Truthful Reverse Auctions for Adaptive Selection via Contextual Multi-Armed Bandits

Pronoy Patra
IIIT, Hyderabad
Hyderabad, India
pronoy.patra@research.iiit.ac.in

Sankarshan Damle
Microsoft Research India
Bangalore, India
t-sandamle@microsoft.com

Manisha Padala
IIT, Gandhinagar
Gandhinagar, India
manisha.padala@iitgn.ac.in

Sujit Gujar*
IIIT, Hyderabad
Hyderabad, India
sujit.gujar@iiit.ac.in

ABSTRACT

We study the problem of selecting Large Language Models (LLMs) for user queries in settings where multiple LLM providers submit the cost of solving a query. From the user's perspective, choosing an optimal model is a sequential, query-dependent decision problem: high-capacity models offer more reliable outputs but are costlier, while lightweight models are faster and cheaper. We formalize this interaction as a reverse auction design problem with contextual online learning, where the user adaptively discovers which model performs best while eliciting costs from competing LLM providers. Existing multi-armed bandit (MAB) mechanisms focus on forward auctions and social welfare, leaving open the challenges of reverse auctions, provider-optimal outcomes, and contextual adaptation. We address this by designing a resampling-based procedure that generalizes truthful forward MAB mechanisms to reverse auctions and prove that any monotone allocation rule with this procedure is truthful. Using this, we propose a contextual MAB algorithm that learns query-dependent model quality with sublinear regret. Our framework unifies mechanism design and adaptive learning, enabling efficient, truthful, and query-aware selection of LLMs.

CCS CONCEPTS

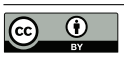
• **Theory of computation → Theory and algorithms for application domains**; **Solution concepts in game theory**; **Convergence and learning in games**.

KEYWORDS

Contextual MAB; Auction Design; Optimal Auction; Learning

ACM Reference Format:
Pronoy Patra, Sankarshan Damle, Manisha Padala, and Sujit Gujar. 2026. Truthful Reverse Auctions for Adaptive Selection via Contextual Multi-Armed Bandits. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/MBRQ7564

*This work was carried out under the ANRF under grant CRG/2022/004980.



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/MBRQ7564

1 INTRODUCTION

Large Language Models (LLMs) [3] increasingly shape the way individuals and organizations make decisions, collaborate, and innovate, reflecting their growing role as general-purpose AI systems [23, 27, 35, 41]. Beyond single-turn tasks, recent work deploys LLMs as *agentic* systems that plan, act, and interact with tools or other LLMs to complete complex objectives [14, 33, 44]. In both single-agent and multi-agent settings, *LLM Providers* (e.g., OpenAI [32] or Anthropic [4]) accelerate adoption by releasing multiple model variants that explicitly trade off capability, latency, and cost [5, 31]. Concretely, modern LLM deployments have established the problem of selecting an appropriate model for a given query. For example, GPT-5 employs a real-time router that dynamically allocates queries between a fast base model and a deeper reasoning model, leveraging signals such as task complexity, tool requirements, and explicit user intent [31]. However, these routing scenarios focus on internal model orchestration within a single provider, where the user may or may not have control over model selection or cost optimization.

Our Goal From the user's perspective, this trade-off remains central: the user derives *utility* from the model's performance while incurring a *cost* for using it. Intuitively, high-capacity models offer more accurate or reliable outputs but are more expensive, whereas lightweight models are faster and cheaper but may underperform on complex tasks. By framing the problem from the user's perspective, we aim to study how a user trades off the quality of a model's output against its usage cost. At the same time, we account for the LLM provider's ability to elicit the cost of solving a query. This formulation provides a foundation for analyzing optimal model selection strategies for the user and the design of mechanisms by the provider.

To formalize this, we focus on two complementary challenges: *elicitation* and *learning*. When the expected performance of available LLMs for a given query is known, the user faces an elicitation problem. Researchers model this as a *reverse auction* where providers submit costs for processing the query. In mechanism design, the objective may be to maximize social welfare, or the buyer's (user's) utility, the latter corresponding to *optimal auction design* [28]. If model performance were known deterministically, designing a user-optimal reverse auction would be straightforward. However, model performance is stochastic and context-dependent;

the user must learn it over time. This introduces a complementary learning challenge, formalized as a (contextual) sequential optimization problem, where the user adaptively identifies which provider performs best for each query. Together, these perspectives unify provider-side cost elicitation with user-side adaptive selection, forming the basis for truthful and efficient LLM allocation mechanisms.

Challenges. Addressing the combined elicitation and learning problem poses several challenges. Generally, researchers use multi-armed bandit (MAB) methods [10] as a standard framework for adaptive or sequential optimization problems. In the context of mechanism design, existing work on MAB mechanisms primarily focuses on traditional forward auctions that measure efficiency in terms of social welfare [7, 8]. These approaches leave two important gaps: (i) no prior work studies MAB mechanisms for reverse auctions, where the user acts as a buyer and multiple LLM providers bid to solve a query, and (ii) no prior work studies MAB mechanisms for optimal auctions that aim to maximize provider utility rather than social welfare. Moreover, to our knowledge, no prior work, except for [2], incorporates contextual information into these mechanisms. Our work addresses these gaps by designing a mechanism that is both truthful and optimal from the LLM provider's perspective, while enabling users to adaptively learn which model best solves each query via a contextual MAB framework.

Our Approach & Contributions

Our Approach. Babaioff et al. [8] and Devanur and Kakade [15] showed that any deterministic auction that directly embeds a MAB algorithm to learn stochastic parameters suffers substantially higher regret compared to its non-strategic counterpart. To address this, Babaioff et al. [8] introduced a randomized *resampling procedure* that preserves truthfulness while retaining the regret guarantees of the underlying MAB algorithm. The key idea is to randomize bids in a manner consistent with Bayesian incentive compatibility (BIC), provided that the MAB algorithm itself is monotone in allocations. However, this framework is limited to *forward* auctions optimizing social welfare. In contrast, we develop an analogous randomized adjustment procedure tailored to *reverse* auctions, where the objective is user-optimal (cost-minimizing) rather than welfare-maximizing. Once such a truthful reverse procedure is available, the remaining challenge lies in designing a contextual MAB algorithm that preserves monotonicity: ensuring that lowering the reported cost does not reduce allocation probability. To this end, we propose TRCM-UCB$_{\mathsf{OPT}}$: a truthful, provider-optimal, reverse contextual MAB algorithm that achieves truthfulness and an $O(\sqrt{T})$ regret bound in reverse, stochastic settings.

Our Contributions. **First,** we formalize the problem of selecting among competing LLM providers as an *optimal reverse contextual MAB mechanism*, where the user solicits costs from providers and adaptively learns which model best solves each query. **Second,** we adapt the theory of optimal auctions to this reverse setting, showing how provider-side incentives can be aligned with user-side selection in our context. **Third,** we design a novel resampling procedure (Algorithm 3), inspired by Babaioff et al. [7] but tailored to reverse, provider-optimal settings rather than to forward, social-welfare–maximizing ones. **Fourth,** we prove that any monotone allocation rule, combined with our resampling procedure, yields a truthful reverse MAB auction (Theorem 3). **Lastly,** we instantiate this result with TRCM-UCB$_{\mathsf{OPT}}$ (Algorithm 2), which leverages contextual information to learn model quality, achieving $O(\sqrt{T})$ regret across T queries. Together, these results establish the first truthful, reverse, contextual MAB auction for adaptive LLM model selection.

2 RELATED WORK

Multi-armed Bandits (MABs). The multi-armed bandit (MAB) problem [10, 37] is a canonical framework for sequential decision-making under uncertainty. Contextual variants [1, 24] extend this framework to settings where the reward distribution depends on observed contextual features.

Reverse Auctions. We model the LLM provider–user interaction as a reverse auction, building on the classic theory of optimal auctions [26, 28]. In a procurement or reverse auction, the buyer (the user, in our setting) seeks to minimize cost while ensuring truthful reporting by the sellers (the LLM providers). Several works extend Myerson's framework [28] to procurement settings [11] and to budget-feasible mechanisms [39]. However, contrary to our setup, these mechanisms generally assume deterministic valuations.

LLMs and Multi-Armed Bandits. Recent advances in LLMs motivate the need for model routing and selection, where a user or system dynamically decides which model to invoke for a given query. Early work focused on static model choice based on latency and cost trade-offs [19, 43], while recent methods use dynamic routers that predict query complexity and select among heterogeneous models [5, 16, 31, 33]. A new line of research also studies the intersection of LLMs with MABs [13, 36, 40, 42]. Dai et al. [13] formulate the multi-LLM selection problem as a cost-aware MAB, optimizing performance–cost trade-offs across heterogeneous LLMs under diverse reward models. Poon et al. [36] introduces the first contextual bandit framework for adaptive LLM selection under unstructured, dynamically evolving prompts, achieving sublinear regret while balancing accuracy and cost. Our formulation complements this line of work by introducing an incentive-aware framework: rather than training a router to minimize latency or maximize accuracy, we design a mechanism that elicits truthful costs from LLM providers.

MABs and Mechanism Design. In many real-world decision-making scenarios, mechanism design and online learning intersect when agents' private information influences the outcome of a learning process. Specifically, in multi-armed bandit (MAB) settings with strategic agents, the learner must balance exploration and exploitation while eliciting truthful information about privately known parameters that impact reward estimation or allocation. However, since these parameters are self-reported, agents may strategically misrepresent them to manipulate the learning process or improve their expected payoffs. To ensure truthful participation while maintaining efficient learning, MAB-based mechanism design integrates incentive-compatible auction mechanisms into the learning framework [7, 8, 15]. Classical formulations consider *forward auctions*, in which an auctioneer repeatedly allocates resources to agents with private values, aiming to maximize social welfare. Babaioff et al. [8] introduced a resampling-based procedure to achieve Bayesian incentive compatibility (BIC) in stochastic MAB environments, a result later extended and refined in [7, 15, 22]. Subsequent work

has explored combining bandit learning with auction design across domains such as crowdsourcing [21], smart grids [38], sponsored search auctions [17], and bidding strategy optimization for displaying ads [18], as well as multi-unit procurement [9]. However, these mechanisms primarily optimize social welfare and operate in forward auction settings. To our knowledge, only Abhishek et al. [2] incorporates contextual information into MAB mechanisms, still within a welfare-maximizing framework. In contrast, our work introduces the first reverse optimal MAB mechanism, aligning provider-side incentives with user-side adaptive learning while ensuring truthfulness and individual rationality. In summary, no prior work studies reverse, contextual MAB mechanisms in which the user is a buyer, eliciting private costs from multiple LLM providers as sellers.

3 PRELIMINARIES

3.1 Model

Setup. We consider a user $\mathfrak{U}$ who submits a sequence of queries over rounds $t = 1, 2, \ldots, T$. At each round t, the user issues a query q_t, and a set of strategic LLM providers, denoted by $[M] = \{1, 2, \ldots, M\}$, compete to process it. Each provider may differ in capability, latency, and cost, reflecting heterogeneous model architectures or inference configurations [5, 31].

Provider Costs and Bids. Each provider $i \in [M]$ has a true private cost ζ_i per token, representing its internal computation cost. For a query q_t, provider i would required $n_{i,q_t} \in \mathbb{N}$ tokens, the actual computation or latency cost incurred by provider i is $c_{i,t} = \zeta_i \cdot n_{i,q_t}$. The strategic provider submits a bid $b_{i,t}$ indicating the price to process the query (the provider submits bids per token; it can be easily converted to a bid for the query). The provider's goal would be to earn as much as possible. In general, the desirable goal in mechanism (or auction) design is to propose (i) an allocation and (ii) a payment rule that incentivizes providers to bid $b_{i,t} = c_{i,t}$.

Allocation ($\mathcal{A}$) and Payments ($\mathcal{P}$). At any round $t \in [T]$, the mechanism determines an allocation $\mathcal{A}_t(b_t)$

$$\mathcal{A}_t(b_t) = \{\mathcal{A}_{1,t}(b_t), \ldots, \mathcal{A}_{M,t}(b_t)\}, \qquad \sum_{i\in[M]} \mathcal{A}_{i,t}(b_t) \le 1,$$

where $\mathcal{A}_{i,t}(b_t) \in \{0, 1\}$ where 1 denotes that bidder/LLM i is selected given reported bids $b_t = (b_{1,t}, \ldots, b_{M,t})$. If provider i is selected, the user makes a payment $p_{i,t}(b_t)$ to i and others do not receive any payments. Let $\mathcal{P}_t : p_t(b_t) = (p_{1,t}(b_t), \ldots, p_{M,t}(b_t))$ capture the payment rule.

Query Context and User Valuation. At round t, query q_t is associated with observable features $x_t \in \mathbb{R}^d$, which may include query embeddings or user-specific attributes (e.g., OpenAI's saved memory feature [30]). Henceforth, we identify query q_t with x_t. For provider i, the expected value to the user of receiving a satisfactory response is modeled as

$$v_{i,t} = \mathbb{E}[r_{i,t} \mid x_t] = \theta_i^\top x_t,$$

where $\theta_i \in \mathbb{R}^d$ denotes a parameter vector specific to provider i, and $r_{i,t}$ denotes the realized response quality for query x_t.

Utilities. Let the user's total utility at round t be,

$$u_{0,t} := \sum_{i\in[M]} \mathcal{A}_{i,t}(b_t) \cdot \big(v_i - p_{i,t}(b_t)\big).$$

The utility of provider i is given by

$$u_{i,t} := \big(p_{i,t}(b_t) - c_{i,t}\big) \cdot \mathcal{A}_{i,t}(b_t) \tag{1}$$

Note. If θ_is are known, $\mathfrak{U}$'s objective at round t is designing an optimal reverse auction. For completeness, we show how to do it in the next subsection. For ease of exposition: (i) we assume n_{i,q_t} to be same for all queries. That is, $n_i = n_{i,q_t} \forall q_t$. Note that, as explained later, the results can be trivially extended for the case when n_{i,q_t}s are different. With this assumption, $c_{i,t}$ has the same support for all t, say $[\underline{c_i}, \overline{c_i}]$ with probability density function being f_i and cumulative distribution function being F_i. Additionally, $b_{i,t}$s become identical across all slots, and hence we can represent them as b_i. (ii) As we would be focusing on a single round, we omit the notation t from the discussion on Optimal Reverse Auction.

3.2 Optimal Reverse Auction

Auctions are typically of two types: (i) forward auction, wherein the auctioneer is a seller, and (ii) reverse auction, where the auctioneer is a buyer. The seminal work by Myerson [28] showed how to design an optimal forward auction; that is, an auction that maximizes the expected revenue. Researchers also extend these results for different reverse auction setups (e.g., Iyengar and Kumar [20]). Here, we show for our settings what an optimal reverse auction is.

As the bids of other LLM providers are random variables for an arbitrary LLM provider i, we work with expected allocations and payments where expectation is w.r.t. the bids of other providers.

Expected Allocation and Payments. We have,

$$\overline{\mathcal{A}_i}(b_i) = \mathbb{E}_{b_{-i}}[\,\mathcal{A}_i(b_i, b_{-i})\,], \qquad \overline{p_i}(b_i) = \mathbb{E}_{b_{-i}}[\,p_i(b_i, b_{-i})\,],$$

where the expectation is taken over the bid profile of all other providers, $b_{-i} = (b_1, b_2, \ldots, b_{i-1}, b_{i+1}, \ldots, b_m)$. The expected utility of provider i with true cost c_i is therefore

$$U_i(b_i|c_i) = -c_i \cdot \overline{\mathcal{A}_i}(b_i) + \overline{p_i}(b_i). \tag{2}$$

That is, the provider incurs a cost c_i to respond to the query irrespective of its bid b_i.

Mechanism Design Objective. The user seeks to maximize her utility u_0 subject to ensuring that providers are incentivized to truthfully reveal their private costs c_i. This requirement of *incentive compatibility* lies at the heart of mechanism design. Additionally, $\mathcal{A}, \mathcal{P}$ should ensure that no provider incurs loss.

Definition 1 (Bayesian Incentive Compatible [29])**.** *An LLM selection mechanism is Bayesian Incentive Compatible (BIC) if for every provider $i \in [M]$,:*

$$U_i(c_i|c_i) \ge U_i(b_i|c_i), \ \forall i \in [M]$$

Definition 2 (Ex Post Individual Rationality (EPIR) [29])**.** *The LLM selection mechanism satisfies Ex-post Individual Rationality (EPIR) if, for every provider $i \in [M]$ and for every possible vector of true costs $c = (c_1, \ldots, c_M)$ drawn from the support of the cost distributions, the utility of provider i is non-negative:*

$$u_i(c) = p_i(c) - c_i \cdot \mathcal{A}_i(c) \ge 0, \quad \forall i \in [M], \forall c \in C$$

where $C = \prod_i [\underline{c_i}, \overline{c_i}]$.

The overall problem may thus be viewed as designing an *optimal reverse auction* mechanism for allocating queries to providers. It is basically designing $\mathcal{A}, \mathcal{P}$ that maximize $\mathfrak{U}$'s expected utility and satisfy the above two properties.

With these definitions in place, we state Myerson's classical characterization of BIC mechanisms for the reverse auction setting.

Theorem 1 (Myerson's Characterization Theorem for Reverse Auction [28])**.** *A mechanism* $(\mathcal{A}, \mathcal{P})$ *is BIC* iff *it satisfies the following two conditions:*

(1) $\overline{\mathcal{A}_i}(\cdot)$ *is non-increasing for all* $i = 1, \ldots, n$.

(2) *The utility takes the form:*

$$U_i(c_i) = U_i(\overline{c_i}) + \int_{c_i}^{\overline{c_i}} \overline{\mathcal{A}_i}(s)\, ds \quad \forall c_i \in C_i;\ \forall i = 1, \ldots, n.$$

Building on this characterization, the optimal reverse auction can be derived. As these steps are analogous to Myerson's original design, we state the results directly. We begin by defining *virtual cost* for a provider i.

$$\Psi_i(c_i) \ = \ c_i + \frac{F_i(c_i)}{f_i(c_i)}, \tag{3}$$

and assume a *regularity condition* that $\Psi_i(\cdot)$ is strictly increasing in c_i[Myerson [28]]. Under regularity condition, an optimal reverse auction's allocation rule selects the provider that maximizes the positive virtual surplus $v_i - \Psi_i(c_i)$, i.e., selects

$$i^* \in \arg\max_i \{v_i - \Psi_i(c_i) \mid v_i - \Psi_i(c_i) \geq 0\}$$

Here, we assume $\{i : v_i - \Psi_i(c_i) \geq 0\}$ is non-empty. Else, the query cannot be assigned to any provider. The payment rule becomes:

$$p_i(c) = \mathcal{A}_i(c)\, c_i + \int_{c_i}^{\overline{c_i}} \mathcal{A}_i(c_{-i}, s_i)\, ds_i$$

Regularity implies a threshold form for the allocation: for fixed c_{-i} there exists a critical value $z_i(c_{-i})$ such that $\mathcal{A}_i(c_{-i}, s_i) = \mathbf{1}\{s_i \leq z_i(c_{-i})\}$. Hence the integral reduces to $\max\{z_i(c_{-i}) - c_i, 0\}$ and, if i wins, the payment becomes the critical threshold $p_i(c) = z_i(c_{-i})$. The mechanism is therefore BIC, individually rational and maximizes $\mathfrak{U}$'s utility. In summary,

Theorem 2 (Optimal Reverse Auction)**.** *Under the regularity assumption that each virtual cost* $\Psi_i(c_i) = c_i + F_i(c_i)/f_i(c_i)$ *is strictly increasing in* c_i*, a mechanism where*

(1) $\mathcal{A}$*: selects LLM provider* i^* *such that:*

$$i^* \in \arg\max_i \{v_i - \Psi_i(c_i) \mid v_i - \Psi_i(c_i) \geq 0\},$$

(2) $\mathcal{P}$*: pays the winner*

$$p_{i^*}(c) = z_{i^*}(c_{-i^*}), z_{i^*}(c_{-i^*}) = \sup\{s :\ v_{i^*} - \Psi_{i^*}(s) \geq \max_{j \neq i^*} \{v_j - \Psi_j(c_j)\}\}$$

and pays all others zero,

is an optimal auction for the user to procure LLM service.

Different n_{i,q_t} for different Queries. Note that, if n_{i,q_t} are different for each query, which is very natural, Ψ_i's would depend upon the query. In that case, we would have to replace Ψ_i as $\Psi_{i,t}$, which in turn depends upon n_{i,q_t}. The overall notation would become clumsy, and hence, we prefer to skip these technicalities for ease of exposition. Furthermore, to introduce an optimal reverse auction, we assumed $v_{i,t} = \mathbb{E}[r_{i,t}|x_t] = \theta_i^\top x_t$ to be known. However, in our LLM provider-user setup, θ_is are unknown but can be learnt.

3.3 Stochastic and Unknown Rewards

History and Learning. In practical settings, the reward obtained from selecting a provider is stochastic. The user $\mathfrak{U}$ is interested in its mean $v_{i,t} \triangleq \mathbb{E}[r_{i,t}|x_t] = \theta_i^\top x_t$ where $\theta_i \in \mathbb{R}^d$ is unknown to $\mathfrak{U}$ and $x_t \in \mathbb{R}^d$ captures the context of the query at round t. Let $r_{I_t,t}$ denote the realized reward from provider $I_t \in [M]$ at round $t \in [T]$. Here $I_t \in [M]$ denotes the provider selected for query x_t in round t. $\mathfrak{U}$ can estimate (learn) θ_i from historical observations. Let h_t capture the history till round t.

$$h_t \ = \ h_{t-1} \cup \{x_t, I_t, r_{I_t,t}\}, \qquad h_0 = \emptyset,$$

We model the interaction as a *contextual multi-armed bandit* (MAB) problem [25], in particular, the above settings is also referred to as linear contextual MAB. Let $\mathcal{ALG}$ be the algorithm that learns θ_is and assigns a new query x_t to one of the providers based on x_t and h_{t-1}.

The quality of a contextual bandit algorithm $\mathcal{ALG}$ is typically measured by its *regret*, defined as the difference between the cumulative expected reward of an oracle that always selects the optimal provider (given full knowledge of θ_i's), and that of the learning algorithm. Formally, if $i_t^* \in \arg\max_i \theta_i^\top x^t$ denotes the optimal provider at round t, the cumulative regret up to horizon T is

$$\mathbb{R}_T(\mathcal{ALG}) = \sum_{t=1}^{T} \left[\left(\theta_{i_t^*}^\top x_t - \psi_{i_t^*}\left(b_{i_t^*}\right) \right) - \left(\theta_{I_t}^\top x_t - \psi_{I_t}\left(b_{I_t}\right) \right) \right], \tag{4}$$

where I_t is the provider chosen by the learning algorithm $\mathcal{ALG}$ at round t. The objective is to design allocation mechanisms that achieve sublinear regret (lower the better), i.e., $R(T) = o(T)$, ensuring asymptotic convergence to the optimal allocation. By appropriately defining rewards, one can easily adapt LinUCB [24] or SupLinUCB [12] for our settings. LinUCB-based algorithm would incur $\Omega(T)$ regret, though it can practically perform better. SUP-LinUCB-based algorithm would incur $O(\sqrt{T})$ regret. [1] The challenge in the SUP-LinUCB is that, when the costs are private, the providers can manipulate the learning algorithm. Thus, there is need to carefully design MAB algorithm and payments [8, 15]. Such an auction design is called *MAB Mechanism Design*.

MAB Mechanism Design. Babaioff et al. [8], Devanur and Kakade [15] address this challenge by first showing that any deterministic truthful MAB mechanism must be exploration-separated, and inevitably suffer a regret in the order of $\Omega(T^{2/3})$. That is, there exists a fundamental trade-off between truthfulness and learning efficiency. To overcome this limitation, Babaioff et al. [7] propose a randomization via random resampling of bids, and demonstrate that it is possible to retain regret guarantees of a *monotone* learning algorithm in the presence of the strategic providers. However, their resampling procedure is designed for forward auctions and aims to optimize social welfare. Contrarily, our goal is to optimize the user's utility, i.e., design an optimal reverse auction.

[1] For completeness, we provide the formal algorithms in the complete version [34].

4 OUR APPROACH

In a nutshell, to design an optimal reverse auction with learnable rewards, we require a monotone allocation rule for contextual MAB with an appropriately defined reward function and a resampling procedure. Towards this, we begin by defining monotonicity of the allocation rule. Then, in Section 4.2, we propose two algorithms: REV-BaseLinUCB-S and REV-SupLinUCB-S-OPT adapted from SupLinCUB [12] to ensure monotonicity. Section 4.3 introduces our resampling procedure, ROSA, and proves that any monotone allocation rule through a contextual MAB algorithm ($\mathcal{ALG}$) combined with ROSA ensures truthful reporting of the costs by the providers while retaining regret guarantees of $\mathcal{ALG}$. Due to space constraints, we omit formal proofs of some of our results. The proofs are available in the complete version [34].

4.1 Ex-Post Monotonicity

The ex-post monotonicity condition ensures that the allocation rule respects the economic intuition that bidding more competitively (i.e., lowering one's declared cost) should not disadvantage a provider in terms of allocation opportunities. More formally,

Definition 3 (Ex Post Monotone). *An allocation rule $\mathcal{A}$ is* ex-post monotone *for a reverse auction if, for every possible sequence of context arrivals $(x_1, x_2, \ldots, x_T)$ and reward realizations, for each provider $i \in [M]$, for all bids b_{-i} of other providers, and for any two possible bids of provider i, $b'_i \leq b_i$, we have:*

$$\mathcal{A}_i(b'_i, t) \geq \mathcal{A}_i(b_i, t) \quad \text{for all rounds } t.$$

Here, $\mathcal{A}_i(b, t)$ denotes the total number of times provider i is allocated in the first t rounds when the bid vector is b.

This property establishes the truthfulness of mechanisms in our stochastic reverse auction setting.

4.2 Monotone MAB Algorithms

In the literature, two of the most popular MAB algorithms for contextual bandits are LinUCB [24] and SupLinUCB [12]. We can leverage these algorithms using net rewards as, $v_{i,t} - \Psi_i(b_i)$; $\Psi_i(b_i)$ is subtracted as $\mathfrak{U}$'s goal is to deploy optimal reverse auction, which selects a provider having the highest $v_{i,t} - \Psi_i(b_i)$. However, with just this update, the allocation won't be monotone. To this extent, we propose REV-BaseLinUCB-S (Algorithm 1) and REV-SupLinUCB-S-OPT (Algorithm 2) that provide monotone allocations.

Difference between REV-SupLinUCB-S-OPT and SupLinUCB. Broadly, REV-SupLinUCB-S-OPT works similarly to SupLinUCB. I.e., it works in stages. Each stage consists of twice the previous number of rounds, thus a total of $O(\log T)$ stages. In each stage, the algorithm maintains an active set of providers. As the stage progresses, the active set continues to eliminate suboptimal providers. However, unlike SupLinUCB, which considers a common θ, that is θ_is are same, the context for each provider is different. As such, REV-SupLinUCB-S-OPT has independent θ_is to be learned for each provider i.

To make SupLinUCB monotone, we decouple learning from the auction by limiting the learning only by the round robin method. Compared to SupLinUCB, we select a provider from the active set one by one (refer to Line 7 of Algorithm 2). Also, we add an exploitation condition, Lines 25-27, wherein we exploit within the current stage. We consider the reward with the confidence subtracted by the virtual cost to match it with the optimal reverse auction setting (as described in Section 3.2). I.e., the selection is done using $\hat{v}_{i,t} + w^s_{i,t} - \Psi_i(b_i)$; $w^s_{i,t}$ is the confidence term in UCB. We defer the proof of the monotonicity of REV-SupLinUCB-S-OPT to Section 5.

Algorithm 1 REV-BaseLinUCB-S

Input: Parameter $\alpha \in \mathbb{R}^+$, History sets $\{H_{i,t}\}_{i\in[M]}$, where $\Lambda_{i,t} \subseteq \{1, 2, \ldots, t-1\}$
Observe context vector $x_t \in \mathbb{R}^d$
for each provider $i \in [M]$ **do**
  $A_{i,t} \leftarrow I_d + \sum_{\tau\in H_{i,t}} x_\tau x_\tau^T$
  $g_{i,t} \leftarrow \sum_{\tau\in H_{i,t}} r_{i,\tau} x_\tau$
  $\hat{\theta}_{i,t} \leftarrow A^{-1}_{i,t} g_{i,t}$
  // Calculate value estimate and confidence width
  $\hat{v}^s_{i,t} \leftarrow \hat{\theta}^\top_{i,t} x_t$
  $w^s_{i,t} \leftarrow \alpha\sqrt{x_t^T A^{-1}_{i,t} x_t}$
end for
Return: $\{\hat{v}^s_{i,t}\}_{i\in[M]}$ and $\{w^s_{i,t}\}_{i\in[M]}$

4.3 ROSA: Resampling Procedure

To avoid exploration-separatedness for truthful mechanisms that incur higher regret, we need a randomized mechanism. Towards this, we randomly update the providers' bids using a *resampling procedure*. Motivated by [7, Definition 4.3], we formally define it for our setting next.

Definition 4 (Reverse Self-Resampling Procedure). *Let I be a nonempty interval in $\mathbb{R}$. A* reverse self-resampling procedure *with support I and resampling probability $\mu \in (0, 1)$ is a randomized algorithm with input $b_i \in I$, random seed rs_i, and output $\widetilde{b_i}(b_i; rs_i) \in I$, that satisfies the following properties:*

(1) *For every fixed rs_i, $\widetilde{b_i}(b_i; rs_i)$ is non-decreasing in b_i.*
(2) *With probability $1 - \mu$, we have $\widetilde{b_i}(b_i; rs_i) = b_i$. Otherwise, $\widetilde{b_i}(b_i; rs_i) > b_i$.*
(3) *Consider the two-variable function*

$$F(a_i, b_i) = \Pr[\, \widetilde{b_i}(b_i; rs_i) < a_i \mid \widetilde{b_i}(b_i; rs_i) > b_i \,], \forall\, a_i > b_i$$

which is called as the distribution function *of the reverse self-resampling procedure. For each b_i, the function $F(\cdot, b_i)$ must be differentiable and strictly increasing on the interval $I \cap (b_i, \overline{c_i})$.*

Algorithm 3 presents a construction for a reverse auction with the self-sampling procedure from Definition 4.

Proposition 1. *ROSA is a reverse self-resampling procedure with support $\mathbb{R}^+$ and resampling probability μ. The distribution function for this procedure is $F(a_i, b_i) = \frac{a_i - b_i}{\overline{c_i} - b_i}$.*

Proof. Properties 1 and 2 in Definition 4 for ROSA follow immediately from the description of Algorithm 3. For Property 3, by conditioning on the event $\widetilde{b_i}(b_i; rs_i) > b_i$, we see that the distribution of $\widetilde{b_i}(b_i; rs_i)$ is uniform in $[b_i, a_i]$. □

Algorithm 2 REV-SupLinUCB-S-OPT

Input: Virtual costs $\Psi(b) = (\Psi_1(b_1), \ldots, \Psi_n(b_n))$
Initialization:
$S_{\max} \leftarrow \lceil \ln T \rceil$
$\Lambda^s_{i,t} \leftarrow \emptyset$ for all $i \in [M], t \in [T], s \in [S_{\max}]$
for $t = 1, 2, \ldots, T$ **do**
  $s \leftarrow 1$ and $\hat{A}_1 \leftarrow [M]$
  $j \leftarrow 1 + (t \mod n)$
  repeat
    Use REV-BaseLinUCB-S with index sets $\{\Lambda^s_{i,t}\}_{i\in[M]}$ and context x_t to get $(\hat{v}^s_{i,t})_{i\in\hat{A}_s}$ and $(w^s_{i,t})_{i\in\hat{A}_s}$.
    if $j \in \hat{A}_s$ and $w^s_{j,t} > 2^{-s}$ **then**
      // Forced exploration
      Select $I_t = j$
      Update history sets for all stages $s' \in [S_{\max}]$:
      $$\Lambda^{s'}_{i,t+1} \leftarrow \begin{cases} \Lambda^{s'}_{i,t} \cup \{t\} & \text{if } s = s' \\ \Lambda^{s'}_{i,t} & \text{otherwise} \end{cases}$$
    else if $w^s_{i,t} \le \frac{1}{\sqrt{T}}\ \forall\ i \in \hat{A}_s$ **then**
      // Pure exploitation (final phase)
      Select $I_t = \arg\max_{i\in\hat{A}_s} \left\{(\hat{v}^s_{i,t} + w^s_{i,t}) - \Psi_i(b_i)\right\}$
      Update index sets for I_t at all stages:
      $\Lambda^{s'}_{I_t,t+1} \leftarrow \Lambda^{s'}_{I_t,t} \quad \forall s' \in [S]$
      $\Lambda_0 \leftarrow \Lambda_0 \cup \{t\}$
    else if $w^s_{i,t} \le 2^{-s}\ \forall\ i \in \hat{A}_s$ **then**
      // Stage advancement
      $\hat{A}_{s+1} \leftarrow \{i \in \hat{A}_s \mid$
        $(\hat{v}^s_{i,t} + w^s_{i,t}) - \Psi_i(b_i) \ge$
        $\max_{a\in\hat{A}_s} \{(\hat{v}^s_{a,t} + w^s_{a,t}) - \Psi_a(b_a)\} - 2\cdot 2^{1-s}\}$
      $s \leftarrow s + 1$
    else
      // Exploitation within the current stage
      Select $I_t = \arg\max_{i\in\hat{A}_s} \left\{(\hat{v}^s_{i,t} + w^s_{i,t}) - \Psi_i(b_i)\right\}$
      $\Lambda^s_{est} \leftarrow \Lambda^s_{est} \cup \{t\}$
    end if
  until I_t is selected
end for

Algorithm 3 ROSA: Reverse OneShot Adjustment, A Non-Recursive Self-Resampling Procedure for Reverse Auction

Require: Bid $b_i \in [0, \infty)$, resampling parameter $\mu \in [0, 1]$, upper cost bound $\overline{c_i}$, random seed rs_i
Sample a random variable γ_i uniformly from $[0, 1]$
Set resampling factor
$$\xi_i = \begin{cases} 1 & \text{with probability } 1-\mu \\ 1 + \gamma_i\left(\frac{\overline{c_i}}{b_i} - 1\right) & \text{with probability } \mu \end{cases}$$
Construct and return the modified bid $\widetilde{b_i}(b_i, rs_i) = \xi_i b_i$

Note that, typically, for each provider we need to use a different random seed rs_i. Hence, for each provider, the resampling procedure is treated as rp_i. Let **RP** be an ensemble of these procedures.

We now illustrate that with a monotone MAB algorithm $\mathcal{ALG}$, one can design a truthful reverse auction using ROSA.

Suppose we are given a monotone allocation rule $\mathcal{ALG}$ and a vector **RP** of reverse self-resampling procedures of each provider $i \in [M]$ that has a resmapling probability $\mu \in (0, 1)$, support $\mathcal{T}_i$,

Rev-GTM: Generic Reverse Transformation Mechanism

Input: Bid vector $b = (b_1, b_2, \ldots, b_M)$
Output: Allocation and payment $(\mathcal{A}, \mathcal{P})$
for each provider $i \in [M]$ **do**
  Obtain $\widetilde{b_i} = \text{ROSA}(b_i, \mu, \overline{c_i}, rs_i)$
end for
Construct modified bid vector $\widetilde{b} = (\widetilde{b_1}, \widetilde{b_2}, \ldots, \widetilde{b_M})$
Allocation: Allocate according to monotone rule $\mathcal{ALG}(\widetilde{b})$
Payment:
for each provider $i \in [M]$ **do**
  if $\xi_i = 1$ **then**
    Set $p_i \leftarrow b_i \cdot \mathcal{A}_i(x)$
  else
    Set $p_i \leftarrow b_i \cdot \mathcal{A}_i(x) + \frac{1}{\mu} \cdot \mathcal{A}_i(x) \cdot (\overline{c_i} - b_i)$
  end if
end for

and output value $\widetilde{b_i}(b_i; rs_i)$. Based on the resampling procedure, we propose a generic transformation that converts $\mathcal{ALG}, \mu, \mathbf{RP}$ into a randomized mechanism $\mathcal{M} = \text{Rev-GTM}(\mathcal{ALG}, \mu, \mathbf{RP})$. Next, we introduce desirable properties of any randomized mechanism.

Definition 5 (Truthfulness in Expectation (Ex-post Incentive Compatibility, EPIC)). *For every provider $i \in [M]$ and for every realization of other providers' bids b_{-i}, truthful reporting of its cost $b_i = c_i$ maximizes the provider's expected utility over the mechanism's internal randomness. That is, $\forall i \in M, \forall t$*

$$u_{i,t}(c_i, b_{-i}|c_i) \ge u_{i,t}(b_i, b_{-i}|c_i) \forall b_i, \forall b_{-i}, \forall h_{t-1}$$

Definition 6 (Universal Ex-post Individual Rationality (EPIR)). *For every realization of the mechanism's randomness, each provider's realized utility is non-negative when bidding truthfully, i.e., $\forall i \in [M], \forall t, \forall h_{t-1}$*

$$u_{i,t}(c_i, b_{-i}|c_i) \ge 0.$$

Now, we state an important result of this section.

Theorem 3. *Consider an arbitrary single-parameter reverse auction domain, and let $\mathcal{ALG}$ be a monotone allocation rule. Suppose we are given an ensemble* **RP** *of self-resampling procedures, where each procedure has resampling probability $\mu \in (0, 1)$. Let the mechanism $\mathcal{M} = (\mathcal{A}, \mathcal{P}) = \text{Rev-GTM}(\mathcal{ALG}, \mu, \mathbf{RP})$ be the mechanism constructed from $\mathcal{ALG}$ via the self-resampling transformation. Then the mechanism $\mathcal{M}$ satisfies the following properties:*

(1) *EPIC, EPIR*
(2) *For m providers and any bid vector b (and any fixed random seed of nature), allocations $\mathcal{A}(b)$ and $\mathcal{ALG}(b)$ are identical with probability at least $1 - M\mu$.*
(3) *If $T = \mathbb{R}^M_+$ (all types are positive), and each $\mathbf{RP}_i$ is the reverse self-resampling procedure, then mechanism $\mathcal{M}$ is ex-post no-positive-transfers, and never pays any provider i more than*

$$b_i \cdot \mathcal{ALG}_i(\widetilde{b}) + (\overline{c_i} - b_i) \cdot \mathcal{ALG}_i(\widetilde{b}) \left(\frac{1}{\mu}\right).$$

Proof Sketch. **(1)** The argument follows the structure of [7]. It suffices to show: (i) the allocation rule $\mathcal{A}$ is monotone, and (ii) the payment rule satisfies Myerson's characterization.

Monotonicity of $\mathcal{A}$ follows from the monotonicity of $\mathcal{ALG}$ and Property 1 of the reverse self-resampling procedure, which ensures that increasing a bid cannot reduce the probability of allocation.

For the payment, the expected payment of provider i is

$$\mathbb{E}[p_i(b)] = \mathbb{E}[b_i \mathcal{ALG}_i(\widetilde{b})] + \mathbb{E}[R_i] = b_i \mathcal{A}_i(b) + \mathbb{E}[R_i].$$

Hence, to match Myerson's characterization, it remains to show

$$\mathbb{E}[R_i] = \int_{c_i}^{\overline{c_i}} \mathcal{A}_i(b_{-i}, u)\, du.$$

This equality holds because the random resampling process, unbiasedly estimates the integral term in Myerson's formula.

Individual rationality follows directly since the expected utility of each provider is non-negative under truthful reporting and monotone allocation.

(2) Since each bid is resampled independently with probability μ, the probability that no bids are resampled equals $1 - M\mu$.

(3) Finally, ex-post no-positive-transfer holds directly from the payment rule. □

TRCM-UCB$_{\mathsf{OPT}}$

We propose that the user $\mathfrak{U}$ deploys the auction as follows: *First*, it collects bids from all LLM providers and updates them using ROSA. *Second*, with updated bids, for every new query, it selects the provider recommended by REV-SupLinUCB-S-OPT and the payments are computed as given Rev-GTM. We refer to such an auction as TRCM-UCB$_{\mathsf{OPT}}$. In summary, TRCM-UCB$_{\mathsf{OPT}}$= Rev-GTM(REV-SupLinUCB-S-OPT, μ, **RP**).

5 TRCM-UCB$_{\mathsf{OPT}}$: THEORETICAL ANALYSIS

We now prove that REV-SupLinUCB-S-OPT is monotone and hence from Theorem 3, TRCM-UCB$_{\mathsf{OPT}}$ is truthful reverse contextual MAB auction that optimizes the user's utility.

5.1 Monotonicity of TRCM-UCB$_{\mathsf{OPT}}$

Theorem 4. *The allocation rule induced by REV-SupLinUCB-S-OPT is ex-post monotone.*

Proof Sketch. Submitting a lower bid b_i^-, provider i ensures it will remain in the active set for at least as many rounds and stages as it would have with the higher bid b_i. During any exploitation step, the provider l with the highest $(\hat{v}_{l,t}^k + w_{l,t}^k) - \Psi_l(b_l)$ is chosen. As $(\hat{v}_{i,t}^k + w_{i,t}^k) - \Psi_i(b_i^-) > (\hat{v}_{i,t}^k + w_{i,t}^k) - \Psi_i(b_i)$ while $(\hat{v}_{j,t}^k + w_{t,t}^k) - \Psi_j(b_j)$ for $j \neq i$ remains unchanged, provider i is strictly more competitive when it lowers its bid. Since a lower bid guarantees that an provider remains eligible for selection for at least as long and makes it a more competitive choice in every selection round, the total number of allocations $\mathcal{A}_i(b_i^-, T)$ must be greater than or equal to $\mathcal{A}_i(b_i, T)$. □

5.2 Regret analysis of REV-SupLinUCB-S-OPT

In this subsection, we prove that REV-SupLinUCB-S-OPT incurs the regret $O(\sqrt{T})$. The proof is similar to SupLinUCB, as our algorithm is derived from it. However, it has certain differences as we need to ensure monotonicity. Still, the regret dependency on T remains the same, albeit, constant factors increase (for REV-SupLinUCB-S-OPT, dependency on the number of providers is quadratic in M, whereas for SupLinUCB, it is $\sqrt{M}$).

The difference in the final regret proof is due to (i) rewards are also bid dependent, which is not the case with [12]. (ii) We need monotonicity, and hence, our active sets retain the providers in the active set for more rounds than that in SupLinUCB. (iii) We need to have an optimal reverse auction To claim the regret guarantees, we need the following lemma.

Lemma 1. *With probability at least $1 - \kappa$, for any round $t \in [T]$ and any stage $s \in [S_{\max}]$, the following hold:*

(1) *For all $i \in [M]$,*

$$(\hat{v}_{i,t}^s + w_{i,t}^s) - \Psi_i(b_i) \;\geq\; v_{i,t} - \Psi_i(b_i) \;\geq\; (\hat{v}_{i,t}^s - w_{i,t}^s) - \Psi_i(b_i).$$

(2) *The optimal provider is never eliminated: $i_t^* \in \hat{A}_s$.*

(3) *For any $i \in \hat{A}_s$,*

$$\left(v_{i_t^*,t} - \Psi_{i_t^*}(b_{i_t^*})\right) - \left(v_{i,t} - \Psi_i(b_i)\right) \leq 2^{3-s}.$$

Proof Sketch. The first part follows directly from Lemma 4 of Chu et al. [12] by subtracting the virtual cost term from both sides of the inequality. For the second part, we use the confidence bound $w_{i,t}^s \leq 2^{-(s-1)}$, which follows from the algorithm's update rule. Combining this bound with Part (1) implies that the optimal provider never satisfies the elimination criterion, and thus remains in the active set. Finally, Part (3) applies Part (1) and the elimination condition to bound the instantaneous regret at round t. □

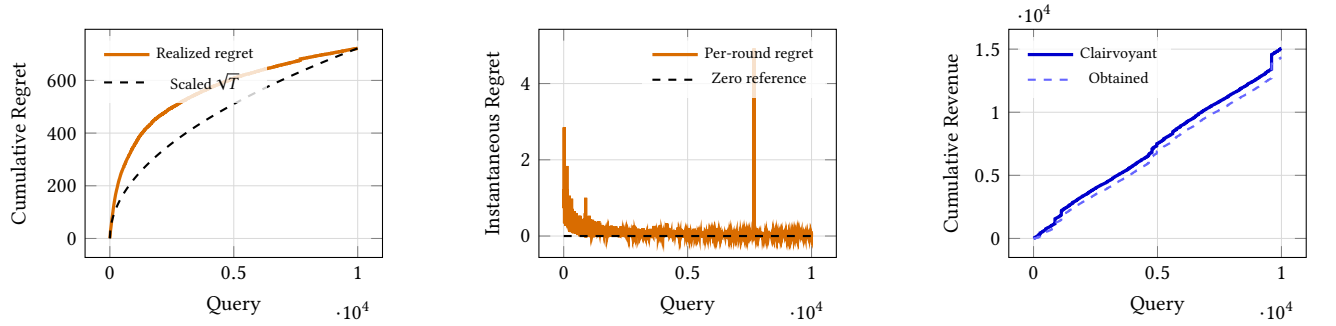
We restrict learning to round-robin ordering only (Lines 10-13, Algorithm 2) and add an additional decision rule for provider selection (Line 27). Note that this additional rule was not used in Chu et al. [12]. Hence, the main challenge is to bound the number of rounds of provider selection, which is done using this decision rule. (We refer to it as Λ_{est}^s in our analysis.) To this extent, we introduce some notations also mentioned in the algorithm:

Let Λ_0 be the set of rounds in which the provider in the pure exploitation phase was selected (Lines [14-17]). Let Λ_{est}^s be the set of rounds the provider was selected in exploitation in the current phase (Lines [25-27]), and let $\Lambda_{T+1}^s = \bigcup_i \Lambda_{i,T+1}^s$. Next, the expression in Claim 1 will be useful to bound the regret.

Claim 1. *At each stage s, $|\Lambda_{est}^s| \leq (n-1) \cdot |\Lambda_{T+1}^s|$.*

Proof. Let's consider M consecutive rounds for any stage s. Assume that provider selection is done as per Lines[25-27] for each of the M rounds. Note that provider selection in this decision block is done if and only if there exists a provider j such that $j \neq k$ (where k is the designated provider for the round), and $w_{j,t}^s > 2^{-s}$. After M consecutive rounds, each provider has received its designated round (as in Line[7]). Hence, if for some provider k, $w_{k,t}^s > 2^{-s}$, then this provider k should be selected on its designated round. Assuming selection in Lines [25–27] occurs for M consecutive rounds leads to a contradiction. More details are in the full version [34]. □

Theorem 5. *REV-SupLinUCB-S-OPT has regret $O\left(M^2\sqrt{dT\ln T}\right)$ with probability at least $1 - \kappa$, if it is run with $\alpha = \sqrt{\frac{1}{2}\ln\frac{2TM}{\kappa}}$.*



(a) Average Cumulative realized regret with scaled $\sqrt{T}$ reference.

(b) Realized per-round regret showing convergence towards zero.

(c) Cumulative actual Revenue vs. clairvoyant benchmark.

Figure 1: Regret and revenue curves for TRCM-UCB$_{\text{OPT}}$ averaged over $N = 40$ random seeds.

Proof Sketch. The proof follows the structure of Theorem 6 in Auer [6], with additional terms introduced to account for differences. Primarily, as our algorithm introduces two new components: round-robin exploration and within-stage exploitation—to ensure monotone MAB allocation, we need to show that with high probability, active sets $\hat{A}_{ss}$ cannot eliminate an optimal provider (proved in Lemma 1). We need to bound that even if $\hat{A}_{ss}$ contains more agents than those in SupLinUCB at any round, (a) the number of times bad sub-optimal providers are used is bounded by a constant with high probability. (b) The sub-optimal arms in active sets may be pulled more frequently, but the regret incurred due to such pulls is bounded. Claim 1 provides an upper bound on how many times a bad sub-optimal provider is used during the within-stage exploitation phase. Lemma 1 provides reward guarantees of a sub-optimal arm in the active set. Finally, the cumulative regret is obtained by summing the instantaneous regret across all selection rounds, and applying Lemma 1 together with Claim 1 completes the bound. □

From Theorems 2, 3, 4, and 5, the reward structure used for arms in REV-SupLinUCB-S-OPT and the definition of TRCM-UCB$_{\text{OPT}}$, we conclude,

Corollary 1. *TRCM-UCB$_{\text{OPT}}$, is EPIC, EPIR, optimal reverse auction that learns θ_is with regret guarantee of $O(\sqrt{T})$.*

6 EXPERIMENTAL ANALYSIS

We study the performance of our mechanism on synthetic data.

Simulation environment. We use $M = 4, d = 5,$ and $T = 100K$, unless noted otherwise. For each round t, we select x_t from a Gaussian distribution [2]: $x_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Here, Σ, the covariance matrix across d-dimensions of the contexts, blends a diagonal scale of 0.2 with off-diagonal correlations of 0.05. Four strategic providers submit bids drawn from provider-specific log-normal virtual-value models: for provider i, the latent parameters (μ_i, σ_i) depend affinely on x_t through a learned context matrix and a scale factor of 0.15. We fix the exploration parameter of REV-SupLinUCB-S-OPT at $\alpha = 0.75$ and regularize the Gram matrices with $\lambda = 1.0$.

Experiments. To obtain statistically stable estimates, we run each configuration over $(N = 40)$ independent random seeds. We fix the μ_is and generate the b_is, contexts, realized rewards, and randomizations for each run.

Implementation Details. The algorithms are implemented in Python 3.13 with `numpy`, `pandas`, and `matplotlib`. All experiments are executed on a single workstation (Windows 11, 32 GB RAM); a 40-run sweep at $T = 100K$ completes in roughly 45s.

Results. We evaluate TRCM-UCB$_{\text{OPT}}$ over $N = 40$ independent random seeds and summarize the averaged trajectories in Figure 1. In the figure, (a) plots the cumulative realized regret together with a scaled $\sqrt{t}$ guide curve. It is clear that our algorithm incurs regret $O(\sqrt{T})$. (b) reports the per-round realized regret, which contracts toward zero. (c) compares the cumulative actual revenue with the optimal revenue that a clairvoyant algorithm–that can predict the expected rewards, i.e., having access to θ. From the figure, it is clear that TRCM-UCB$_{\text{OPT}}$ quickly approaches the performance of the clairvoyant algorithm.

7 CONCLUSION

We introduced the first truthful optimal reverse contextual MAB mechanism for adaptive LLM model selection, integrating provider-side cost elicitation with user-side learning. Our randomized resampling procedure, ROSA, extends prior forward-auction frameworks to reverse, cost-minimizing settings while preserving truthfulness and sublinear regret. The resulting TRCM-UCB$_{\text{OPT}}$ mechanism aligns incentives between users and LLM providers and achieves $O(\sqrt{T})$ regret. More broadly, our work establishes a foundation for mechanism design in multi-model AI ecosystems, where learning and economic incentives must co-evolve.

Future Work. Future extensions could incorporate minimum thresholds on the costs reported by LLM providers. Another promising direction is to handle uncertainty in query token requirements and user-side budget limits. Incorporating such practical and external constraints would enhance the applicability of our reverse optimal contextual MAB framework to real LLM marketplaces.

Acknowledgments. Sujit Gujar gratefully acknowledges support from ANRF under grant CRG/2022/004980.

[2]Complete version [34] depicts a similar trend for the Exponential distribution.

REFERENCES

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* 24 (2011).

[2] Kumar Abhishek, Shweta Jain, and Sujit Gujar. 2020. Designing Truthful Contextual Multi-Armed Bandits based Sponsored Search Auctions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1732–1734.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Anthropic. 2025. Anthropic Claude Models. https://www.anthropic.com.

[5] Anthropic. 2025. Claude Opus 4 & Claude Sonnet 4 System Card. https://www.anthropic.com/claude-4-system-card

[6] Peter Auer. 2003. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* 3, null (March 2003), 397–422.

[7] Moshe Babaioff, Robert D Kleinberg, and Aleksandrs Slivkins. 2015. Truthful mechanisms with implicit payment computation. *Journal of the ACM (JACM)* 62, 2 (2015), 1–37.

[8] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. 2009. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*. 79–88.

[9] Satyanath Bhat, Shweta Jain, Sujit Gujar, and Y. Narahari. 2019. An optimal bidimensional multi-armed bandit auction for multi-unit procurement. *Annals of Mathematics and Artificial Intelligence* 85, 1 (2019), 1–19. https://doi.org/10.1007/s10472-018-9611-0

[10] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.

[11] Matthew Cary, Abraham D Flaxman, Jason D Hartline, and Anna R Karlin. 2008. Auctions for structured procurement.. In *SODA*, Vol. 8. 304–313.

[12] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*. PMLR, 208–214.

[13] Xiangxiang Dai, Jin Li, Xutong Liu, Anqi Yu, and John Lui. 2024. Cost-effective online multi-llm selection with versatile reward models. *arXiv preprint arXiv:2405.16587* (2024).

[14] Sankarshan Damle and Boi Faltings. 2025. LLMs for Resource Allocation: A Participatory Budgeting Approach to Inferring Preferences. In *ECAI*. 3743–3750.

[15] Nikhil R. Devanur and Sham M. Kakade. 2009. The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce* (Stanford, California, USA) *(EC '09)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/1566374.1566388

[16] Tao Feng, Yanzhen Shen, and Jiaxuan You. 2024. Graphrouter: A graph-based router for llm selections. *arXiv preprint arXiv:2410.03834* (2024).

[17] Guoju Gao, He Huang, Mingjun Xiao, Jie Wu, Yu-E Sun, and Sheng Zhang. 2021. Auction-based combinatorial multi-armed bandit mechanisms with strategic arms. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[18] Mengzhuo Guo, Wuqi Zhang, Congde Yuan, Binfeng Jia, Guoqing Song, Hua Hua, Shuangyang Wang, and Qingpeng Zhang. 2024. A Bayesian Multi-Armed Bandit Algorithm for Bid Shading in Online Display Advertising. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4506–4513.

[19] Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031* (2024).

[20] Garud Iyengar and Anuj Kumar. 2008. Optimal procurement mechanisms for divisible goods with capacitated suppliers. *Review of Economic Design* 12, 2 (01 Jun 2008), 129–154. https://doi.org/10.1007/s10058-008-0046-7

[21] Shweta Jain, Ganesh Ghalme, Satyanath Bhat, Sujit Gujar, and Y. Narahari. 2016. A Deterministic MAB Mechanism for Crowdsourcing with Logarithmic Regret and Immediate Payments. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 86–94.

[22] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Y. Narahari. 2018. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence* 254 (2018), 44–63. https://doi.org/10.1016/j.artint.2017.10.001

[23] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *NeurIPS* 36 (2023), 51991–52008.

[24] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.

[25] Tyler Lu, David Pal, and Martin Pal. 2010. Contextual Multi-Armed Bandits. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 9)*, Yee Whye Teh and Mike Titterington (Eds.). PMLR, Chia Laguna Resort, Sardinia, Italy, 485–492. https://proceedings.mlr.press/v9/lu10a.html

[26] Eric Maskin and John Riley. 1984. Monopoly with incomplete information. *The RAND Journal of Economics* 15, 2 (1984), 171–196.

[27] Microsoft. 2023. Introducing Microsoft 365 Copilot – your copilot for work. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

[28] Roger B Myerson. 1981. Optimal auction design. *Mathematics of operations research* 6, 1 (1981), 58–73.

[29] Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory*. Cambridge University Press, Cambridge.

[30] OpenAI. 2024. Memory and New Controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/.

[31] OpenAI. 2025. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf.

[32] OpenAI. 2025. OpenAI Language Models. https://openai.com.

[33] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[34] Pronoy Patra, Sankarshan Damle, Manisha Padala, and Sujit Gujar. 2026. Truthful Reverse Auctions for Adaptive Selection via Contextual Multi-Armed Bandits. arXiv:2602.14476 [cs.GT] https://arxiv.org/abs/2602.14476

[35] Yury Pinsky. 2023. Google Bard: New Features Update (September 2023). https://blog.google/products/gemini/google-bard-new-features-update-sept-2023/

[36] Manhin Poon, XiangXiang Dai, Xutong Liu, Fang Kong, John Lui, and Jinhang Zuo. 2025. Online Multi-LLM Selection via Contextual Bandits under Unstructured Context Evolution. *arXiv preprint arXiv:2506.17670* (2025).

[37] Herbert Robbins. 1952. Some aspects of the sequential design of experiments. (1952).

[38] Jain Shweta and Gujar Sujit. 2020. A multiarmed bandit based incentive mechanism for a subset selection of customers for demand response in smart grids. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2046–2053.

[39] Yaron Singer. 2010. Budget feasible mechanisms. In *2010 IEEE 51st Annual Symposium on foundations of computer science*. IEEE, 765–774.

[40] Varul Srivastava, Sankarshan Damle, and Manisha Padala. 2025. Robust In-Context Learning via Multi-Armed Bandit-Based Partition Selection. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*. https://openreview.net/forum?id=NzERpVgzsC

[41] Yu Su, Diyi Yang, Shunyu Yao, and Tao Yu. 2024. Language Agents: Foundations, Prospects, and Risks. In *EMNLP: Tutorial Abstracts*. 17–24.

[42] Jiahang Sun, Zhiyong Wang, Runhan Yang, Chenjun Xiao, John C.S. Lui, and Zhongxiang Dai. 2025. Large Language Model-Enhanced Multi-Armed Bandits. In *Workshop on Reasoning and Planning for Large Language Models*. https://openreview.net/forum?id=aUpWZ3B734

[43] Jiayi Yuan, Yifan Lu, Rixin Liu, Yu-Neng Chuang, Hongyi Liu, Shaochen Zhong, Yang Sui, Guanchu Wang, Jiarong Xing, and Xia Hu. 2025. Who Routes the Router: Rethinking the Evaluation of LLM Routing Systems. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*. https://openreview.net/forum?id=EEPostHMtF

[44] Pengyu Zhao, Zijian Jin, and Ning Cheng. 2023. *arXiv preprint arXiv:2309.14365* (2023).