Cost-Aware Model Selection and Adaptive Reasoning in Large Language Models via Online Learning

Doctoral Consortium

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

ABSTRACT

Large Language Models (LLMs) have shown strong performance in generative tasks, yet challenges remain in their evaluation, alignment with human preferences, and improvement of reasoning under limited feedback and computational budgets. Addressing these challenges requires learning frameworks that operate under uncertainty and resource constraints, making online learning a natural foundation. However, many generative AI problems do not map directly to existing online learning formulations. My research bridges this gap by grounding generative AI problems in principled online learning frameworks while developing new theoretical insights. In recent work, we proposed a cost-aware Track-and-Stop–type algorithm for dueling bandits to identify the best model with minimum evaluation cost. Building on this, we study two further directions: (i) allocating a fixed training budget to fine-tune the optimal model from a set of candidates, which maps naturally to decreasing bandits with pseudo-regret minimization; and (ii) improving LLM reasoning by formulating hybrid test-time scaling as a Markov Decision Process, enabling principled allocation of inference-time computation.

KEYWORDS

Online Learning, Dueling Bandits, Reinforcement Learning, Large Language Models, Fine-tuning, Test-Time Scaling, MDPs

ACM Reference Format:
Sarvesh Gharat. 2026. Cost-Aware Model Selection and Adaptive Reasoning in Large Language Models via Online Learning: Doctoral Consortium. 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/WSXD8440

1 INTRODUCTION

Large language models (LLMs) have shown strong empirical performance across a wide range of generative tasks. As these models are increasingly deployed in practice, a natural question arises: given several available models, which one should be used for a particular task? In many settings, practitioners have access to a finite set of candidate models that differ in architecture, training data, and scale, yet lack a reliable way to determine which model will perform best from a human perspective. In practice, this decision is typically made through static evaluation on benchmarks or held-out datasets. Such evaluations rely on automated metrics and fixed test sets, which are increasingly affected by data contamination and leakage [11], particularly for large models trained on web-scale corpora [3, 5]. Moreover, benchmark scores often fail to capture aspects of model quality that matter to human users in open-ended generative tasks, such as usefulness, coherence, or reasoning quality [8]. As a result, benchmark-based rankings can be unreliable indicators of which model is actually preferred by humans.

This mismatch between automated metrics and human judgments motivates alternative evaluation paradigms. In open-ended generative tasks, it is often difficult to define calibrated scalar performance measures, whereas humans are far more reliable at expressing relative preferences between model outputs. Preference-based evaluation has therefore emerged as an alternative, where models are compared using human or human-aligned judgments rather than absolute scores [2, 12]. While this approach better reflects human preferences, it introduces new challenges. Collecting preference feedback is costly, and the cost of querying different models can vary substantially. Consequently, evaluation is often constrained by a limited budget, raising the question of how model comparisons should be allocated to identify the best model efficiently. This setting is inherently sequential, since feedback from earlier comparisons can guide which comparisons are most informative later on.

Identifying a suitable model, however, does not fully resolve the problem. In many applications, the selected model must be adapted to the target task through fine-tuning. Fine-tuning typically improves performance gradually as more compute is spent, and the final performance depends on how the available training budget is allocated. Importantly, the model that performs best before fine-tuning need not be the one that performs best after fine-tuning. When the training budget is limited, one must therefore decide both which model to fine-tune and how to allocate resources over time, without knowing in advance which choice will yield the best final performance. Existing approaches address this challenge using heuristic strategies, such as committing the entire budget to a single model chosen based on pre-training or evaluation performance, or applying early stopping rules [1]. These approaches do not address the more fundamental question of how close one can get, under a limited budget, to the performance that would be achieved if the optimal candidate were fine-tuned using the full budget.

A related issue arises at inference time, particularly when improving the reasoning capabilities of models. Recent work shows that allocating additional computation during inference—either through sequential test-time scaling [9], where reasoning is extended over multiple steps, or through parallel test-time scaling [10], where

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/WSXD8440

multiple reasoning paths are executed independently—can substantially improve reasoning performance. These approaches represent complementary ways of spending inference-time compute, but both incur non-trivial computational cost. In practice, inference-time budgets are often fixed, and existing methods rely on fixed or heuristic rules for deciding how much computation to use, without adapting to task difficulty or available resources.

In this work, we study these questions by treating model evaluation, adaptation, and inference-time reasoning as sequential decision problems under explicit budget constraints. Rather than relying on static benchmarks or fixed heuristics, we adopt an online learning perspective that enables decisions to be made adaptively based on observed feedback. We begin by studying how to identify the best model from a finite set using preference-based feedback under a limited evaluation budget. We then consider how to allocate a fixed training budget across candidate models to approach the performance of the best model that would be obtained under full fine-tuning. Finally, we study how inference-time computation can be allocated across sequential and parallel test-time scaling to improve reasoning performance under cost constraints. Together, these settings provide a unified view of how limited resources can be used effectively when interacting with large language models.

2 CURRENT RESEARCH

In our recent work [6], we study the problem of identifying the best model from a finite set under preference-based feedback and heterogeneous evaluation costs. This problem is motivated by large language model (LLM) evaluation settings, where automated metrics are often unreliable and querying different models can incur substantially different computational or monetary costs. We formalize it as a cost-aware dueling bandit with a fixed-confidence objective, where the learner observes noisy pairwise preferences between model outputs and must identify the Condorcet winner while minimizing the total comparison cost.

Our first contribution is an information-theoretic lower bound on the expected cost incurred by any δ-probably correct algorithm in this setting. Exploiting the structure induced by the Condorcet assumption, we show that this lower bound admits a closed-form characterization, in contrast to prior cost-aware bandit results that rely on implicit or numerically computed optimization problems. This characterization provides insight into how optimal sampling strategies should trade off statistical difficulty and evaluation cost across different model comparisons.

Building on this insight, we propose a cost-aware Track-and-Stop style algorithm that adaptively allocates comparisons to match the optimal cost proportions. A key technical challenge arises because the optimal allocation need not be unique, complicating direct extensions of classical tracking-based methods. We address this by designing a sampling rule that tracks a set of optimal allocations and a generalized likelihood ratio–based stopping rule tailored to best-arm identification under dueling feedback. We prove that the proposed algorithm is δ-probably correct and asymptotically achieves the lower bound on cost as $\delta \to 0$.

Finally, we validate our approach on both synthetic instances and real-world LLM evaluation datasets derived from head-to-head comparisons [4]. Across all settings, our method consistently reduces evaluation cost relative to cost-unaware and heuristic baselines, demonstrating the practical relevance of cost-aware preference-based model selection. This work provides a principled foundation for resource-efficient evaluation under preference feedback and serves as a starting point for studying adaptation and inference-time decision-making under explicit budget constraints.

3 ONGOING AND FUTURE DIRECTIONS

Building on our cost-aware preference-based evaluation results, we study two budgeted sequential decision problems: (i) allocating training actions and cheap inference samples to minimize pseudo-regret, and (ii) test-time scaling framed as an MDP for adaptive allocation of inference compute across reasoning steps under cost constraints.

3.1 Budgeted model adaptation with cheap inference feedback

In the first setting, we consider a finite set of K candidate models whose performance evolves through costly fine-tuning actions, where the best model after fine-tuning is not known a priori. At any intermediate training level, the learner may collect comparatively cheap inference-time samples to refine performance estimates before committing further training resources. The learner must therefore make sequential decisions that involve selecting a model and choosing between advancing training or gathering additional inference feedback. We formulate this interaction as an online learning problem with heterogeneous action costs and study policies that exploit cheap inference samples to reduce uncertainty and improve decisions. Preliminary analysis suggests that strategically allocating inference samples at intermediate training levels can significantly reduce pseudo-regret, even when no further training is performed.

3.2 Cost-aware test-time scaling via an MDP formulation

In the second setting, we study how inference-time computation should be allocated to improve model reasoning under a fixed budget. Prior work shows that increasing test-time compute can improve performance through sequential test-time scaling, which extends a single reasoning trace over multiple steps [9], and parallel test-time scaling, which executes multiple independent reasoning traces and aggregates their outputs [10]. Sequential scaling can exhibit diminishing returns and even performance degradation beyond a certain thinking budget due to error accumulation [7], while parallel scaling incurs higher computational cost and does not allow information sharing across traces.

Motivated by these complementary limitations, we study a hybrid test-time scaling strategy that adaptively combines sequential and parallel computation. We model the inference process as a Markov Decision Process (MDP), where each intermediate reasoning trace corresponds to a state, and actions correspond to either extending the current trace or branching into multiple parallel traces with different costs. This formulation enables adaptive allocation of inference-time compute based on the evolving reasoning trajectory and remaining budget, while capturing the trade-off between exploration via parallel branching and refinement via sequential extension.

REFERENCES

[1] Sebastian Pineda Arango, Fabio Ferreira, Arlind Kadra, Frank Hutter, and Josif Grabocka. 2023. Quick-tune: Quickly learning which pretrained model to finetune and how. *arXiv preprint arXiv:2306.03828* (2023).

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[3] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.

[4] 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).

[5] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[6] Sarvesh Gharat, Nikhil Karamchandani, and Jayakrishnan Nair. 2026. Cost-Aware Best Arm Identification via Dueling Feedback with Applications to Large Language Models. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*. International Foundation for Autonomous Agents and Multiagent Systems. https://doi.org/10.65109/WSXD8440

[7] Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. 2025. Does Thinking More always Help? Understanding Test-Time Scaling in Reasoning Models. *arXiv preprint arXiv:2506.04210* (2025).

[8] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters, and Malka N Halgamuge. 2025. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence* (2025).

[9] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 20286–20332.

[10] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? *arXiv preprint arXiv:2502.12215* (2025).

[11] Xin Zhou, Martin Weyssow, Ratnadira Widyasari, Ting Zhang, Junda He, Yunbo Lyu, Jianming Chang, Beiqi Zhang, Dan Huang, and David Lo. 2025. Lessleak-bench: A first investigation of data leakage in llms across 83 software engineering benchmarks. *arXiv preprint arXiv:2502.06215* (2025).

[12] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).