Cost-Aware Best Arm Identification via Dueling Feedback with Applications to Large Language Models

Extended Abstract

Sarvesh Gharat
Centre for Machine Intelligence and Data Science, IIT Bombay
Mumbai, India
sarveshgharat19@gmail.com

Nikhil Karamchandani
Department of Electrical Engineering, IIT Bombay
Mumbai, India
nikhilk@ee.iitb.ac.in

Jayakrishnan Nair
Department of Electrical Engineering, IIT Bombay
Mumbai, India
jayakrishnan.nair@ee.iitb.ac.in

ABSTRACT

Motivated by the problem of selecting the best large language model (LLM) under heterogeneous querying costs, we study a variant of the multi-armed bandit problem with (i) dueling feedback, where pairwise comparisons provide preference signals, and (ii) heterogeneous sampling costs. Assuming the existence of a Condorcet winner—an assumption we validate empirically on real-world datasets—we propose a Track-and-Stop style algorithm for best-arm identification with prescribed confidence. We prove that the algorithm almost surely achieves asymptotically optimal cost as the error probability tends to zero. Experiments on synthetic and real-world datasets demonstrate consistent improvements over classical cost-unaware methods and existing cost-aware baselines.

KEYWORDS

Dueling Bandits, Best Arm Identification, Cost-Aware Learning, Large Language Models (LLMs)

ACM Reference Format:

Sarvesh Gharat, Nikhil Karamchandani, and Jayakrishnan Nair. 2026. Cost-Aware Best Arm Identification via Dueling Feedback with Applications to Large Language Models: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/GEKA7634

1 INTRODUCTION

Sequential decision-making under uncertainty is a fundamental challenge in machine learning, with the multi-armed bandit (MAB) model serving as its canonical abstraction. While classical bandits assume access to scalar rewards, many real-world applications provide feedback only via pairwise preferences. This preference-based feedback has emerged as a more reliable signal in human-centered evaluation tasks, as individuals often find it easier to answer "Which of these two options is better?" than to assign consistent absolute scores [3, 8, 9].

Such preference-based feedback is naturally modeled using the dueling bandit framework, where the learner observes noisy pairwise comparisons rather than scalar rewards [10, 11]. An instance of a dueling bandit problem is characterized by a preference matrix $\mathbb{P} = [[p_{i,j}]]$, where $p_{i,j}$ reflects the probability that arm i wins over arm j in a duel.

While originally motivated by information retrieval and recommendation systems, preference-based learning has recently become central to the evaluation of large language models (LLMs). Platforms such as Chatbot Arena [1] rely on head-to-head human comparisons to rank models, as automatic metrics often suffer from data leakage and misalignment with human judgment. In this context, each arm corresponds to a model, and comparisons are obtained through human preference feedback, naturally giving rise to a dueling bandit formulation.

A critical, yet often overlooked factor in these applications is that the comparisons may incur heterogeneous costs; for instance, generating responses from different models can require vastly different computational or monetary resources. Despite this, existing dueling bandit algorithms almost universally assume uniform costs per comparison.

This gap motivates our study of cost-aware best-arm identification under dueling feedback. We consider a setting where each arm has a known sampling cost, and the cost of a comparison is the sum of the costs of the involved arms. Assuming the existence of a Condorcet winner, we characterize the instance-dependent optimal cost complexity for identifying the best arm. We derive a closed-form expression for the optimal cost allocation and propose a cost-aware Track-and-Stop algorithm that is δ-probably correct and asymptotically optimal. Finally, we validate our approach on synthetic data and real-world LLM evaluation benchmarks, demonstrating a consistent reduction in total sampling costs.

2 PROBLEM FORMULATION

We consider a general L-armed dueling bandit problem where a learner is given access to L arms, denoted by $\mathcal{L} := [L]$. Unique to this work, each arm $i \in \mathcal{L}$ has an associated sampling cost $c_i > 0$, where the vector of costs $C := (c_1, c_2, \ldots, c_L)$ is known apriori. At each time step $t = 1, 2, \ldots$, the learner selects a pair of arms $a_t := (a_{t,1}, a_{t,2})$ with $a_{t,1}, a_{t,2} \in \mathcal{L}$. Upon selecting this pair, the learner incurs a cost $c_{a_{t,1}} + c_{a_{t,2}}$ and observes the outcome of the duel between them. For each pair (i, j), the outcome is an independent Bernoulli random variable with probability $p_{i,j}$, where $p_{j,i} = 1 - p_{i,j}$.

We assume the existence of a Condorcet winner, i.e., an arm $a^* \in \mathcal{L}$ such that $p_{a^*,j} > 1/2$ for all $j \neq a^*$. The learner does not know the preference matrix $\mathbb{P}$ apriori and must identify a^* through sequential sampling of arm pairs. Given a confidence parameter $\delta \in (0, 1)$, an algorithm is said to be δ-probably correct (δ-PC) if

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/GEKA7634

$\mathbb{P}(\hat{a}_{\tau_\delta} \neq a^*) \leq \delta$, where τ_δ denotes its stopping time and $\hat{a}_{\tau_\delta}$ its recommendation. The objective is to design δ-PC algorithms that minimize the expected total sampling cost

$$J(\tau_\delta) := \sum_{t=1}^{\tau_\delta} (c_{a_{t,1}} + c_{a_{t,2}})$$

3 LOWER BOUND

Let $\mathcal{K} = \{(i, j) : 1 \leq i < j \leq L\}$ denote the set of arm pairs. For an instance $\nu = (C, \mathbb{P})$, let $\epsilon_{\text{alt}}(\nu)$ denote the set of valid alternative instances (having a Condorcet winner) with the same cost vector but a different best arm:

$$\epsilon_{\text{alt}}(\nu) = \{\nu' = (C, \mathbb{P}') : a^*(\nu') \neq a^*(\nu)\}.$$

A standard change-of-measure argument (as in [2]) yields an information-theoretic lower bound in terms of a min–max optimization over pairwise sampling weights and alternative instances [4, 6, 7]. Such bounds are typically not directly interpretable and the optimal sampling fractions rarely admit closed-form characterizations. In contrast, exploiting the Condorcet structure of the present formulation, we show that the optimal cost complexity simplifies to an explicit closed form as given in Theorem 3.1.

Theorem 3.1. *Given an error threshold $\delta \in (0, 1)$, let $J(\tau_\delta)$ denote the cost incurred by a δ-PC algorithm. Then*

$$\frac{\mathbb{E}_\nu[J(\tau_\delta)]}{\log(1/4\delta)} \geq \sum_{m \in \mathcal{L} \setminus \{a^*(\nu)\}} \frac{1}{\max_{n \neq m} \left(\frac{1}{c_m + c_n} \, d\big(p_{m,n}, \max(0.5, p_{m,n})\big) \right)}.$$

Theorem 3.1 provides an interpretable instance-dependent characterization of the optimal cost complexity and yields a closed-form description of the optimal sampling allocation. This structural simplification plays a central role in the design of our algorithm in the next section.

4 ALGORITHM

We now present the Dueling Bandit Cost-Aware Track-and-Stop (DCTAS) algorithm. Let $N_{i,j}(t)$ denote the number of duels of pair (i, j) up to time t, and $\hat{p}_{i,j}(t)$ the empirical preference estimates. Using $\hat{\mathbb{P}}(t)$ and the known costs C, DCTAS computes the estimated optimal pull fractions $\alpha(t)$ from the closed-form allocation characterized in Section 3.

Tracking rule. At time t, if there exists a pair $(i, j) \in \mathcal{K}$ such that $N_{i,j}(t) < \sqrt{t}$, a least-sampled pair is selected. Otherwise, DCTAS selects the pair whose empirical pull count most lags its target allocation:

$$a_t = \arg\min_{(i,j) \in \mathcal{K}} \left(N_{i,j}(t) - \sum_{s=1}^{t} \alpha_{i,j}(s) \right).$$

Stopping rule. To decide when to stop, DCTAS employs a Chernoff-style stopping rule. For arms i, j, let

$$Z_{i,j}(t) = \sum_{k \in \mathcal{L}} N_{j,k}(t) \, d\big(\hat{p}_{j,k}(t), \max\{0.5, \hat{p}_{j,k}(t)\}\big) - \sum_{k \in \mathcal{L}} N_{i,k}(t) \, d\big(\hat{p}_{i,k}(t), \max\{0.5, \hat{p}_{i,k}(t)\}\big).$$

Next, let $Z(t) = \max_{i \in \mathcal{L}} \min_{j \neq i} Z_{i,j}(t)$. The algorithm stops at

$$\tau_\delta = \inf\{t : Z(t) > \beta(t, \delta)\},$$

and returns the arm $i = \arg max_{i \in \mathcal{L}} \min_{j \neq i} Z_{i,j}(t)$.

Further, the following theorem establishes correctness.

Theorem 4.1. *For any $\delta \in (0, 1)$, DCTAS is δ-PC and satisfies*

$$\limsup_{\delta \to 0} \frac{J(\tau_\delta)}{\log(1/\delta)} \leq c^*(\nu),$$

where $c^*(\nu)$ is given in Theorem 3.1.

Thus, DCTAS asymptotically matches the cost lower bound, establishing optimal cost complexity under heterogeneous comparison costs.

5 NUMERICAL EXPERIMENTS

We evaluate the empirical performance of DCTAS against several baselines on both synthetic and real-world dueling bandit instances derived from LLM evaluation data.

Baselines. We compare with: (i) *TAS*, the cost-unaware Track-and-Stop obtained by setting uniform costs; (ii) *CRR*, a cost-aware round-robin scheme that selects the pair minimizing $N_{i,j}(t)(c_i + c_j)$ with the same stopping rule as DCTAS; (iii) *DPCA*, a cost-aware explore–verify algorithm based on Karnin [5]; and (iv) *DCTAC*, which uses the same tracking rule as DCTAS but a confidence-interval stopping rule.

Synthetic instance. We first consider a three-arm synthetic instance with a unique Condorcet winner and heterogeneous costs $C = (k, 1, 1)$, where $k \in [1, 19]$. Across 500 trials with $\delta = 0.01$, DCTAS consistently incurs lower identification cost than TAS. As the cost of the best arm increases, the optimal allocation shifts toward cheaper comparisons, and DCTAS adapts its sampling accordingly, while TAS continues to oversample expensive pairs.

Real-world datasets. We next evaluate on Chatbot Arena datasets [1], which provided pairwise preference outcomes and query costs for LLMs across text-to-image, text-to-text, vision, and search tasks. Across all datasets (100 trials, $\delta = 10^{-10}$), DCTAS consistently reduces cost relative to TAS and other baselines, confirming the benefit of cost-aware tracking under realistic LLM cost heterogeneity. Moreover, DCTAC achieves the lowest empirical cost, indicating that replacing the Chernoff stopping rule with a confidence-interval criterion can enable earlier termination in practice while preserving the advantages of the cost-aware sampling rule.

6 CONCLUSION

We studied cost-aware best-arm identification in dueling bandits motivated by heterogeneous LLM querying costs. Under a Condorcet assumption, we derived an explicit instance-dependent lower bound on optimal sampling cost and proposed DCTAS, a cost-aware Track-and-Stop algorithm with a Chernoff stopping rule. We showed that DCTAS is δ-PC and asymptotically cost-optimal. Further experiments on synthetic and real-world LLM preference datasets illustrate the benefits of cost-aware allocation under heterogeneous comparison costs.

Extended version. All proofs, detailed experimental results, and plots validating the claims are provided in the extended version of this paper.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support via the State Bank of India grant: "Dynamic LLM Optimization and Fine-Tuning via Dueling Bandits."

REFERENCES

[1] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132* (2024).

[2] Aurélien Garivier and Emilie Kaufmann. 2016. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*. PMLR, 998–1027.

[3] Kai Hui and Klaus Berberich. 2017. Low-cost preference judgment via ties. In *European Conference on Information Retrieval*. Springer, 626–632.

[4] Yassir Jedra and Alexandre Proutiere. 2020. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems* 33 (2020), 10007–10017.

[5] Zohar S Karnin. 2016. Verification based solution for structured mab problems. *Advances in Neural Information Processing Systems* 29 (2016).

[6] Riccardo Poiani, Marc Jourdan, Emilie Kaufmann, and Rémy Degenne. 2024. Best-arm identification in unimodal bandits. *arXiv preprint arXiv:2411.01898* (2024).

[7] Kota Srinivas Reddy, PN Karthik, Nikhil Karamchandani, and Jayakrishnan Nair. 2023. Best arm identification in bandits with limited precision sampling. In *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 1466–1471.

[8] Kevin Roitero, Alessandro Checco, Stefano Mizzaro, and Gianluca Demartini. 2022. Preferences on a budget: Prioritizing document pairs when crowdsourcing relevance judgments. In *Proceedings of the ACM Web Conference 2022*. 319–327.

[9] Xinyi Yan, Chengxi Luo, Charles LA Clarke, Nick Craswell, Ellen M Voorhees, and Pablo Castells. 2022. Human preferences as dueling bandits. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 567–577.

[10] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. 2012. The k-armed dueling bandits problem. *J. Comput. System Sci.* 78, 5 (2012), 1538–1556.

[11] Yisong Yue and Thorsten Joachims. 2009. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*. 1201–1208.