CoPSI eliminates configurations whose status is already clear and concentrates the remaining budget near the Pareto boundary. The cost-insensitive PSI-SR also benefits from successive elimination, but CoPSI attains lower errors at the largest budget (0.004 vs. 0.006 on GSM8K and 0.148 vs. 0.164 on PIQA). The modest difference between the two elimination methods is consistent with the limited cost dispersion among the selected configurations, in contrast to the large gains observed under strong cost heterogeneity in Section 5.1. Appendix B.4 further shows that these conclusions are unchanged when the realized, rather than the mean, token cost of each evaluation is charged.

6 Conclusion

In this paper, we formulated budgeted LLM configuration evaluation as a cost-aware multi-objective bandit problem and studied two complementary goals: online configuration selection and fixed-budget Pareto configuration identification. For online selection, we proposed CoHV-UCB, which selects configurations according to an optimistic hypervolume-per-cost index, and showed that both its hypervolume-efficiency regret and standard regret are of order $O\big(\sum_{i\neq i^\star} C_{\lambda,H}^2 \log B/\Delta_i\big)$, matching the budget and gap dependence of single-objective budgeted bandits. For fixed-budget identification, we proposed CoPSI, which sets its sampling targets according to the aggregate cost of the unresolved configurations and eliminates configurations using empirical Pareto gaps, and proved that its error probability is at most $2K^2 D \exp\big(-B/(256 L_{K,\lambda} H_{\boldsymbol{\mu},c})\big)$, which reduces to the standard fixed-budget rate when all costs are equal. Experiments on synthetic instances and LLM configuration data showed that CoHV-UCB attains lower regret than cost-insensitive, single-objective, and non-adaptive baselines, and that CoPSI attains lower identification error than PSI-SR and uniform allocation, with the largest gains under heterogeneous costs. These results demonstrate the importance of jointly accounting for evaluation cost and multiple objectives in LLM configuration evaluation.

Limitations and future work. Our analysis has several limitations that suggest directions for future work. First, CoPSI assumes deterministic and known costs; although our experiments indicate robustness to stochastic token costs (Appendix B.4), a guarantee for random costs remains open. Second, we treat configurations as independent arms; exploiting the structure shared among models, prompts, and decoding parameters may further improve sample efficiency. Third, extending the framework to contextual or non-stationary settings would support personalized evaluation and evolving deployment environments, and more robust Pareto identification algorithms are needed when classification gaps are small or early estimates are unreliable.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising the manuscript, checking its technical consistency, and developing the experimental code. The authors take full responsibility for the problem formulation, theoretical claims, experimental methodology, and reported results, including all text, code, and figures prepared with AI assistance.

REPRODUCIBILITY STATEMENT

Appendix A specifies the synthetic instances, data preprocessing, replay construction, budgets, evaluation metrics, and hyperparameters used in all experiments, and Appendix B reports the supplementary results. Appendices C–E provide complete proofs of all theoretical results.

REFERENCES

Rin Ashizawa, Yoichi Hirose, Nozomu Yoshinari, Kento Uchida, and Shinichi Shirakawa. Bandit-based prompt design strategy selection improves prompt optimizers. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20799–20817, 2025.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pp. 41–53, 2010.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(11):397–422, 2002.

Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 939–947, 2016.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Proceedings of The 27th Conference on Learning Theory*, pp. 1109–1134, 2014.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM*, 65(3):1–55, 2018.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

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

Élise Crepon, Aurélien Garivier, and Wouter M Koolen. Sequential learning of the Pareto front for multi-objective bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 3583–3591, 2024.

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

Madalina M. Drugan and Ann Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks*, pp. 1–8, 2013.

Kehua Feng, Keyan Ding, Tan Hongzhi, Kede Ma, Zhihua Wang, Shuangquan Guo, Cheng Yuzhou, Ge Sun, Guozhou Zheng, Qiang Zhang, and Huajun Chen. Sample-efficient human evaluation of large language models via maximum discrepancy competition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10913–10947, 2025.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25*, pp. 3212 – 3220, 2012.

Zhi Hong, Qian Zhang, Jiahang Sun, Zhiwei Shang, Mingze Kong, Xiangyi Wang, Yao Shu, and Zhongxiang Dai. MASPOB: Bandit-based prompt optimization for multi-agent systems with graph neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47, 2022.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of The 27th Conference on Learning Theory*, pp. 423–439, 2014.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed pareto set identification. In *Advances in Neural Information Processing Systems 36*, 2023.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit Pareto set identification: the fixed budget setting. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 2548–2556, 2024.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

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

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Haifang Li and Yingce Xia. Infinitely many-armed bandits with budget constraints. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 2182–2188, 2017.

Percy Liang, Rishi Bommasani, Tony Lee, and et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835–8856.

Xiaoqiang Lin, Zhaoxuan Wu, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. Use your INSTINCT: INSTruction optimization for LLMs usIng neural bandits coupled with transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 30317–30345, 2024.

Shang Liu, Jiashuo Jiang, and Xiaocheng Li. Non-stationary bandits with knapsacks. In *Advances in Neural Information Processing Systems 35*, pp. 16522–16532, 2022.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3080–3086, 2019.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *The 13th International Conference on Learning Representations*, 2025.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of large language models. In *Advances in Neural Information Processing Systems 37*, pp. 22483–22512, 2024.

Manhin Poon, XiangXiang Dai, Xutong Liu, Fang Kong, John C. S. Lui, and Jinhang Zuo. Online multi-llm selection via contextual bandits under unstructured context evolution, 2025. URL https://arxiv.org/abs/2506.17670.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7957–7968, 2023.

Shuang Qiu, Dake Zhang, Rui Yang, Boxiang Lyu, and Tong Zhang. Traversing Pareto optimal policies: Provably efficient multi-objective reinforcement learning. *arXiv preprint arXiv:2407.17466*, 2024.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 1760–1770, 2018.

Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. Efficient prompt optimization through the lens of best arm identification. In *Advances in Neural Information Processing Systems 37*, pp. 99646–99685, 2024a.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8601–8629, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, pp. 24824–24837, 2022.

Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. Prompt optimization with ease? efficient ordering-aware automated selection of exemplars. In *Advances in Neural Information Processing Systems 37*, pp. 122706–122740, 2024.

Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 3960–3966, 2015.

Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits with multiple plays. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 2210–2216, 2016.

Qiuyi (Richard) Zhang. Optimal scalarizations for sublinear hypervolume regret. In *Advances in Neural Information Processing Systems 37*, pp. 39963–39999, 2024.

Richard Zhang and Daniel Golovin. Random hypervolume scalarizations for provable multi-objective black box optimization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11096–11105, 2020.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36*, pp. 46595–46623, 2023.

Jin Peng Zhou, Christian K Belardi, Ruihan Wu, Travis Zhang, Carla P Gomes, Wen Sun, and Kilian Q Weinberger. On speeding up language model evaluation. In *The 13th International Conference on Learning Representations*, 2025.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. A survey on efficient inference for large language models, 2024. URL https://arxiv.org/abs/2404.14294.

A Experimental Details

This appendix describes the experimental protocol shared by all experiments (Appendix A.1), the controlled synthetic instances (Appendix A.2), the LLM configuration data and their preprocessing (Appendix A.3), and the budgets and hyperparameters (Appendix A.4) used in Section 5.

A.1 Common Protocol and Evaluation Metrics

All comparisons use paired replays. In each repetition, we pre-generate one arm-wise feedback stream and let every method draw its observations from this common stream, so that differences between methods are not driven by independent sampling noise. All figures report the mean over repetitions together with 95% confidence intervals.

Online selection is evaluated by the hypervolume-efficiency regret R_B^{HV} and, in Appendix B.1, by the standard hypervolume regret R_B^{std}, both defined in Section 3.1. Exact Pareto identification is evaluated by the empirical misidentification probability $\Pr(\widehat{\mathcal{P}}_B \neq \mathcal{P}^\star)$, i.e., the fraction of repetitions in which the returned set differs from the true Pareto set.

To assess the quality of recent decisions, which cumulative regret cannot reveal because it is nondecreasing in the number of pulls, we use the moving-average per-pull regret. Let $\Delta_{i_s} = \nu_{i^\star} - \nu_{i_s}$ be the hypervolume-efficiency gap incurred at pull s. With window length $w = 50$, the moving-average regret at pull t is

$$\overline{\Delta}_t^{(w)} = \frac{1}{\min\{w, t\}} \sum_{s=\max\{1,t-w+1\}}^{t} \Delta_{i_s}.$$

When exact identification is uninformative because of ties or near-zero gaps, we measure approximate Pareto recovery by the F1 score between the estimated and true Pareto sets,

$$\mathrm{F1} = \frac{2\left|\widehat{\mathcal{P}}_B \cap \mathcal{P}^\star\right|}{\left|\widehat{\mathcal{P}}_B\right| + \left|\mathcal{P}^\star\right|},$$

which is the harmonic mean of Pareto precision and Pareto recall. It equals one only under exact recovery and provides a graded measure otherwise.

A.2 Synthetic Instances

Each pull returns two binary coordinates whose means $(\mu_i^{(1)}, \mu_i^{(2)})$ are listed in Table 1, together with the heterogeneous costs. In the online instance, arm 2 has the largest hypervolume but is the most expensive, arm 1 maximizes the hypervolume per cost, and arm 3 has the largest first-objective mean. In the identification instance, the Pareto set consists of arms 1–3; arms 4–6 are dominated with a coordinate gap of 0.03, whereas arms 7–12 are dominated with large gaps.

Table 1: Reward means and heterogeneous costs of the controlled synthetic instances.

Online selection								
Arm	1	2	3	4	5	6	7	8
$\mu_i^{(1)}$	0.72	0.92	0.96	0.40	0.82	0.62	0.68	0.45
$\mu_i^{(2)}$	0.72	0.92	0.40	0.96	0.62	0.82	0.68	0.45
c_i	1	8	1	1	3	3	2	1

Pareto set identification												
Arm	1	2	3	4	5	6	7	8	9	10	11	12
$\mu_i^{(1)}$	0.82	0.60	0.71	0.79	0.57	0.68	0.42	0.30	0.38	0.34	0.31	0.22
$\mu_i^{(2)}$	0.60	0.82	0.71	0.57	0.79	0.68	0.30	0.42	0.34	0.38	0.31	0.36
c_i	1	1	1	1	1	1	16	16	16	16	16	16

The budget sweeps use the costs in Table 1, with normalized budgets $\rho = B/\sum_i c_i \in \{2, 4, 8, 16, 32, 64\}$ for online selection and $\rho \in \{25, 50, 100, 200, 400, 800\}$ for identification. In the cost-ratio sweeps, the reward means are unchanged and $\kappa \in \{1, 2, 4, 8, 16\}$. For online selection,

the largest-hypervolume arm (arm 2) has cost one and every other arm has cost $1/\kappa$; for identification, arms 1–6 have cost $1/\kappa$ and the clearly dominated arms 7–12 have cost one. The budgets are fixed at $B = 64$ for online selection and $B = 1500$ for identification, so that all instances share the same cost scale and $\kappa = 1$ recovers homogeneous costs.

For each arm, we construct a replay population of 20,000 observations with the stated marginal means. Online experiments use 200 paired repetitions, deterministic costs, a reward-confidence scale of 0.25 (see Appendix A.4), and an initialization fraction of 0.05; identification experiments use 300 paired repetitions. No parameter is tuned on the reported synthetic outcomes.

A.3 LLM Configuration Data

The datasets and configurations are described in Section 5.2. Table 2 summarizes their configuration-level statistics, and Figure 3 shows the accuracy–efficiency trade-offs of individual configurations, where both axes are maximization objectives and efficiency is measured by the normalized inverse latency. This subsection also details how the raw evaluation logs are converted into bandit feedback.

Table 2: Configuration-level statistics of the GSM8K and PIQA configuration sets.

Dataset	Models	Configurations	Accuracy	Latency (s)	Tokens	$\lvert\mathcal{P}^\star\rvert$
GSM8K	9	108	0.014–0.165	0.432–8.854	122.1–146.4	7
PIQA	11	132	0.462–0.571	0.443–3.049	93.5–149.9	7

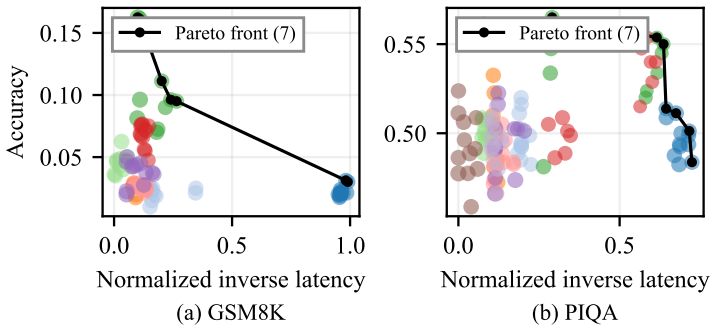


Figure 3: Configuration landscapes on GSM8K and PIQA. Each point is a configuration; point size is proportional to its mean token consumption, and the black curve marks the empirical Pareto front. Both axes are maximization objectives.

We split the 1,000 instances once into 200 profiling instances and 800 evaluation instances. The profiling split determines the latency percentiles (τ_5, τ_{95}), the maximum token cost $c_{\max}$, and the deterministic arm cost c_i, defined as the mean token count on the profiling split. For configuration i and evaluation instance q, a pull returns the binary accuracy, the token cost $c_{i,q}$, and the normalized inverse latency

$$r^{\text{eff}}_{i,q} = \text{clip}\left(\frac{t^{-1}_{i,q} - \tau^{-1}_{95}}{\tau^{-1}_{5} - \tau^{-1}_{95}}, 0, 1\right),$$

where $t_{i,q}$ is the observed latency. Online methods internally use the normalized cost $\widetilde{c}_{i,q} = c_{i,q}/c_{\max}$, whereas all figures report actual token budgets. Feedback is generated by arm-wise bootstrap sampling from the evaluation split.

The 14-arm subset for exact Pareto identification is formed independently of the outcomes from GPT-2, CodeLlama-7B, Tulu-7B, Tulu-2-7B, Gemma-7B, LLaMA-7B, and Mistral-7B at temperatures 0.5 and 1.0, with generation length 128 and direct prompting. Its minimum population gaps are 0.005 on GSM8K and 0.010 on PIQA. Gap ties are resolved with a relative tolerance of 10^{-10} and an absolute tolerance of 10^{-12}.

A.4 Budgets and Hyperparameters

We parameterize the budget by $\rho = B/(K\bar{c})$, where $\bar{c}$ is the mean arm cost. On the LLM data, online experiments use $\rho \in \{2, 4, 8, 16, 32\}$ with 100 paired replays; exact identification on the 14-arm subsets uses $\rho \in \{20, 50, 100, 200, 500, 1000, 2000, 5000\}$ with 500 paired repetitions; approximate recovery on the complete sets (Appendix B.3) uses $\rho \in \{10, 20, 50, 100, 200\}$ with 100 repetitions; and the stochastic-cost comparison (Appendix B.4) uses 500 repetitions for each method and budget.

In the online experiments, the reward and cost confidence radii are scaled as

$$\beta^r_{i,t} = s_r\sqrt{\alpha \log(T_B)/n_{i,t}}, \qquad \beta^c_{i,t} = s_c\sqrt{\alpha \log(T_B)/n_{i,t}},$$

and each arm is initialized with $n_{\text{init}} = \max\{1, \lfloor \eta\rho \rfloor\}$ pulls. The scales are selected using the profiling split only, over $s_r \in \{0.003, 0.01, 0.03, 0.1\}$, $s_c \in \{0.001, 0.003, 0.01\}$, and $\eta \in \{0.05, 0.1\}$; this selects $(s_r, s_c, \eta) = (0.01, 0.01, 0.05)$ on both datasets, and the same calibration is used for all three UCB-based methods. All online methods use the profiling mean cost to check the feasibility of each selection, so that no method exceeds the remaining budget.

B Additional Experimental Results

This appendix complements Section 5 with the standard hypervolume regret (Appendix B.1), within-run online trajectories (Appendix B.2), approximate Pareto recovery on the complete configuration sets (Appendix B.3), and robustness to stochastic token costs (Appendix B.4).

B.1 Standard Hypervolume Regret

Figure 4 complements the efficiency-regret results in Figure 2(a)–(b) with the standard budgeted regret R^{std}_B. Across all five budgets and both datasets, CoHV-UCB attains the lowest mean standard regret among the four methods, and its advantage is most pronounced on PIQA at large budgets. At $\rho = 32$ on PIQA, the regrets of CoHV-UCB, HV-UCB, Accuracy-Cost-UCB, and uniform allocation are 161.07, 447.35, 241.79, and 1517.77, respectively. CoHV-UCB thus reduces the standard regret by 64.0% relative to HV-UCB, by 33.4% relative to Accuracy-Cost-UCB, and by 89.4% relative to uniform allocation, which separately demonstrates the value of accounting for token cost, retaining the efficiency objective, and adapting the allocation to observed rewards.

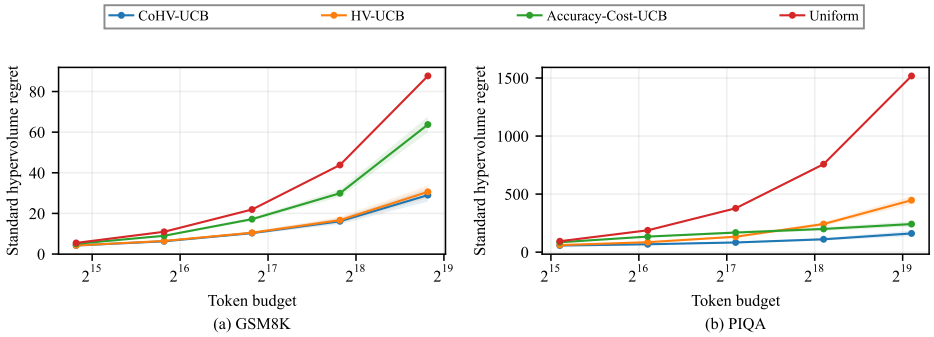


Figure 4: Standard hypervolume regret versus the token budget on GSM8K and PIQA, averaged over 100 paired replays. Shaded regions are 95% confidence intervals.

The corresponding GSM8K regrets at $\rho = 32$ are 29.01, 30.62, 63.75, and 87.69. The two hypervolume-based UCB methods are close because the highest-hypervolume configuration is also highly competitive in hypervolume per cost on this dataset. Nevertheless, CoHV-UCB remains the best method and reduces regret by 54.5% relative to Accuracy-Cost-UCB and by 66.9% relative to uniform allocation. These results show that cost-aware multi-objective selection improves not only the efficiency-normalized criterion used to construct its index, but also the cumulative hypervolume collected under the same token budget, as predicted by Corollary 1.

B.2 Online Decision Trajectories

To examine performance during the decision process rather than only at the stopping budget, Figure 5 fixes $\rho = 32$ and records each method after every pull. The curves are aligned by consumed tokens, so the methods are compared at equal resource expenditure even when they complete different numbers of evaluations.

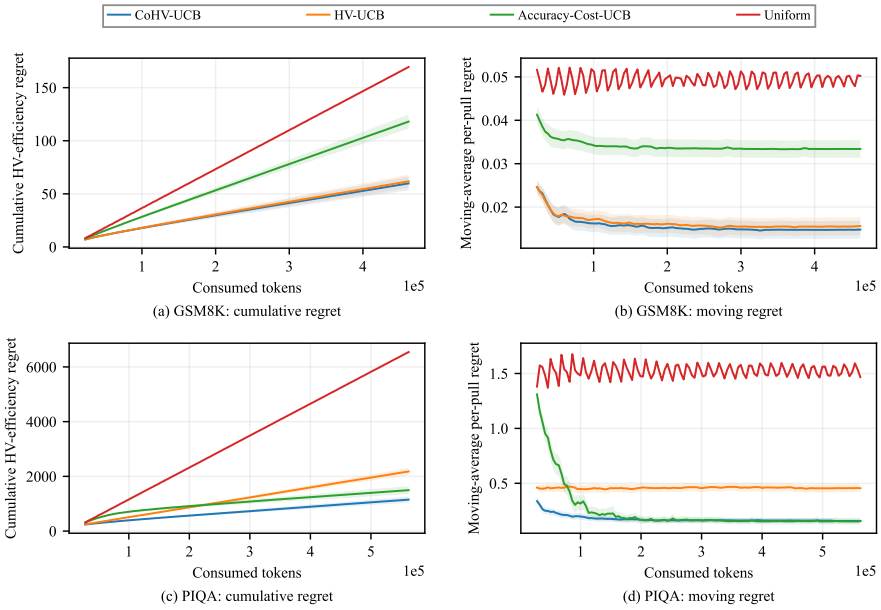


Figure 5: Online decision trajectories at the fixed budget $\rho = 32$. The left column shows the cumulative hypervolume-efficiency regret, and the right column shows the moving-average per-pull regret over the most recent 50 pulls. Horizontal axes report consumed tokens. Shaded regions are 95% confidence intervals over 100 paired replays.

On PIQA, the separation between methods emerges early and persists throughout most of the trajectory. At the end of the budget, CoHV-UCB has a cumulative regret of 1149.32, compared with 2179.81 for HV-UCB, 1495.65 for Accuracy-Cost-UCB, and 6544.07 for uniform allocation. Its final moving-average regret is 0.156, approximately one third of that of HV-UCB (0.455) and one tenth of that of uniform allocation (1.466). It is comparable to that of Accuracy-Cost-UCB (0.157), but CoHV-UCB achieves a substantially smaller cumulative regret, indicating that it combines strong recent decisions with better performance over the entire trajectory.

The GSM8K trajectories exhibit the same ordering relative to the single-objective and non-adaptive baselines. The final cumulative regret of CoHV-UCB is 59.85, compared with 118.04 for Accuracy-Cost-UCB and 169.55 for uniform allocation, and the corresponding moving-average regrets are 0.0148, 0.0334, and 0.0502. CoHV-UCB and HV-UCB remain close, with the latter ending at a cumulative regret of 61.81 and a moving-average regret of 0.0156, which is consistent with the limited cost dispersion discussed in Section 5. Although the moving averages do not decrease monotonically to zero, CoHV-UCB consistently avoids the much larger recent losses incurred by accuracy-only and uniform selection.

B.3 Approximate Pareto Recovery on the Complete Configuration Sets

The complete GSM8K and PIQA configuration sets contain ties and near-zero gaps, which make exact identification uninformative at the available sample sizes. Figure 6 therefore reports the Pareto F1 score on all 108 and 132 configurations. CoPSI outperforms uniform allocation at every budget on both datasets. The advantage is already visible at $\rho = 10$, where the F1 scores are 0.323 versus 0.265 on GSM8K and 0.250 versus 0.148 on PIQA; adaptive elimination is thus useful even when the budget is too small to recover the full front accurately. At $\rho = 200$, CoPSI reaches 0.727 on

GSM8K and 0.486 on PIQA, whereas uniform allocation reaches 0.677 and 0.409, which correspond to absolute gains of 0.050 and 0.077. The larger relative gain on PIQA (18.7%) is notable because PIQA contains more configurations and a denser set of near-boundary configurations: by rejecting configurations whose status is already clear, CoPSI reserves more evaluations for the ambiguous part of the front.

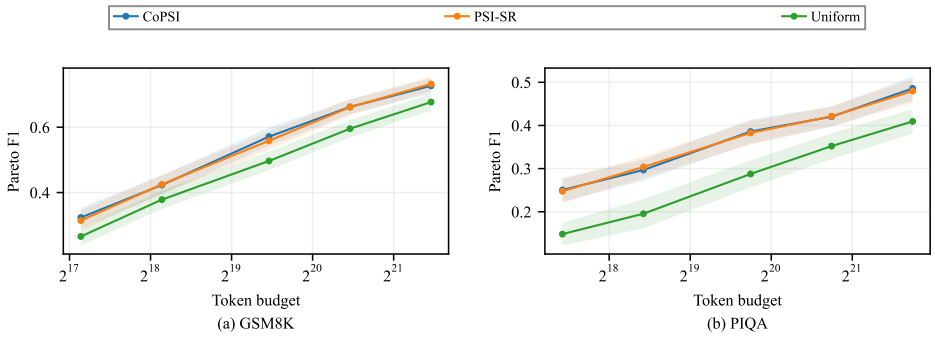


Figure 6: Pareto F1 score on all 108 GSM8K and 132 PIQA configurations over 100 repetitions. Shaded regions are 95% confidence intervals.

CoPSI and the cost-insensitive PSI-SR are closer, as expected from the moderate cost dispersion of these data. Their F1 scores at $\rho = 200$ are 0.727 and 0.732 on GSM8K and 0.486 and 0.480 on PIQA. Neither method dominates the other across all budgets, but both consistently improve over uniform allocation. Incorporating configuration-dependent costs therefore preserves the statistical benefit of successive elimination while making the allocation compatible with a token budget.

B.4 Robustness to Stochastic Token Costs

Theorem 2 assumes deterministic costs, and the identification results in Section 5 accordingly charge the profiling mean c_i on every pull. To assess robustness to this assumption, we instead charge all methods the realized cost $c_{i,q}$ of each evaluation. CoPSI continues to use the profiling estimate c_i to set its sampling targets, PSI-SR retains its cost-insensitive allocation, and uniform allocation performs only complete sampling rounds so that every arm receives the same number of observations. All methods use the same pre-generated feedback stream in each repetition.

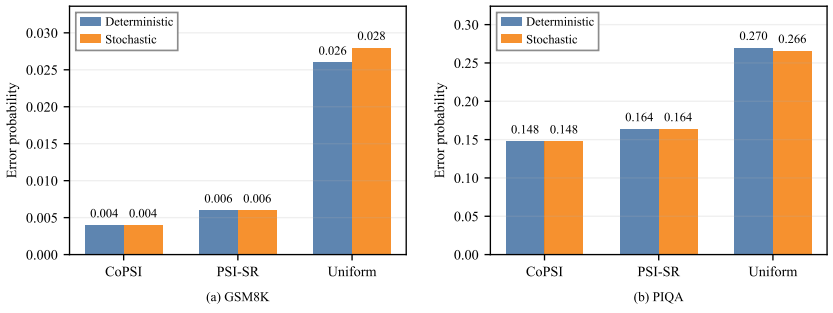


Figure 7: Pareto identification error on the 14-arm subsets at $\rho = 5000$ under deterministic and stochastic token costs over 500 repetitions. Values above the bars are empirical error probabilities.

Figure 7 shows that realized token variability causes no material degradation for any method. On GSM8K, the errors under deterministic and stochastic costs are both 0.004 for CoPSI and both 0.006 for PSI-SR, while that of uniform allocation changes only from 0.026 to 0.028. On PIQA, the corresponding values are 0.148 in both settings for CoPSI, 0.164 in both settings for PSI-SR, and 0.270 versus 0.266 for uniform allocation. This insensitivity arises because the phase-wise sampling targets of both elimination methods are computed from profiling costs, and at this large budget the

realized token fluctuations average out over many pulls without triggering the budget stopping rule before the targets are reached. Consequently, the two cost settings usually evaluate the same feedback samples and return the same Pareto estimate. Uniform allocation shows a small difference because its number of complete sampling rounds is determined directly by the realized costs.

Under stochastic costs, CoPSI retains the lowest error on both datasets. On GSM8K, its error of 0.004 is 33.3% lower than that of PSI-SR (0.006) and 85.7% lower than that of uniform allocation (0.028); on PIQA, its error of 0.148 is 9.8% lower than that of PSI-SR and 44.4% lower than that of uniform allocation. These results support using profiling means for cost-aware allocation when the exact token charge of an individual evaluation is unknown in advance.

C Proof of Theorem 1

We first bound the expected hypervolume that any budgeted policy can collect (Lemma 1); this lemma is also used in the proof of Corollary 1. We then prove Theorem 1 in four steps.

Lemma 1 (Optimal benchmark upper bound) *Consider the budgeted multi-objective bandit problem in which pulling arm i incurs a random cost $c_{i,t} \in [\lambda, 1]$ with mean μ_i^c and yields the scalar hypervolume utility $H_i = \mathrm{HV}(\boldsymbol{\mu}_i)$. Let $i^\star = \mathrm{argmax}_{i\in[K]} H_i/\mu_i^c$ and $\nu_{i^\star} = H_{i^\star}/\mu_{i^\star}^c$. Assume that the cost samples of each arm are i.i.d. across pulls and that, conditional on the history and the selected arm i, the next cost has mean μ_i^c. Allowing a policy to include the boundary pull that first exhausts the budget, every policy that selects arm I_t at pull t satisfies*

$$\mathbb{E}\Big[\sum_{t=1}^{T(B)} H_{I_t}\Big] \le \nu_{i^\star}(B+1).$$

In particular, $R_B^\star \le \nu_{i^\star}(B+1)$.

Proof of Lemma 1. Fix an admissible policy, let $T(B)$ be its number of pulls, and let $N_i(B) = \sum_{t=1}^{T(B)} \mathbb{I}\{I_t = i\}$ be the number of pulls of arm i. By the definition of $i^\star$, every arm satisfies $H_i \le \nu_{i^\star}\mu_i^c$. Since I_t is measurable with respect to the history before pull t and the next cost has conditional mean $\mu_{I_t}^c$, the tower property gives

$$\mathbb{E}\Big[\sum_{t=1}^{T(B)} H_{I_t}\Big] = \sum_{i=1}^{K} H_i\,\mathbb{E}[N_i(B)] \le \nu_{i^\star}\sum_{i=1}^{K}\mu_i^c\,\mathbb{E}[N_i(B)] = \nu_{i^\star}\,\mathbb{E}\Big[\sum_{t=1}^{T(B)} c_{I_t,t}\Big].$$

Immediately before the last pull, the cumulative cost is at most B, and every pull costs at most one; hence $\sum_{t=1}^{T(B)} c_{I_t,t} \le B+1$ almost surely, even when the boundary pull is included. Substituting this bound into the display above proves the claim for every admissible policy, and taking the supremum over policies yields $R_B^\star \le \nu_{i^\star}(B+1)$. □

Step 1: A uniform concentration event. For each arm i, objective d, and sample size s, let $\hat\mu_{i,s}^{(d)}$ and $\hat\mu_{i,s}^c$ denote the empirical means of the first s samples of the d-th objective and of the cost of arm i, respectively. Define the good event

$$\mathcal{E} = \Big\{|\hat\mu_{i,s}^{(d)} - \mu_i^{(d)}| \le \sqrt{\tfrac{\alpha\log T_B}{s}} \text{ and } |\hat\mu_{i,s}^c - \mu_i^c| \le \sqrt{\tfrac{\alpha\log T_B}{s}},\ \ \forall i\in[K],\ d\in[D],\ s\le T_B\Big\}.$$

Since rewards and costs lie in $[0,1]$, Hoeffding's inequality gives $\Pr\big(|\hat\mu_{i,s}^{(d)} - \mu_i^{(d)}| > \sqrt{\alpha\log T_B/s}\big) \le 2T_B^{-2\alpha}$ for any fixed i, d, s, and the same bound holds for the cost. A union bound over arms, the D objectives and the cost coordinate, and all sample sizes $s \le T_B$ yields

$$\Pr(\mathcal{E}^c) \le 2K(D+1)T_B\cdot T_B^{-2\alpha} = 2K(D+1)T_B^{1-2\alpha}.$$

Because each pull costs at least λ, at most T_B pulls occur before the budget is exhausted, so the failure event contributes at most $T_B\Pr(\mathcal{E}^c) \le 2K(D+1)T_B^{2-2\alpha}$ to the expected number of pulls of any fixed arm.

Step 2: Optimism of the optimal arm. On $\mathcal{E}$, for any arm i and any $t \le T_B$, we have $\mu_i^{(d)} \le \bar{r}_{i,t}^{(d)} + \beta_{i,t}$ for all $d \in [D]$. Since $\mu_i^{(d)} \le 1$, clipping preserves this inequality, i.e., $\mu_i^{(d)} \le U_{i,t}^{(d)}$, and the coordinate-wise monotonicity of the hypervolume gives $H(\boldsymbol{\mu}_i) \le H(\mathbf{U}_{i,t})$. Moreover, $\bar{c}_{i,t} - \beta_{i,t} \le \mu_i^c$ on $\mathcal{E}$ and $\lambda/2 \le \mu_i^c$, so $\underline{c}_{i,t} = \max\{\lambda/2, \bar{c}_{i,t} - \beta_{i,t}\} \le \mu_i^c$. Applying both inequalities to $i^\star$,

$$I_{i^\star,t} = \frac{H(\mathbf{U}_{i^\star,t})}{\underline{c}_{i^\star,t}} \ge \frac{H(\boldsymbol{\mu}_{i^\star})}{\mu_{i^\star}^c} = \nu_{i^\star},$$

so the index of the optimal arm is optimistic on the good event.

Step 3: An upper bound on the index of a suboptimal arm. Fix $i \ne i^\star$. On $\mathcal{E}$, $U_{i,t}^{(d)} = \min\{1, \bar{r}_{i,t}^{(d)} + \beta_{i,t}\} \le \mu_i^{(d)} + 2\beta_{i,t}$ for every d, so $\|\mathbf{U}_{i,t} - \boldsymbol{\mu}_i\|_\infty \le 2\beta_{i,t}$, and the Lipschitz continuity of H on $[0,1]^D$ gives $H(\mathbf{U}_{i,t}) - H(\boldsymbol{\mu}_i) \le 2D\beta_{i,t}$. For the denominator, $\bar{c}_{i,t} \ge \mu_i^c - \beta_{i,t}$ on $\mathcal{E}$, hence $\underline{c}_{i,t} \ge \max\{\lambda/2, \mu_i^c - 2\beta_{i,t}\}$ and $\mu_i^c - \underline{c}_{i,t} \le 2\beta_{i,t}$. Adding and subtracting $H(\boldsymbol{\mu}_i)/\underline{c}_{i,t}$,

$$I_{i,t} - \nu_i = \frac{H(\mathbf{U}_{i,t}) - H(\boldsymbol{\mu}_i)}{\underline{c}_{i,t}} + \frac{H(\boldsymbol{\mu}_i)\,(\mu_i^c - \underline{c}_{i,t})}{\underline{c}_{i,t}\,\mu_i^c} \le \frac{2D\beta_{i,t}}{\lambda/2} + \frac{2\beta_{i,t}}{(\lambda/2)\lambda} = \Big(\frac{4D}{\lambda} + \frac{4}{\lambda^2}\Big)\beta_{i,t},$$

where we used $\underline{c}_{i,t} \ge \lambda/2$, $\mu_i^c \ge \lambda$, and $H(\boldsymbol{\mu}_i) \le 1$. With $C_{\lambda,H} = \frac{4D}{\lambda} + \frac{4}{\lambda^2}$, we conclude that $I_{i,t} \le \nu_i + C_{\lambda,H}\beta_{i,t}$.

Step 4: Bounding the number of pulls of a suboptimal arm. Let $n_i(B)$ denote the number of times arm i is pulled before the budget B is exhausted. If arm $i \ne i^\star$ is selected at round t on $\mathcal{E}$, then $I_{i,t} \ge I_{i^\star,t}$, and Steps 2–3 give $\nu_i + C_{\lambda,H}\beta_{i,t} \ge \nu_{i^\star}$, i.e., $C_{\lambda,H}\beta_{i,t} \ge \Delta_i$. Hence arm i cannot be selected whenever $C_{\lambda,H}\beta_{i,t} \le \Delta_i/2$, which, since $\beta_{i,t} = \sqrt{\alpha \log T_B / n_{i,t}}$, holds as soon as

$$n_{i,t} \ge N_i := \left\lceil \frac{4\alpha C_{\lambda,H}^2 \log T_B}{\Delta_i^2} \right\rceil.$$

Thus, on $\mathcal{E}$, $n_i(B) \le 1 + N_i$, where the additional 1 accounts for the possible boundary pull. On $\mathcal{E}^c$, the number of pulls is at most T_B. Combining the two cases yields the pull-count bound

$$\mathrm{E}[n_i(B)] \le (1 + N_i)\Pr(\mathcal{E}) + T_B \Pr(\mathcal{E}^c) \le 1 + \left\lceil \frac{4\alpha C_{\lambda,H}^2 \log T_B}{\Delta_i^2} \right\rceil + 2K(D+1)T_B^{2-2\alpha}. \quad (2)$$

Finally, by the definition of the hypervolume-efficiency regret, $\mathrm{E}[R_B^{\mathrm{HV}}] = \sum_{i \ne i^\star} \Delta_i \mathrm{E}[n_i(B)] = O\big(\sum_{i \ne i^\star} C_{\lambda,H}^2 \log B/\Delta_i\big)$, where we used $\log T_B = O(\log B)$, treating λ as a problem-dependent constant, and $\alpha \ge 2$, which makes the last term in (2) $O(1)$. □

D Proof of Corollary 1

The proof follows the standard regret decomposition for budgeted multi-armed bandits, with the scalar reward of arm i replaced by its hypervolume utility H_i. By Lemma 1, we have $R_B^\star \le \nu_{i^\star}(B+1)$. Since H_i is deterministic once the arm is fixed, $\mathrm{E}[\sum_{t=1}^{T(B)} H_{i_t}] = \sum_{i=1}^K H_i \mathrm{E}[n_i(B)]$, and therefore

$$\mathrm{E}[R_B^{\mathrm{std}}] \le \nu_{i^\star}(B+1) - \sum_{i=1}^K H_i \mathrm{E}[n_i(B)].$$

Let $C_{T(B)} = \sum_{t=1}^{T(B)} c_{i_t,t}$ be the total cost consumed before stopping. By the definition of $T(B)$, $C_{T(B)} \le B < C_{T(B)+1}$, and since every cost is at most one, $C_{T(B)} > B-1$. Taking expectations and using the optional-sampling identity for adaptively sampled bounded costs gives $\sum_{i=1}^K \mu_i^c \mathrm{E}[n_i(B)] \ge B - 1$, i.e., $B + 1 - \sum_{i=1}^K \mu_i^c \mathrm{E}[n_i(B)] \le 2$. Consequently,

$$\begin{aligned}\mathrm{E}[R_B^{\mathrm{std}}] &\le \nu_{i^\star}\Big(B + 1 - \sum_{i=1}^K \mu_i^c \mathrm{E}[n_i(B)]\Big) + \sum_{i=1}^K (\nu_{i^\star}\mu_i^c - H_i)\mathrm{E}[n_i(B)] \\ &\le 2\nu_{i^\star} + \sum_{i \ne i^\star} \mu_i^c \Delta_i\, \mathrm{E}[n_i(B)],\end{aligned}$$

where we used $\nu_{i^\star}\mu^c_{i^\star} - H_{i^\star} = 0$ and $\nu_{i^\star}\mu^c_i - H_i = \mu^c_i(\nu_{i^\star} - \nu_i) = \mu^c_i\Delta_i$ for $i \neq i^\star$. Substituting the pull-count bound (2) established in the proof of Theorem 1 yields a finite-budget bound on $\mathrm{E}[R^{\mathrm{std}}_B]$. Finally, since $\mu^c_i \leq 1$, the leading term satisfies $\mu^c_i\Delta_i \cdot C^2_{\lambda,H}\log B/\Delta^2_i \leq C^2_{\lambda,H}\log B/\Delta_i$, which gives the stated bound. □

E Proof of Theorem 2

The proof separates the deterministic classification argument from the cost-aware sampling calculation. The elimination rule of CoPSI is the successive-rejects version of Empirical Gap Elimination (EGE-SR): in each phase it removes an active arm with the largest empirical classification gap and, among ties, removes an empirically dominated arm whenever possible. The only difference from standard EGE-SR (Kone et al., 2024) is the cost-aware choice of the sampling target n_r. We first show that the elimination rule classifies every configuration correctly whenever the empirical pairwise margins are sufficiently accurate (Step 1), and then show that the cost-aware sampling schedule is budget-feasible (Step 2) and guarantees this accuracy with high probability (Steps 3–5).

Step 1: Deterministic correctness of the elimination rule. We first state the deterministic property of EGE-SR used in the proof.

Lemma 2 (Deterministic correctness of EGE-SR) *Consider EGE-SR with the empirical classification gap in (1), and let $k_r = |\mathcal{A}_r|$. Suppose that, for every phase r and every active pair $i, j \in \mathcal{A}_r$,*

$$\left|\widehat{m}_r(i,j) - m(i,j)\right| \leq \frac{\gamma_{(k_r)}}{8} \qquad \text{and} \qquad \left|\widehat{M}_r(i,j) - M(i,j)\right| \leq \frac{\gamma_{(k_r)}}{8}.$$

Then every removed arm is classified correctly and the final surviving arm is Pareto-optimal. Hence $\widehat{\mathcal{P}}_B = \mathcal{P}^\star$.

Proof of Lemma 2. For each dominated arm i, fix $i^\dagger \in \operatorname{argmax}_{j\in\mathcal{P}^\star} m(i,j)$. This choice satisfies $m(i,i^\dagger) = \gamma_i$: starting from any arm that dominates i and repeatedly moving to one of its dominators must terminate at a Pareto-optimal arm, and every move can only increase the coordinate-wise domination margin, so the maximum in $\Delta^-_i = \max_{j\neq i} m(i,j)$ is attained by a Pareto-optimal arm. Fix a phase r, write $k = k_r$ and $\varepsilon = \gamma_{(k)}/8$. We first record three deterministic consequences of the assumed pairwise error bounds.

Gap representation. For every active arm,

$$\widehat{\gamma}_{i,r} = \max\{\widehat{\Delta}^-_{i,r},\ \widehat{\Delta}^+_{i,r}\}. \tag{3}$$

Indeed, if i is empirically dominated, then $\widehat{\Delta}^-_{i,r} > 0$ and $\widehat{\Delta}^+_{i,r} \leq 0$; if it is empirically Pareto-optimal, the two inequalities are reversed. Thus (3) follows from (1).

Lower bound on active empirical gaps. Suppose that every active dominated arm i has its fixed Pareto dominator $i^\dagger$ active. Then

$$\widehat{\gamma}_{i,r} \geq \gamma_i - 2\varepsilon, \qquad i \in \mathcal{A}_r. \tag{4}$$

For a dominated arm, $i^\dagger \in \mathcal{A}_r$ and the pairwise error bound give $\widehat{\Delta}^-_{i,r} \geq \widehat{m}_r(i,i^\dagger) \geq m(i,i^\dagger) - \varepsilon = \gamma_i - \varepsilon$. For a Pareto-optimal arm i and any active competitor $j \in \mathcal{A}_r \setminus \{i\}$, the pairwise error bound gives $\widehat{M}_r(i,j) \geq M(i,j) - \varepsilon$ and $\widehat{M}_r(j,i)^+ \geq M(j,i)^+ - \varepsilon$. Moreover, whenever j is dominated, an arm attaining Δ^-_j remains active by assumption, and if j is Pareto-optimal then $(\Delta^-_j)^+ = 0$; in either case $(\widehat{\Delta}^-_{j,r})^+ \geq (\Delta^-_j)^+ - \varepsilon$. Consequently,

$$\widehat{M}_r(i,j) \wedge \left[\widehat{M}_r(j,i)^+ + \left(\widehat{\Delta}^-_{j,r}\right)^+\right] \geq \left\{M(i,j) \wedge \left[M(j,i)^+ + \left(\Delta^-_j\right)^+\right]\right\} - 2\varepsilon.$$

Since a minimum over active competitors is at least the population minimum over all competitors, the definition $\Delta^+_i = \gamma_i$ gives $\widehat{\Delta}^+_{i,r} \geq \gamma_i - 2\varepsilon$, and (4) follows from (3).

A missed empirical domination has a small population margin. If i is dominated, both i and $i^\dagger$ are active, and $i^\dagger$ does not empirically dominate i, then $\widehat{m}_r(i,i^\dagger) \leq 0$, and the pairwise error bound yields

$$\gamma_i = m(i,i^\dagger) \leq \varepsilon. \tag{5}$$

We now establish, by induction over the phases, the dominator-preservation invariant

$$\mathcal{Q}_r: \quad i \in \mathcal{A}_r \setminus \mathcal{P}^\star \implies i^\dagger \in \mathcal{A}_r. \tag{6}$$

It holds at $r = 1$ because $\mathcal{A}_1 = [K]$. Assume that it holds at phase r and, for a contradiction, suppose that $e_r = i^\dagger$ while the corresponding dominated arm i remains active. Among the k active arms there exists an arm a with $\gamma_a \geq \gamma_{(k)}$, so (4) and the maximality of e_r give

$$\widehat{\gamma}_{i^\dagger,r} \geq \widehat{\gamma}_{a,r} \geq \gamma_{(k)} - 2\varepsilon = 6\varepsilon. \tag{7}$$

If $i^\dagger$ is empirically dominated, its population Pareto-optimality implies $m(i^\dagger, j) \leq 0$ for every j, so $\widehat{\gamma}_{i^\dagger,r} = \widehat{\Delta}^-_{i^\dagger,r} \leq \varepsilon$, contradicting (7). Suppose instead that $i^\dagger$ is empirically Pareto-optimal. It cannot empirically dominate i: otherwise i is empirically dominated and, using i as a competitor in the minimum defining $\widehat{\Delta}^+_{i^\dagger,r}$, we get $\widehat{\gamma}_{i^\dagger,r} \leq \widehat{M}_r(i, i^\dagger)^+ + (\widehat{\Delta}^-_{i,r})^+ = \widehat{\gamma}_{i,r}$; either the inequality is strict, in which case $i^\dagger$ is not selected, or it is an equality, in which case the tie rule selects the empirically dominated arm i, and both alternatives contradict the removal of $i^\dagger$. Hence (5) gives $\gamma_i \leq \varepsilon$. Moreover, $M(i, i^\dagger) = -m(i, i^\dagger) = -\gamma_i \leq 0$, so $\widehat{M}_r(i, i^\dagger)^+ \leq \varepsilon$, and the uniform pairwise bound gives $(\widehat{\Delta}^-_{i,r})^+ \leq \gamma_i + \varepsilon \leq 2\varepsilon$. Using i as a competitor once more yields $\widehat{\gamma}_{i^\dagger,r} = \widehat{\Delta}^+_{i^\dagger,r} \leq 3\varepsilon$, again contradicting (7). Therefore $\mathcal{Q}_{r+1}$ holds, and (6) holds at every phase.

It remains to show that the arm removed in each phase is classified correctly. The invariant and (4) imply, exactly as in (7), that

$$\widehat{\gamma}_{e_r,r} \geq 6\varepsilon. \tag{8}$$

If e_r is empirically dominated but truly Pareto-optimal, then $\widehat{\gamma}_{e_r,r} = \widehat{\Delta}^-_{e_r,r} \leq \varepsilon$, a contradiction. If e_r is empirically Pareto-optimal but truly dominated, its fixed dominator $e_r^\dagger$ is active by (6) and does not empirically dominate e_r, so (5) gives $\gamma_{e_r} \leq \varepsilon$. Since $M(e_r, e_r^\dagger) = -\gamma_{e_r} \leq 0$, the pairwise error bound and the first term in the minimum defining $\widehat{\Delta}^+_{e_r,r}$ give $\widehat{\gamma}_{e_r,r} = \widehat{\Delta}^+_{e_r,r} \leq \widehat{M}_r(e_r, e_r^\dagger)^+ \leq \varepsilon$, again contradicting (8). Thus every accepted arm is truly Pareto-optimal and every rejected arm is truly dominated. Finally, after $K - 1$ phases, a dominated final survivor would, by (6), require a distinct active Pareto dominator, which is impossible because $\mathcal{A}_K$ is a singleton. The final survivor is therefore Pareto-optimal, and the returned set is exactly $\mathcal{P}^\star$. □

It remains to prove that the cost-aware sampling schedule guarantees the conditions of Lemma 2 with high probability.

Step 2: Budget feasibility. At phase r, CoPSI sets $n_r = \lfloor B/(L_{K,\lambda} C_r) \rfloor$ with $C_r = C(\mathcal{A}_r)$. Since exactly one arm is removed per phase and every arm has positive cost, C_r is decreasing and n_r is nondecreasing. The algorithm stops after phase $K - 1$, whose active set contains two arms, and returns the singleton $\mathcal{A}_K$ without further sampling. Hence, with $n_0 = 0$, summation by parts and $n_r \leq B/(L_{K,\lambda} C_r)$ give

$$\sum_{r=1}^{K-1} C_r(n_r - n_{r-1}) = C_{K-1} n_{K-1} + \sum_{r=1}^{K-2} (C_r - C_{r+1}) n_r \leq \frac{B}{L_{K,\lambda}} \Big(1 + \sum_{r=1}^{K-2} \frac{C_r - C_{r+1}}{C_r}\Big).$$

At phase r the active set has size $k_r = K - r + 1$; the removed arm costs at most 1 and the remaining $k_r - 1$ arms cost at least $(k_r - 1)\lambda$ in total, so $(C_r - C_{r+1})/C_r \leq 1/(1 + (k_r - 1)\lambda)$. Therefore the total cost is at most

$$\frac{B}{L_{K,\lambda}} \Big(1 + \sum_{r=1}^{K-2} \frac{1}{1 + (k_r - 1)\lambda}\Big) \leq \frac{B}{L_{K,\lambda}} \Big(1 + \sum_{\ell=2}^{K} \frac{1}{1 + (\ell - 1)\lambda}\Big) = B,$$

and the sampling schedule is feasible under budget B.

Step 3: A phase-wise lower bound on the number of samples. By the definition of $H_{\boldsymbol{\mu},c}$, every realized active set satisfies $C_r = C(\mathcal{A}_r) \leq H_{\boldsymbol{\mu},c} \gamma^2_{(k_r)}$. Since rewards lie in $[0, 1]$, every classification gap is at most one, so $C_r \leq H_{\boldsymbol{\mu},c}$. Under the assumption $B \geq 2L_{K,\lambda} H_{\boldsymbol{\mu},c}$, we have $B/(L_{K,\lambda} C_r) \geq 2$, and the elementary inequality $\lfloor x \rfloor \geq x/2$ for $x \geq 2$ gives $n_r \geq B/(2L_{K,\lambda} C_r)$. Multiplying by $\gamma^2_{(k_r)}$,

$$n_r \gamma^2_{(k_r)} \geq \frac{B \gamma^2_{(k_r)}}{2L_{K,\lambda} C_r} \geq \frac{B}{2L_{K,\lambda} H_{\boldsymbol{\mu},c}}. \tag{9}$$

Step 4: Concentration at adaptive sampling targets. For each arm i and coordinate d, expose in advance an infinite i.i.d. sequence of rewards with mean $\mu_i^{(d)}$, and let $\overline{\mu}_{i,n}^{(d)}$ be the empirical mean of its first n observations. Although the active set and the targets n_r are data-dependent, Ville's inequality applied to the Hoeffding exponential supermartingales for the upper and lower tails yields the time-uniform bound

$$\Pr\Big(\exists\, n \geq N:\ \big|\overline{\mu}_{i,n}^{(d)} - \mu_i^{(d)}\big| > x\Big) \leq 2\exp(-2Nx^2). \tag{10}$$

For each phase r, define $\mathcal{E}_r = \{\forall i \in \mathcal{A}_r, \forall d \in [D] : |\widehat{\mu}_{i,r}^{(d)} - \mu_i^{(d)}| \leq \gamma_{(k_r)}/16\}$ and $\mathcal{E} = \bigcap_{r=1}^{K-1} \mathcal{E}_r$. Since each pairwise margin is a minimum or maximum of coordinate-wise differences, on $\mathcal{E}_r$ every active pair satisfies $|\widehat{m}_r(i,j) - m(i,j)| \leq \gamma_{(k_r)}/8$ and $|\widehat{M}_r(i,j) - M(i,j)| \leq \gamma_{(k_r)}/8$. By (9), every realized target satisfies $n_r \geq N := B/(2L_{K,\lambda}H_{\boldsymbol{\mu},c}\gamma_{(k_r)}^2)$. Applying (10) with $x = \gamma_{(k_r)}/16$ and a union bound over at most K active arms and D coordinates,

$$\Pr(\mathcal{E}_r^c) \leq 2KD\exp\Big(-2\cdot\frac{B}{2L_{K,\lambda}H_{\boldsymbol{\mu},c}\gamma_{(k_r)}^2}\cdot\frac{\gamma_{(k_r)}^2}{16^2}\Big) = 2KD\exp\Big(-\frac{B}{256L_{K,\lambda}H_{\boldsymbol{\mu},c}}\Big).$$

A union bound over the $K-1$ phases gives $\Pr(\mathcal{E}^c) \leq 2K^2D\exp\big(-B/(256L_{K,\lambda}H_{\boldsymbol{\mu},c})\big)$.

Step 5: Concluding the proof. On $\mathcal{E}$, the pairwise conditions of Lemma 2 hold in every phase. Hence all arms removed by CoPSI are classified correctly, the final survivor in $\mathcal{A}_K$ is Pareto-optimal, and $\widehat{\mathcal{P}}_B = \widehat{\mathcal{P}} \cup \mathcal{A}_K = \mathcal{P}^\star$. Therefore

$$\Pr\big(\widehat{\mathcal{P}}_B \neq \mathcal{P}^\star\big) \leq \Pr(\mathcal{E}^c) \leq 2K^2D\exp\Big(-\frac{B}{256L_{K,\lambda}H_{\boldsymbol{\mu},c}}\Big).$$

□