LLM-PriorCB: Textual Contextual Bandits with LLM-Induced Priors

Geon-Hyeong Kim[1*], **Yu Jin Kim**[1*], **June Yong Yang**[1], **Woohyung Lim**[1], **Youngsoo Jang**[2†], **Moontae Lee**[1†]
[1]LG AI Research, [2]UNIST

Abstract

Many real-world decision-making problems, including recommendation, tool selection, and model routing, require choosing among actions described in natural language while observing feedback only for the selected action. Contextual bandits provide a principled framework for exploration under such partial feedback, but standard implementations typically start from uninformative reward-model priors and can adapt slowly in cold-start or evolving-action regimes. Large language models (LLMs) offer useful semantic knowledge from reward-free text, yet using an LLM directly as the online decision maker does not by itself provide calibrated uncertainty or principled exploration. We study textual contextual bandits, where both contexts and actions are represented by text, action descriptions are available without reward labels, and the action set may evolve over time. We propose LLM-PriorCB, a two-stage framework that separates semantic prior construction from online learning. In the offline stage, an LLM estimates rewards for reward-free context-action text pairs, and these estimates are distilled into action-specific prior parameters. In the online stage, LLM-PriorCB applies a disjoint linear UCB rule initialized by these action-specific priors, and updates only the selected action's parameter using observed rewards, without further LLM queries. We derive prior-dependent regret guarantees for the resulting disjoint linear bandit, showing how prior misspecification affects the confidence radius while preserving no-regret learning. Experiments on MovieLens and GAIA show that LLM-PriorCB reduces cumulative regret relative to LinUCB, CBLI, and prompted LLM policy baselines.

1 Introduction

Many real-world applications make decisions from natural-language inputs by selecting among alternatives also described in natural language, including recommender systems [21, 14, 26], LLM routers [18, 7], and tool-use or agentic systems [22, 4, 5]. These settings naturally involve bandit feedback, since only the outcome of the selected action is observed, and often feature evolving action sets as new items, models, APIs, or tools are introduced [15, 28]. Thus, effective decision making requires both exploration under partial feedback and rapid generalization to new actions.

LLMs are attractive in this setting because they encode broad linguistic, factual, and commonsense knowledge from large-scale pretraining [9, 20, 6], and have been used for recommendation, routing, and tool selection [21, 22, 18, 7]. However, directly using an LLM as the decision maker does not solve online learning [19]. Standard LLM pipelines lack decision-relevant posterior uncertainty, principled exploration from partial feedback, and regret guarantees. Moreover, empirical studies show that native in-context exploration by LLM agents is brittle and benefits from explicit algorithmic

[*] Equal contribution. [†] Equal advising.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

support [12, 17]. Semantic prior knowledge alone is therefore insufficient for reliable decision making under partial feedback.

Contextual bandits provide principled exploration strategies and theoretical guarantees for partial-feedback learning [15, 8, 1, 2, 13], with recent extensions to neural function approximation and large action spaces [27, 28]. While modern text encoders can provide useful semantic representations for contexts and actions, existing bandit methods still typically initialize reward models from uninformative priors and adapt them only through sparse online feedback. As a result, even with pretrained embeddings, the learner may require substantial interaction to identify reward-relevant structure, particularly in cold-start settings and environments with evolving action spaces.

Prior work shows that warm-start signals can reduce early regret, but must be robust to mismatch with online rewards [25]. Recent evidence also suggests that LLM-generated information can provide useful prior knowledge for contextual bandits [3]. This raises our key question: can we use the semantic prior knowledge of LLMs while preserving the uncertainty-aware online learning behavior of contextual bandits?

We formalize *textual contextual bandits*, where both contexts and actions are natural-language descriptions and the available action set may evolve over time. At each round, the learner observes a textual context and a set of text-described actions, selects one action, and observes only its reward. The goal is to use textual semantics to improve early decisions while allowing online feedback to correct wrong prior beliefs.

We propose `LLM-PriorCB`, a framework that integrates LLM-induced priors with online contextual bandit algorithms. The LLM constructs prior beliefs over context-action pairs from their textual descriptions, while posterior updates rely on observed rewards. Crucially, the LLM does not directly choose actions. Instead, it provides an informative initialization, while the bandit algorithm governs online action selection and exploration. When the prior is informative, this initialization improves sample efficiency and reduces early-stage regret. When the prior is misspecified, online reward feedback can progressively correct the induced beliefs.

Our contributions are:

- We introduce *textual contextual bandits*, covering textual contexts, textual actions, partial feedback, and evolving action sets.
- We propose `LLM-PriorCB`, which converts LLM-induced semantic information into priors for online contextual bandit learning.
- We prove the asymptotic optimality of `LLM-PriorCB`, even under prior misspecification.
- We empirically show that `LLM-PriorCB` achieves stable and robust performance.

2 Preliminaries

This section introduces the contextual bandit notation used throughout the paper and recalls the feature-based reward models on which our method builds.

Contextual bandits. We consider a contextual bandit over horizon T. At each round $t \in [T]$, the learner observes a context c_t and a finite available action set $\mathcal{A}_t$, selects $a_t \in \mathcal{A}_t$, and observes only the reward of the selected action. For each action a, let $Y_t(a)$ denote its potential reward and define

$$\mu(c_t, a) := \mathbb{E}[Y_t(a) \mid c_t], \qquad y_t = Y_t(a_t) = \mu(c_t, a_t) + \eta_t,$$

where η_t is conditionally mean-zero noise. The optimal action, optimal expected reward, instantaneous regret, and cumulative regret are defined as

$$a_t^* \in \arg\max_{a \in \mathcal{A}_t} \mu(c_t, a), \qquad \mu_t^* = \mu(c_t, a_t^*), \qquad r_t = \mu_t^* - \mu(c_t, a_t), \qquad R_T = \sum_{t=1}^{T} r_t.$$

A policy selects actions using the past interaction history and the current context-action set.

Feature-based models. Many contextual bandit algorithms represent each context-action pair by a feature vector $z_{t,a} \in \mathbb{R}^D$. Linear methods assume

$$\mu(c_t, a) = \mu(z_{t,a}) = z_{t,a}^\top \theta^*,$$

where $\theta^* \in \mathbb{R}^D$ is an unknown parameter vector. They then use confidence sets or posterior sampling for exploration, as in LinUCB and Thompson sampling [15, 2]. For our purpose, the key limitation is the lack of informative reward-aware priors over the bandit parameters. Even with strong pretrained embeddings, contextual bandit algorithms typically initialize reward models from uninformative priors and must adapt them using sparse online feedback.

3 Textual Contextual Bandits

We define a textual contextual bandit as a contextual bandit in which both contexts and actions are described by natural language. At round t, the learner observes a textual context $c_t \in \mathcal{C}$ and an available action set $\mathcal{A}_t$, where each action $a \in \mathcal{A}_t$ has a textual description $s_a \in \mathcal{S}$. Let

$$\psi : \mathcal{C} \cup \mathcal{S} \to \mathbb{R}^d$$

denote a text embedding model that maps textual contexts and action descriptions into vector representations, and let

$$\Psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^D$$

denote a joint feature map that combines context and action representations. The textual information is embedded and combined as

$$z_t = \psi(c_t), \qquad z_a = \psi(s_a), \qquad z_{t,a} = \Psi(z_t, z_a)$$

and the expected reward is written as $\mu(z_{t,a}) = \mu(c_t, a)$. Accordingly,

$$a_t^* \in \arg\max_{a \in \mathcal{A}_t} \mu(z_{t,a}), \qquad \mu_t^* = \mu(z_{t,a_t^*}), \qquad r_t = \mu_t^* - \mu(z_{t,a_t}), \qquad R_T = \sum_{t=1}^{T} r_t.$$

For analysis, we consider either a linear model

$$\mu(z_{t,a}) = z_{t,a}^\top \theta^*$$

or a disjoint arm-specific model

$$\mu(z_{t,a}) = z_{t,a}^\top \theta_a^*.$$

In the disjoint setting, $\theta_{0,a}$ denotes the prior parameter for arm a, and $\hat{\theta}_{t,a}$ denotes its online estimate. LLM-PriorCB uses the LLM as a source of prior information rather than as the final decision maker. Specifically, the LLM uses textual descriptions to construct representations or prior parameters, while the bandit algorithm selects actions and updates estimates from observed rewards. This separation allows the learner to exploit semantic knowledge before sufficient feedback is available while preserving uncertainty-aware exploration and regret analysis.

Before online interaction, the learner may access a reward-free textual dataset

$$\mathcal{D}_0 = (\mathcal{C}_0, \mathcal{A}_0, \{s_a\}_{a \in \mathcal{A}_0})$$

which contains textual contexts and action descriptions but no logged reward tuples (c, a, y). In our framework, $\mathcal{D}_0$ is used only to extract LLM-induced prior information before deployment.

4 LLM-PriorCB

We propose `LLM-PriorCB`, a two-stage framework for textual contextual bandits. In the offline stage, an LLM is used to construct action-specific prior parameters from the reward-free textual dataset $\mathcal{D}_0$ defined in Section 3. In the online stage, a prior-centered UCB algorithm selects actions and updates its estimates using observed rewards. After online interaction begins, the LLM is no longer queried. Thus, the LLM provides only prior information, while exploration and adaptation are handled by the bandit algorithm.

4.1 Offline Construction of LLM-Induced Priors

We focus on the disjoint linear model from Section 3, where each action a has an unknown parameter θ_a^* and reward model $\mu(z_{t,a}) = z_{t,a}^\top \theta_a^*$. The goal of the offline stage is to construct an informative prior parameter $\theta_{0,a} \in \mathbb{R}^D$ for each action a.

For each offline context $c_i \in \mathcal{C}_0$ and action $a \in \mathcal{A}_0$, we query an LLM $\mathcal{M}$ with the textual pair (c_i, s_a) to obtain an estimated expected reward. When using M samples, we average the LLM outputs as

$$\widehat{\mu}_{i,a} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mu}_{i,a}^{(m)}.$$

Here, $\widehat{\mu}_{i,a}^{(m)}$ denotes the m-th LLM-generated estimate for (c_i, s_a).

We introduce a prior encoder $f_\phi : \mathbb{R}^d \to \mathbb{R}^D$ that maps the action embedding $z_a = \psi(s_a)$ to the prior parameter $\theta_{0,a} = f_\phi(z_a)$. The encoder is trained to make the induced linear predictor match the LLM-generated reward estimates:

$$\min_\phi \frac{1}{|\mathcal{C}_0|\,|\mathcal{A}_0|} \sum_{c_i \in \mathcal{C}_0} \sum_{a \in \mathcal{A}_0} \left(\widehat{\mu}_{i,a} - z_{i,a}^\top f_\phi(z_a)\right)^2. \tag{1}$$

After training, we set $\theta_{0,a} = f_\phi(z_a)$. For a newly introduced online action $a \notin \mathcal{A}_0$, we initialize its prior in the same way from its textual description s_a.

4.2 Online Learning with Prior-Centered UCB

Given the LLM-induced priors $\{\theta_{0,a}\}$, online learning proceeds with an action-wise prior-centered UCB rule. For each action a, initialize

$$A_{0,a} = \lambda I_D, \qquad b_{0,a} = \lambda \theta_{0,a}.$$

At round t, define the current estimate

$$\hat{\theta}_{t-1,a} = A_{t-1,a}^{-1} b_{t-1,a}.$$

For each available action $a \in \mathcal{A}_t$, compute

$$U_t(a) = z_{t,a}^\top \hat{\theta}_{t-1,a} + \beta_{t-1}^{\text{dis}}(\delta) \sqrt{z_{t,a}^\top A_{t-1,a}^{-1} z_{t,a}},$$

where $\beta_{t-1}^{\text{dis}}(\delta)$ is an exploration coefficient controlling the uncertainty bonus. The learner selects

$$a_t \in \arg\max_{a \in \mathcal{A}_t} U_t(a).$$

After observing y_t, only the selected action is updated:

$$A_{t,a_t} = A_{t-1,a_t} + z_{t,a_t} z_{t,a_t}^\top, \qquad b_{t,a_t} = b_{t-1,a_t} + y_t z_{t,a_t}.$$

For all $a \neq a_t$, set $A_{t,a} = A_{t-1,a}$ and $b_{t,a} = b_{t-1,a}$.

5 Theoretical Derivation

Theorem 5.1 (Cumulative regret analysis)**.** *Suppose Assumptions A.1, A.2, A.3, A.4, and A.5 hold. Suppose that at every round $t = 1, \ldots, T$, the selected arm a_t satisfies*

$$a_t \in \arg\max_{a \in \mathcal{A}_t} \left[z_{t,a}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) \sqrt{z_{t,a}^\top A_{t-1}^{-1} z_{t,a}} \right].$$

Then, with probability at least $1 - \delta$,

$$R_T := \sum_{t=1}^{T} r_t \leq 2\beta_T(\delta) \sqrt{2T c_\lambda \log \frac{\det(A_T)}{\det(\lambda I_D)}},$$

where

$$\beta_t(\delta) := R\sqrt{2\log\frac{\det(A_t)^{1/2}}{\det(\lambda I_D)^{1/2}\delta}} + \sqrt{\lambda}S_\Delta$$

and

$$c_\lambda := \max\left\{1, \frac{L^2}{\lambda}\right\}.$$

The above theorem can be specialized to LLM-PriorCB as follows.

Corollary 5.2 (Cumulative regret analysis for LLM-PriorCB)**.** *Suppose the assumptions in Theorem 5.1 hold. Consider the disjoint linear model*

$$\mu(z_{t,a}) = z_{t,a}^\top \theta_a^*, \qquad z_{t,a} \in \mathbb{R}^D, \quad a \in [K],$$

with arm-specific prior parameters $\theta_{0,a} \in \mathbb{R}^D$. Then the arm-wise LLM-PriorCB rule induced by this block representation satisfies, with probability at least $1-\delta$,

$$R_T \le 2\beta_T^{\mathrm{dis}}(\delta)\sqrt{2Tc_\lambda \log\frac{\prod_{a=1}^K \det(A_{T,a})}{\lambda^{Kd}}},$$

where

$$A_{t,a} := \lambda I_d + \sum_{i=1}^t \mathbf{1}\{a_i = a\} z_{i,a} z_{i,a}^\top$$

and

$$\beta_t^{\mathrm{dis}}(\delta) := R\sqrt{2\log\frac{\prod_{a=1}^K \det(A_{t,a})^{1/2}}{\lambda^{Kd/2}\delta}} + \sqrt{\lambda}S_\Delta.$$

6 Experiments

We evaluate our method on two benchmark tasks constructed from the MovieLens dataset [10] and the GAIA dataset [16].

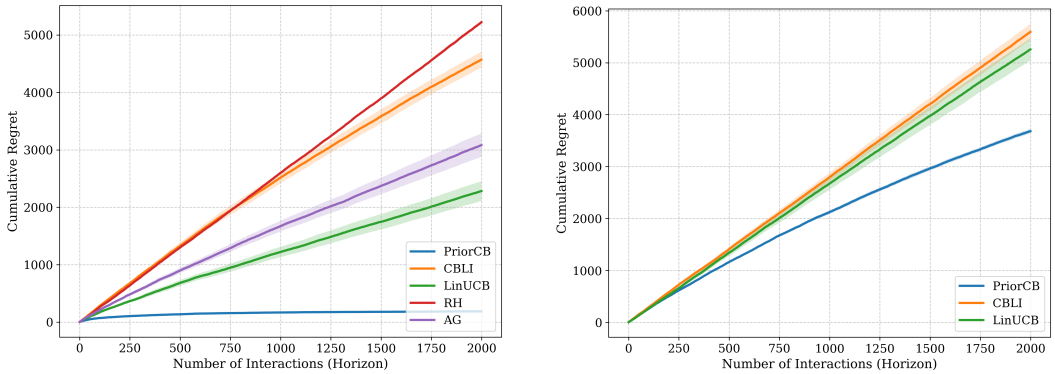
6.1 MovieLens

We adopt a semi-synthetic contextual bandit task based on the MovieLens dataset [10], following an experimental design introduced in prior work, most notably EVOLvE [17]. This semi-synthetic setup provides controlled access to key components of the environment, such as oracle context features and ground-truth rewards, while preserving realistic user and movie information derived from real-world data.

Dataset We use the MovieLens-100k dataset [10], a widely adopted benchmark for recommendation system evaluation. It consists of 100,000 explicit ratings on a 5-point scale, collected from $N = 943$ users on 1,682 movies. Each user has provided at least 20 ratings and is associated with demographic attributes, including age, gender, occupation, and zip code. Each movie is described by metadata such as its title, release year, and genre tags. Together, these features provide a natural testbed for studying contextual decision-making with textual side information. Following prior work, we restrict the action set $\mathcal{A}$ to the top-K most-watched movies, where $K = 10$ corresponds to a relatively easy setting and $K = 30$ induces a more challenging problem.

Environment To construct the ground-truth reward function, we apply singular value decomposition (SVD) to the user–movie rating matrix $P \in \mathbb{R}^{N\times K}$ and obtain a low-rank approximation [11]. Specifically, we approximate P as $\tilde{P} = U\Sigma V^\top$, where U and V denote user and movie embedding matrices, respectively. The reward for recommending movie a to user i is computed as $\mu(z_{i,a}) = u_i^\top \Sigma v_a$.

At each time step, a user is sampled, and the agent receives contextual information according to the current information regime. The agent selects one movie from a finite action set and receives a scalar reward corresponding to the user's preference.



(a) Oracle Regime

(b) Text-Only Regime

Figure 1: Performance as a function of the interaction horizon for $K = 10$. PriorCB (ours) consistently achieves significantly lower cumulative regret compared to baseline methods, demonstrating strong sample efficiency in the oracle regime. We observe a similar qualitative trend in the text-only regime, indicating that the advantage of PriorCB persists even when only textual information is provided. Shaded regions indicate $\pm$ one standard error over 10 independent runs.

Information Regimes We consider two information regimes. *(i) Oracle Context Regime*: the agent has direct access to latent user context features u_i obtained from the low-rank approximation, which are also used to generate rewards. *(ii) Text-Only Context Regime*: the latent user context features are hidden, and the agent must infer effective context features solely from textual descriptions of user demographic information. The Oracle Context Regime closely resembles the classical contextual bandit setting, whereas the Text-Only Context Regime more closely reflects practical real-world scenarios.

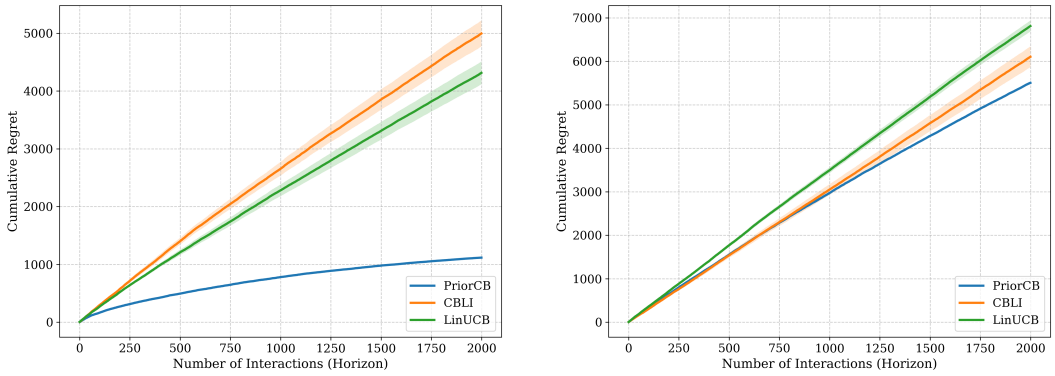
Baselines and Experimental Setup In the Oracle Context Regime, we compare our method against CBLI [3], standard LinUCB [15], and two prompted LLM policy baselines, RH and AG, both of which prompt an LLM to directly select actions [17]. RH represents past interactions as raw $(\text{context}, \text{action}, \text{reward})$ tuples and provides them, together with the current context and available actions, as the LLM prompt. AG augments the prompt with LinUCB-derived exploitation values and exploration bonuses, so that the LLM receives explicit algorithmic guidance for action selection. In the Text-Only Context Regime, we focus on baselines whose action-selection rules are fully specified under the same text-derived feature representation. We therefore report CBLI [3] and LinUCB [15]. We omit RH and AG because they are prompted LLM policy baselines requiring LLM inference at every round, and adapting AG's oracle-context LinUCB guidance to text-only features would introduce an additional implementation variant beyond the original baseline.

For prior construction, we use QWEN3-4B-INSTRUCT-2507 [24] as the LLM $\mathcal{M}$ and QWEN3-EMBEDDING-8B as the text encoder ψ. We select a subset of 100 users from the dataset as $\mathcal{D}_0$. For each user i and movie a in $\mathcal{D}_0$, we estimate the corresponding rating by prompting the LLM $\mathcal{M}$ with the user's demographic attributes c_i and the movie's metadata s_a. Each user–movie pair is queried 10 times, and the resulting predictions are averaged to obtain prior reward estimates. These estimated ratings are used both for prior construction in our algorithm and for CBLI pretraining.

For the action encoder, the mapping function f_ϕ is implemented as a two-layer multilayer perceptron (MLP). We train f_ϕ by minimizing the MSE objective defined in Equation 1. During online learning, users are sampled from a dataset disjoint from $\mathcal{D}_0$. All online experiments are conducted over 10 random seeds, and performance is evaluated using cumulative regret. Details of the prompting templates and training are provided in Appendix D.1.

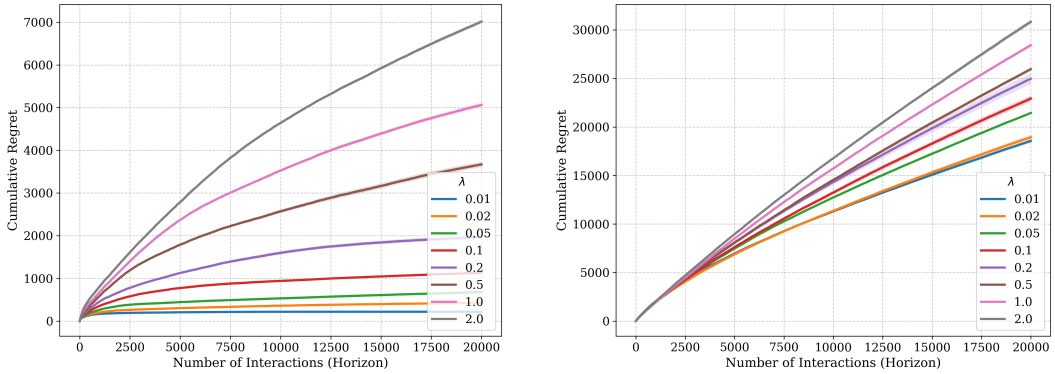
6.2 Experimental results

Experiments with $K = 10$ We first evaluate all methods in an oracle regime, where the agent is provided with oracle-quality action representations. This setting isolates the effect of the learning al-



(a) Oracle Regime (b) Text-Only Regime

Figure 2: Performance as a function of the interaction horizon when $K = 10$ in prior construction and $K = 30$ when online learning. Shaded regions indicate $\pm$ one standard error over 10 independent runs.



(a) Oracle Regime (b) Text-Only Regime

Figure 3: We report cumulative regret of LLM-PriorCB for $\lambda \in [0.01, 2.0]$ under (a) the oracle regime and (b) the text-only regime. Smaller λ consistently yields lower regret, while larger λ degrades performance, indicating that overly confident priors can suppress exploration and hinder effective online adaptation.

gorithm itself by removing confounding factors arising from noisy or mis-specified textual encodings of the context. As shown in Figure 1(a), LLM-PriorCB (abbreviated as PriorCB in figures) achieves significantly lower cumulative regret than all baseline methods across the entire horizon, demonstrating strong sample efficiency. This result indicates that LLM-PriorCB can effectively exploit informative representations while maintaining sufficient exploration. In contrast, LinUCB performs poorly in this setting, as it learns action representations entirely from scratch and therefore requires a substantial amount of exploration. CBLI also exhibits limited exploration due to its inherent design, leading to suboptimal performance even under oracle access. As a result, its convergence is slower than that of the naive LinUCB. Furthermore, RH and AG fail to achieve competitive performance, suggesting that relying solely on LLM-based decision making without principled online learning and exploration mechanisms is insufficient.

We then evaluate the methods in a text-only regime (Figure 1(b)), where context and action representations are derived solely from natural language descriptions. This setting is closer to real-world problem scenarios, and despite the additional noise and ambiguity introduced by textual representations, LLM-PriorCB maintains a clear advantage over competing methods, indicating that its benefits are not limited to idealized oracle settings. In Appendix B, we report additional results over an extended horizon of up to 20,000 steps. Together, these results demonstrate that LLM-PriorCB

effectively leverages prior information both when representations are reliable and when they must be inferred from noisy textual descriptions.

Experiments with dynamic action set Furthermore, we evaluate the methods in an arm-evolving setting, where the available action set changes over time. Specifically, the offline textual dataset $\mathcal{D}_0$ contains prior information for only $K = 10$ actions, while online learning is conducted with an expanded action set of $K = 30$ actions. Thus, many actions appearing during online interaction are not included in the offline prior construction stage. As shown in Figure 2, LLM-PriorCB consistently outperforms baseline methods despite this mismatch between the offline prior set and the online action set. While this experiment represents only a limited evaluation of evolving-action scenarios, the results suggest that LLM-PriorCB can effectively accommodate newly introduced actions by leveraging textual prior information.

Ablation studies We further conduct an ablation study on the exploration parameter λ, which controls the strength of the prior in LLM-PriorCB As shown in Figure 3 smaller values of λ consistently lead to lower cumulative regret in this experiments. This observation suggests that overly trusting the prior can hinder effective exploration, preventing the agent from adequately incorporating online feedback. This behavior is consistent with the poor performance of CBLI observed in earlier experiments, where limited exploration due to strong reliance on prior estimates leads to suboptimal outcomes. Additional ablation studies on the feature dimension, rank, and the number of actions are provided in Appendix B.

6.3 GAIA

Dataset We conduct experiments on the GAIA [16] benchmark, a dataset designed to evaluate General AI Assistants on realistic user queries that require a combination of fundamental abilities. The questions are formulated to have a brief, unambiguous correct answer, making verification straightforward. We construct a text-only subset of GAIA by filtering out instances that require visual understanding. After filtering, the subset contains 127 questions from validation set and 230 questions from test set.

Environment We formulate tool selection as a contextual bandit with a set of $K = 3$ tools. The three arms correspond to reason-only (using only internal knowledge without any web search; low cost), search-light (using single web search with one query; mid cost), and search-deep (using three web search with three distinct queries; high cost). Detailed descriptions are provided in the Appendix D.2.2. At each time step, the agent receives a question, and the agent selects one tool from a set of K tools, after which it receives a scalar reward defined by a pre-designed function. Specifically, we set the expected reward for tool a to question c_i as $\mu(z_{t,a}) = accuracy - cost$, where $accuracy \in \{0, 1\}$ is computed via quasi-exact match against the ground-truth answer, and the $cost$ is a fixed tool-dependent penalty.

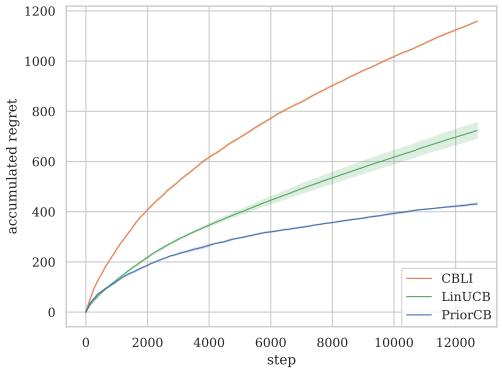
Baelines We consider two baselines: naive LinUCB [15] with froze embedding backbone, CBLI [3].

Experimental setup We evaluate gpt-5-mini [23] and gpt-5.2 on GAIA benchmark. These models are used as prior LLMs for offline prior construction, while online learning and evaluation are performed on the validation set under a contextual bandit.

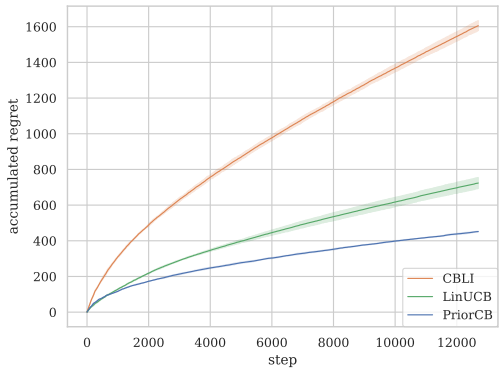
For prior construction, we use the GAIA benchmark test set as $\mathcal{D}_0$. For each question c_i and tool a, we obtain an estimated score by prompting an LLM with the question together with the descriptions of all available tools. The full prompt template is provided in Appendix D.2.2. Subsequently, we train a two-layer MLP action encoder on the generated dataset, using QWEN/QWEN3-EMBEDDING-8B as a frozen embedding backbone. The experimental details are explained in Appendix D.2.3.

For online learning, we use the GAIA benchmark validation set. Due to the limited number of validation instances, we repeat the online evaluation for 100 iterations (with reshuffling and different random seeds).

Experimental results Figure 4 reports the cumulative regret of PriorCB, LinUCB, and CBLI in the main GAIA experiment. Overall, PriorCB achieves substantially lower cumulative regret than



(a) Cumulative regret of Prior from different eval contexts



(b) Cumulative regret of Prior from same eval contexts

Figure 4: Performance as a function of the interaction horizon when $K = 3$ using gpt-5.2 with medium reasoning effort. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels.

the baselines, indicating that the LLM-induced prior provides useful initialization for subsequent online learning. Figure 4(a) and Figure 4(b) correspond to priors constructed from a different context set and from the evaluation context set, respectively. In both cases, prior construction uses only reward-free textual information and does not access environment rewards, labels, or answers. CBLI shows substantial variation depending on the dataset used for prior construction. By contrast, PriorCB remains stable across the two prior-construction settings, while LinUCB is also stable but lacks the benefit of LLM-induced initialization. Additional results for the prompted LLM policy baseline RH [17] are provided in Appendix C.1. Since RH requires LLM inference at every decision step, we report it separately from the main long-horizon comparison. Appendix C.1 also includes RH results with GPT-5.2 and GPT-5-mini, together with the long-horizon GPT-5-mini results for LinUCB, CBLI, and PriorCB. The average reward results are reported in Appendix C.2.

7 Conclusion

We studied *textual contextual bandits*, where both contexts and actions are described in natural language and the action set may evolve over time. To bridge LLM-based decision making and principled bandit learning, we proposed LLM-PriorCB, a two-stage framework that uses an LLM only to construct reward-free offline priors from textual information, and then performs online learning solely through a standard contextual bandit algorithm. Theoretically, we establish prior-dependent cumulative regret guarantees for disjoint linear contextual bandits, showing that well-aligned offline priors can reduce the regret bound, while prior misspecification enters only through explicit arm-wise bias terms. Across experiments on MovieLens and GAIA, LLM-PriorCB consistently achieved lower regret than baselines under both oracle and text-only regimes, degraded gracefully under mis-specified priors, and showed promising performance when the action set expanded between prior construction and online learning. Our results highlight the importance of separating prior knowledge extraction from online adaptation, enabling stable exploration and reliable improvement from sparse bandit feedback.

There are several directions for future work. First, while our evolving-arm experiment provides initial evidence of robustness, broader evaluations under more diverse non-stationary action dynamics would further clarify the limits of the approach. Second, extending the prior construction stage to incorporate richer uncertainty estimates (e.g., calibrated prior covariances) and exploring tighter integrations with contextual bandit variants beyond linear models are promising avenues. Finally, applying the framework to real-world deployments with latency and cost constraints remains an important step toward practical LLM-assisted online decision-making systems.

References

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

[2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[3] Parand Alamdari, Yanshuai Cao, and Kevin Wilson. Jump starting bandits with llm-generated prior knowledge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19821–19833, 2024.

[4] Anthropic. Introducing the model context protocol. Anthropic News, 2024.

[5] Anthropic. Agent skills. Claude API Documentation, 2025.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024.

[8] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[10] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[11] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[12] Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

[13] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[14] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. Large language models for generative recommendation: A survey and visionary discussions. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 10146–10159, 2024.

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

[16] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

[17] Allen Nie, Yi Su, Bo Chang, Jonathan N Lee, Ed H Chi, Quoc V Le, and Minmin Chen. Evolve: Evaluating and optimizing llms for exploration. *arXiv preprint arXiv:2410.06238*, 2024.

[18] 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. *arXiv preprint arXiv:2406.18665*, 2024.

[19] Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do llm agents have regret? a case study in online learning and games. *arXiv preprint arXiv:2403.16843*, 2024.

[20] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019.

[21] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *Proceedings of the 17th ACM conference on recommender systems*, pages 890–896, 2023.

[22] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551, 2023.

[23] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka,

Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2025. URL `https://arxiv.org/abs/2601.03267`.

[24] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

[25] Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand N Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301*, 2019.

[26] Wenlin Zhang, Chuhan Wu, Xiangyang Li, Yuhao Wang, Kuicai Dong, Yichao Wang, Xinyi Dai, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. Llmtreerec: Unleashing the power of large language models for cold-start recommendations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 886–896, 2025.

[27] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International conference on machine learning*, pages 11492–11502. PMLR, 2020.

[28] Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pages 27428–27453. PMLR, 2022.

Algorithm 1 Prior-centered LinUCB

Require: Prior parameter $\theta_0 \in \mathbb{R}^D$, regularization $\lambda > 0$, confidence level $\delta \in (0, 1)$

Initialize $A_0 \leftarrow \lambda I_D$
Initialize $b_0 \leftarrow \lambda \theta_0$
for $t = 1, 2, \ldots, T$ **do**
  Observe context c_t and available arm set $\mathcal{A}_t$
  for all $a \in \mathcal{A}_t$ **do**
    Construct joint feature $z_{t,a} \leftarrow \Psi(\psi(c_t), \psi(s_a))$
  end for
  Compute $\hat{\theta}_{t-1} \leftarrow A_{t-1}^{-1} b_{t-1}$
  Select

$$a_t \leftarrow \arg\max_{a \in \mathcal{A}_t} \left[z_{t,a}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) \sqrt{z_{t,a}^\top A_{t-1}^{-1} z_{t,a}} \right]$$

  Observe reward y_t
  Update

$$A_t \leftarrow A_{t-1} + z_{t,a_t} z_{t,a_t}^\top$$

  Update

$$b_t \leftarrow b_{t-1} + y_t z_{t,a_t}$$

end for

A Theoretical Derivation

For the theoretical derivation, we analyze the sequence generated by an arbitrary adaptive policy. The proof below does not depend on the internal implementation of Algorithm 1, except when we explicitly impose the UCB selection rule for the instantaneous regret bound.

Assumption A.1 (Adaptive bandit protocol)**.** At each round t, before observing the reward y_t, the learner observes the context c_t and a nonempty finite arm set $\mathcal{A}_t$. For each $a \in \mathcal{A}_t$, the joint feature is

$$z_{t,a} = \Psi(\psi(c_t), \psi(s_a)) \in \mathbb{R}^D.$$

The learner selects $a_t \in \mathcal{A}_t$ before observing y_t.

Let $\mathcal{H}_t$ denote the pre-reward sigma-field at round t, containing all past observations and the current context, arm set, and selected arm:

$$\mathcal{H}_t = \sigma\big(\theta_0,\, c_1, \mathcal{A}_1, a_1, y_1,\, \ldots,\, c_{t-1}, \mathcal{A}_{t-1}, a_{t-1}, y_{t-1},\, c_t, \mathcal{A}_t, a_t\big).$$

Then $x_t := z_{t,a_t}$ is $\mathcal{H}_t$-measurable.

Assumption A.2 (Linear reward model)**.** There exists a fixed but unknown parameter

$$\theta^* \in \mathbb{R}^D$$

such that for every context-arm pair,

$$\mu(z_{t,a}) = z_{t,a}^\top \theta^*.$$

The optimal arm at round t is

$$a_t^* \in \arg\max_{a \in \mathcal{A}_t} \mu(z_{t,a}),$$

and the optimal expected reward is

$$\mu_t^* = \mu(z_{t,a_t^*}).$$

Assumption A.3 (Bounded features)**.** There exists $L > 0$ such that

$$\|z_{t,a}\|_2 \le L$$

for all t and all $a \in \mathcal{A}_t$. Equivalently,

$$\|z_{t,a}\|_{(\lambda I_D)^{-1}}^2 \le \frac{L^2}{\lambda}.$$

Algorithm 2 LLM-PriorCB

Require: Offline textual dataset $\mathcal{D}_0 = (\mathcal{C}_0, \mathcal{A}_0, \{s_a\}_{a\in\mathcal{A}_0})$, LLM $\mathcal{M}$, number of LLM samples M, embedding map ψ, joint feature map Ψ, prior encoder f_ϕ, regularization $\lambda > 0$, confidence level $\delta \in (0, 1)$, horizon T

Stage I: Offline prior construction

for all $c_i \in \mathcal{C}_0$ **do**
  $z_i \leftarrow \psi(c_i)$
  for all $a \in \mathcal{A}_0$ **do**
    $z_a \leftarrow \psi(s_a)$, $z_{i,a} \leftarrow \Psi(z_i, z_a)$
    Query $\mathcal{M}$ M times with (c_i, s_a) to obtain $\{\widehat{\mu}_{i,a}^{(m)}\}_{m=1}^M$
    $\widehat{\mu}_{i,a} \leftarrow \frac{1}{M}\sum_{m=1}^M \widehat{\mu}_{i,a}^{(m)}$
  end for
end for
Train f_ϕ by minimizing Eq. equation 1
for all $a \in \mathcal{A}_0$ **do**
  $\theta_{0,a} \leftarrow f_\phi(\psi(s_a))$
end for

Stage II: Online learning

Initialize the set of initialized arms $\mathcal{I} \leftarrow \emptyset$
for $t = 1, 2, \ldots, T$ **do**
  Observe context c_t and available arm set $\mathcal{A}_t$
  $z_t \leftarrow \psi(c_t)$
  for all $a \in \mathcal{A}_t$ **do**
    $z_a \leftarrow \psi(s_a)$, $z_{t,a} \leftarrow \Psi(z_t, z_a)$
    if $a \notin \mathcal{I}$ **then**
      $\theta_{0,a} \leftarrow f_\phi(z_a)$
      $A_a \leftarrow \lambda I_D$
      $b_a \leftarrow \lambda\theta_{0,a}$
      $\mathcal{I} \leftarrow \mathcal{I} \cup \{a\}$
    end if
    $\hat{\theta}_a \leftarrow A_a^{-1} b_a$
    $U_t(a) \leftarrow z_{t,a}^\top \hat{\theta}_a + \beta_{t-1}^{\text{dis}}(\delta)\sqrt{z_{t,a}^\top A_a^{-1} z_{t,a}}$
  end for
  Select $a_t \leftarrow \arg\max_{a\in\mathcal{A}_t} U_t(a)$
  Observe reward y_t
  Update $A_{a_t} \leftarrow A_{a_t} + z_{t,a_t} z_{t,a_t}^\top$
  Update $b_{a_t} \leftarrow b_{a_t} + y_t z_{t,a_t}$
end for

Assumption A.4 (Conditional sub-Gaussian noise)**.** The observed reward satisfies

$$y_t = x_t^\top \theta^* + \eta_t.$$

The noise satisfies

$$\mathbb{E}[\eta_t \mid \mathcal{H}_t] = 0,$$

and for all $s \in \mathbb{R}$,

$$\mathbb{E}[\exp(s\eta_t) \mid \mathcal{H}_t] \le \exp\left(\frac{s^2 R^2}{2}\right).$$

Assumption A.5 (Bounded prior misspecification)**.** There exists a known constant $S_\Delta > 0$ such that

$$\|\theta^* - \theta_0\|_2 \le S_\Delta.$$

We write

$$\Delta^* := \theta^* - \theta_0.$$

Definition A.6. For any adaptively chosen sequence $\{a_t\}_{t\ge1}$, define

$$A_t = \lambda I_D + \sum_{i=1}^t x_i x_i^\top, \qquad b_t = \lambda\theta_0 + \sum_{i=1}^t y_i x_i,$$

where $x_i = z_{i,a_i}$.

The prior-centered ridge estimator is

$$\hat{\theta}_t = A_t^{-1} b_t.$$

Since $\lambda > 0$,

$$A_t \succeq \lambda I_D \succ 0,$$

so A_t is symmetric positive definite and A_t^{-1} is well-defined.

Lemma A.7 (Equivalence to residual ridge regression)**.** *Define*

$$\hat{\Delta}_t := \hat{\theta}_t - \theta_0, \qquad \Delta^* := \theta^* - \theta_0.$$

Then

$$\hat{\Delta}_t = A_t^{-1} \sum_{i=1}^{t} x_i \left(y_i - x_i^\top \theta_0 \right).$$

Equivalently, $\hat{\Delta}_t$ is the ridge estimator obtained by regressing the residualized rewards

$$\tilde{y}_i := y_i - x_i^\top \theta_0$$

on x_i, with regularization centered at zero:

$$\hat{\Delta}_t = \arg\min_{\Delta \in \mathbb{R}^D} \left\{ \sum_{i=1}^{t} \left(\tilde{y}_i - x_i^\top \Delta \right)^2 + \lambda \|\Delta\|_2^2 \right\}.$$

Proof. By definition,

$$\hat{\Delta}_t = \hat{\theta}_t - \theta_0 = A_t^{-1} b_t - \theta_0 = A_t^{-1}(b_t - A_t \theta_0).$$

Now,

$$\begin{aligned} b_t - A_t \theta_0 &= \lambda \theta_0 + \sum_{i=1}^{t} y_i x_i - \left(\lambda I_D + \sum_{i=1}^{t} x_i x_i^\top \right) \theta_0 \\ &= \sum_{i=1}^{t} y_i x_i - \sum_{i=1}^{t} x_i x_i^\top \theta_0 \\ &= \sum_{i=1}^{t} x_i \left(y_i - x_i^\top \theta_0 \right). \end{aligned}$$

Therefore,

$$\hat{\Delta}_t = A_t^{-1} \sum_{i=1}^{t} x_i \left(y_i - x_i^\top \theta_0 \right).$$

The optimization form follows from the normal equation of ridge regression:

$$\left(\lambda I_D + \sum_{i=1}^{t} x_i x_i^\top \right) \Delta = \sum_{i=1}^{t} x_i \tilde{y}_i.$$

The left-hand matrix is A_t, so the unique minimizer is exactly

$$\hat{\Delta}_t = A_t^{-1} \sum_{i=1}^{t} x_i \tilde{y}_i.$$

□

Lemma A.8 (Self-normalized bound)**.** *Under Assumptions A.1 and A.4, with probability at least $1 - \delta$, simultaneously for all $t \geq 0$,*

$$\left\| \sum_{i=1}^{t} x_i \eta_i \right\|_{A_t^{-1}} \leq R \sqrt{2 \log \frac{\det(A_t)^{1/2}}{\det(\lambda I_D)^{1/2} \delta}}.$$

Proof. This is the standard self-normalized concentration inequality for vector-valued martingales. In our notation, x_t is measurable before observing y_t, and η_t is conditionally R-sub-Gaussian given the pre-reward sigma-field $\mathcal{H}_t$. Therefore, the process

$$\sum_{i=1}^{t} x_i \eta_i$$

satisfies the conditions of the self-normalized martingale inequality with regularization matrix λI_D. For more details, see Theorem 1 in [1]. □

Lemma A.9 (Confidence ellipsoid)**.** *Suppose Assumptions A.1, A.2, A.4, and A.5 hold. Define*

$$\beta_t(\delta) = R\sqrt{2\log\frac{\det(A_t)^{1/2}}{\det(\lambda I_D)^{1/2}\delta}} + \sqrt{\lambda} S_\Delta.$$

Then, with probability at least $1-\delta$*, simultaneously for all* $t \geq 0$*,*

$$\|\hat{\theta}_t - \theta^*\|_{A_t} \leq \beta_t(\delta).$$

Proof. By Lemma A.7,

$$\hat{\Delta}_t = A_t^{-1} \sum_{i=1}^{t} x_i \left(y_i - x_i^\top \theta_0\right).$$

Using the linear model,

$$y_i = x_i^\top \theta^* + \eta_i,$$

we have

$$y_i - x_i^\top \theta_0 = x_i^\top (\theta^* - \theta_0) + \eta_i = x_i^\top \Delta^* + \eta_i.$$

Therefore,

$$\begin{aligned}\hat{\Delta}_t &= A_t^{-1} \sum_{i=1}^{t} x_i \left(x_i^\top \Delta^* + \eta_i\right) \\ &= A_t^{-1} \left(\sum_{i=1}^{t} x_i x_i^\top \Delta^* + \sum_{i=1}^{t} x_i \eta_i\right).\end{aligned}$$

Since

$$A_t = \lambda I_D + \sum_{i=1}^{t} x_i x_i^\top,$$

we have

$$\sum_{i=1}^{t} x_i x_i^\top = A_t - \lambda I_D.$$

Thus,

$$\begin{aligned}\hat{\Delta}_t &= A_t^{-1} \left((A_t - \lambda I_D)\Delta^* + \sum_{i=1}^{t} x_i \eta_i\right) \\ &= \Delta^* - \lambda A_t^{-1} \Delta^* + A_t^{-1} \sum_{i=1}^{t} x_i \eta_i.\end{aligned}$$

Hence,

$$\hat{\Delta}_t - \Delta^* = A_t^{-1} \left(\sum_{i=1}^{t} x_i \eta_i - \lambda \Delta^*\right).$$

Since

$$\hat{\Delta}_t - \Delta^* = \hat{\theta}_t - \theta^*,$$

we obtain

$$\begin{aligned}\|\hat{\theta}_t - \theta^*\|_{A_t} &= \left\| A_t^{-1} \left(\sum_{i=1}^{t} x_i \eta_i - \lambda \Delta^* \right) \right\|_{A_t} \\ &= \left\| \sum_{i=1}^{t} x_i \eta_i - \lambda \Delta^* \right\|_{A_t^{-1}} \\ &\leq \left\| \sum_{i=1}^{t} x_i \eta_i \right\|_{A_t^{-1}} + \|\lambda \Delta^*\|_{A_t^{-1}}.\end{aligned}$$

For the second term, since $A_t \succeq \lambda I_D$, we have

$$A_t^{-1} \preceq \frac{1}{\lambda} I_D.$$

Therefore,

$$\begin{aligned}\|\lambda \Delta^*\|_{A_t^{-1}}^2 &= \lambda^2 {\Delta^*}^\top A_t^{-1} \Delta^* \\ &\leq \lambda^2 {\Delta^*}^\top \left(\frac{1}{\lambda} I_D \right) \Delta^* \\ &= \lambda \|\Delta^*\|_2^2 \\ &\leq \lambda S_\Delta^2.\end{aligned}$$

Thus,

$$\|\lambda \Delta^*\|_{A_t^{-1}} \leq \sqrt{\lambda} S_\Delta.$$

By Lemma A.8, with probability at least $1 - \delta$, simultaneously for all $t \geq 0$,

$$\left\| \sum_{i=1}^{t} x_i \eta_i \right\|_{A_t^{-1}} \leq R \sqrt{2 \log \frac{\det(A_t)^{1/2}}{\det(\lambda I_D)^{1/2} \delta}}.$$

Combining the two inequalities gives

$$\|\hat{\theta}_t - \theta^*\|_{A_t} \leq R \sqrt{2 \log \frac{\det(A_t)^{1/2}}{\det(\lambda I_D)^{1/2} \delta}} + \sqrt{\lambda} S_\Delta = \beta_t(\delta).$$

□

Definition A.10. For a fixed horizon T, define the confidence event

$$\mathcal{E}_T(\delta) := \left\{ \forall t \in \{0, \ldots, T\},\ \|\hat{\theta}_t - \theta^*\|_{A_t} \leq \beta_t(\delta) \right\}.$$

By Lemma A.9,

$$\Pr(\mathcal{E}_T(\delta)) \geq 1 - \delta.$$

Lemma A.11 (Instantaneous regret under UCB selection)**.** *On the event $\mathcal{E}_T(\delta)$, fix a round t. Suppose the selected arm a_t satisfies*

$$a_t \in \arg\max_{a \in \mathcal{A}_t} \left[z_{t,a}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) \sqrt{z_{t,a}^\top A_{t-1}^{-1} z_{t,a}} \right].$$

Then the instantaneous regret

$$r_t := \mu_t^* - \mu(z_{t,a_t})$$

satisfies

$$r_t \leq 2\beta_{t-1}(\delta) \sqrt{z_{t,a_t}^\top A_{t-1}^{-1} z_{t,a_t}}.$$

Consequently, with probability at least $1 - \delta$, the above bound holds simultaneously for all rounds t in which the UCB selection rule is used.

Proof. For any arm $a \in \mathcal{A}_t$, define

$$w_{t,a} := \sqrt{z_{t,a}^\top A_{t-1}^{-1} z_{t,a}}.$$

By Cauchy's inequality under the A_{t-1}-norm,

$$\begin{aligned}
\left| z_{t,a}^\top (\hat{\theta}_{t-1} - \theta^*) \right| &= \left| \left(A_{t-1}^{-1/2} z_{t,a} \right)^\top \left(A_{t-1}^{1/2} (\hat{\theta}_{t-1} - \theta^*) \right) \right| \\
&\leq \left\| A_{t-1}^{-1/2} z_{t,a} \right\|_2 \left\| A_{t-1}^{1/2} (\hat{\theta}_{t-1} - \theta^*) \right\|_2 \\
&= w_{t,a} \|\hat{\theta}_{t-1} - \theta^*\|_{A_{t-1}} \\
&\leq \beta_{t-1}(\delta) w_{t,a}.
\end{aligned}$$

Therefore, for every $a \in \mathcal{A}_t$, we obtain two upper-confidence inequalities:

$$z_{t,a}^\top \theta^* \leq z_{t,a}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) w_{t,a},$$

and

$$z_{t,a}^\top \hat{\theta}_{t-1} \leq z_{t,a}^\top \theta^* + \beta_{t-1}(\delta) w_{t,a}.$$

Let

$$U_t(a) := z_{t,a}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) w_{t,a}.$$

By the first upper-confidence inequality,

$$\mu(z_{t,a_t^*}) = z_{t,a_t^*}^\top \theta^* \leq U_t(a_t^*).$$

Since a_t maximizes the UCB score,

$$U_t(a_t^*) \leq U_t(a_t).$$

Thus,

$$\mu(z_{t,a_t^*}) \leq U_t(a_t).$$

Expanding $U_t(a_t)$ gives

$$\mu(z_{t,a_t^*}) \leq z_{t,a_t}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) w_{t,a_t}.$$

Using the second confidence inequality for the selected arm a_t,

$$z_{t,a_t}^\top \hat{\theta}_{t-1} \leq z_{t,a_t}^\top \theta^* + \beta_{t-1}(\delta) w_{t,a_t}.$$

Therefore,

$$\begin{aligned}
\mu(z_{t,a_t^*}) &\leq z_{t,a_t}^\top \theta^* + 2\beta_{t-1}(\delta) w_{t,a_t} \\
&= \mu(z_{t,a_t}) + 2\beta_{t-1}(\delta) w_{t,a_t}.
\end{aligned}$$

Rearranging,

$$r_t = \mu(z_{t,a_t^*}) - \mu(z_{t,a_t}) \leq 2\beta_{t-1}(\delta) w_{t,a_t}.$$

Substituting the definition of w_{t,a_t} yields

$$r_t \leq 2\beta_{t-1}(\delta) \sqrt{z_{t,a_t}^\top A_{t-1}^{-1} z_{t,a_t}},$$

□

Theorem 5.1 (Cumulative regret analysis)**.** *Suppose Assumptions A.1, A.2, A.3, A.4, and A.5 hold. Suppose that at every round $t = 1, \ldots, T$, the selected arm a_t satisfies*

$$a_t \in \arg\max_{a \in \mathcal{A}_t} \left[z_{t,a}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) \sqrt{z_{t,a}^\top A_{t-1}^{-1} z_{t,a}} \right].$$

Then, with probability at least $1 - \delta$,

$$R_T := \sum_{t=1}^{T} r_t \leq 2\beta_T(\delta) \sqrt{2T c_\lambda \log \frac{\det(A_T)}{\det(\lambda I_D)}},$$

where

$$\beta_t(\delta) := R \sqrt{2 \log \frac{\det(A_t)^{1/2}}{\det(\lambda I_D)^{1/2} \delta}} + \sqrt{\lambda} S_\Delta$$

and

$$c_\lambda := \max \left\{ 1, \frac{L^2}{\lambda} \right\}.$$

Proof. We prove the bound on the event $\mathcal{E}_T(\delta)$, which holds with probability at least $1-\delta$. Define

$$w_t := x_t^\top A_{t-1}^{-1} x_t = \|x_t\|_{A_{t-1}^{-1}}^2 .$$

We prove the regret bound on the event $\mathcal{E}_T(\delta)$. By Lemma A.11, for every $t = 1, \dots, T$,

$$r_t \le 2\beta_{t-1}(\delta)\sqrt{x_t^\top A_{t-1}^{-1} x_t} = 2\beta_{t-1}(\delta)\sqrt{w_t}.$$

Since

$$A_t = A_{t-1} + x_t x_t^\top ,$$

we have

$$A_t \succeq A_{t-1}.$$

Moreover, by the matrix determinant lemma,

$$\det(A_t) = \det(A_{t-1})\left(1 + x_t^\top A_{t-1}^{-1} x_t\right) = \det(A_{t-1})(1 + w_t).$$

Hence $\det(A_t)$ is nondecreasing in t, and therefore β_t is also nondecreasing. Thus,

$$\beta_{t-1}(\delta) \le \beta_T(\delta)$$

for all $t \le T$.

Therefore,

$$R_T = \sum_{t=1}^{T} r_t \le 2\beta_T(\delta) \sum_{t=1}^{T} \sqrt{w_t}.$$

By Cauchy's inequality,

$$\sum_{t=1}^{T} \sqrt{w_t} \le \sqrt{T \sum_{t=1}^{T} w_t}.$$

Thus,

$$R_T \le 2\beta_T(\delta)\sqrt{T \sum_{t=1}^{T} w_t}.$$

It remains to upper bound $\sum_{t=1}^{T} w_t$. By Assumption A.3,

$$\|x_t\|_2 \le L.$$

Since

$$A_{t-1} \succeq \lambda I_D,$$

we have

$$A_{t-1}^{-1} \preceq \frac{1}{\lambda} I_D.$$

Therefore,

$$w_t = x_t^\top A_{t-1}^{-1} x_t \le \frac{\|x_t\|_2^2}{\lambda} \le \frac{L^2}{\lambda}.$$

Since

$$c_\lambda = \max\left\{1, \frac{L^2}{\lambda}\right\},$$

we have for every t,

$$w_t \le c_\lambda \min\{1, w_t\}.$$

Indeed, if $w_t \le 1$, then

$$w_t \le c_\lambda w_t = c_\lambda \min\{1, w_t\},$$

and if $w_t > 1$, then

$$w_t \le \frac{L^2}{\lambda} \le c_\lambda = c_\lambda \min\{1, w_t\}.$$

Now use the standard elliptical potential argument. For every $u \geq 0$,

$$\min\{1, u\} \leq 2\log(1+u).$$

Therefore,

$$\sum_{t=1}^{T} w_t \leq c_\lambda \sum_{t=1}^{T} \min\{1, w_t\} \leq 2c_\lambda \sum_{t=1}^{T} \log(1+w_t).$$

Using the determinant identity above,

$$\sum_{t=1}^{T} \log(1+w_t) = \sum_{t=1}^{T} \log \frac{\det(A_t)}{\det(A_{t-1})} = \log \frac{\det(A_T)}{\det(A_0)}.$$

Since

$$A_0 = \lambda I_D,$$

we obtain

$$\sum_{t=1}^{T} w_t \leq 2c_\lambda \log \frac{\det(A_T)}{\det(\lambda I_D)}.$$

Substituting this into the previous regret bound gives

$$R_T \leq 2\beta_T \sqrt{T \cdot 2c_\lambda \log \frac{\det(A_T)}{\det(\lambda I_D)}}.$$

Hence,

$$R_T \leq 2\beta_T \sqrt{2Tc_\lambda \log \frac{\det(A_T)}{\det(\lambda I_D)}}.$$

Since the event $\mathcal{E}_T(\delta)$ holds with probability at least $1-\delta$, the theorem follows. □

Corollary 5.2 (Cumulative regret analysis for LLM-PriorCB)**.** *Suppose the assumptions in Theorem 5.1 hold. Consider the disjoint linear model*

$$\mu(z_{t,a}) = z_{t,a}^\top \theta_a^*, \qquad z_{t,a} \in \mathbb{R}^D, \quad a \in [K],$$

with arm-specific prior parameters $\theta_{0,a} \in \mathbb{R}^D$. Then the arm-wise LLM-PriorCB rule induced by this block representation satisfies, with probability at least $1-\delta$,

$$R_T \leq 2\beta_T^{\mathrm{dis}}(\delta) \sqrt{2Tc_\lambda \log \frac{\prod_{a=1}^{K} \det(A_{T,a})}{\lambda^{Kd}}},$$

where

$$A_{t,a} := \lambda I_d + \sum_{i=1}^{t} \mathbf{1}\{a_i = a\} z_{i,a} z_{i,a}^\top$$

and

$$\beta_t^{\mathrm{dis}}(\delta) := R\sqrt{2\log \frac{\prod_{a=1}^{K} \det(A_{t,a})^{1/2}}{\lambda^{Kd/2}\delta}} + \sqrt{\lambda} S_\Delta.$$

Proof. Let $e_a \in \mathbb{R}^K$ denote the a-th standard basis vector and define the block-lifted feature

$$\widetilde{z}_{t,a} = e_a \otimes z_{t,a} \in \mathbb{R}^{Kd}.$$

Also define the stacked true and prior parameters

$$\Theta^* = (\theta_1^{*\top}, \ldots, \theta_K^{*\top})^\top, \qquad \Theta_0 = (\theta_{0,1}^\top, \ldots, \theta_{0,K}^\top)^\top.$$

Then, for every arm a,

$$\widetilde{z}_{t,a}^\top \Theta^* = z_{t,a}^\top \theta_a^* = \mu(z_{t,a}).$$

Thus the disjoint model is a special case of the linear model in Theorem 5.1, with feature $\widetilde{z}_{t,a}$ and parameter Θ^*.

Define the global prior-centered design matrix and response vector in the block representation by

$$\widetilde{A}_t = \lambda I_{Kd} + \sum_{i=1}^{t} \widetilde{z}_{i,a_i} \widetilde{z}_{i,a_i}^\top,$$

$$\widetilde{b}_t = \lambda \Theta_0 + \sum_{i=1}^{t} y_i \widetilde{z}_{i,a_i}.$$

Since $\widetilde{z}_{i,a_i}$ has nonzero entries only in the block corresponding to arm a_i, we have

$$\widetilde{A}_t = \operatorname{diag}\left(A_{t,1}, \ldots, A_{t,K}\right),$$

and

$$\widetilde{b}_t = (b_{t,1}^\top, \ldots, b_{t,K}^\top)^\top.$$

Therefore,

$$\widetilde{A}_t^{-1} \widetilde{b}_t = (\hat{\theta}_{t,1}^\top, \ldots, \hat{\theta}_{t,K}^\top)^\top.$$

Moreover, for any candidate arm a,

$$\widetilde{z}_{t,a}^\top \widetilde{A}_{t-1}^{-1} \widetilde{z}_{t,a} = z_{t,a}^\top A_{t-1,a}^{-1} z_{t,a},$$

and

$$\widetilde{z}_{t,a}^\top \widetilde{A}_{t-1}^{-1} \widetilde{b}_{t-1} = z_{t,a}^\top \hat{\theta}_{t-1,a}.$$

Hence the UCB rule in Theorem 5.1 coincides with the arm-wise PriorCB-Disjoint rule.

Finally,

$$\det(\widetilde{A}_t) = \prod_{a=1}^{K} \det(A_{t,a}), \qquad \det(\lambda I_{Kd}) = \lambda^{Kd}.$$

Also,

$$\|\Theta^* - \Theta_0\|_2 = \left(\sum_{a=1}^{K} \|\theta_a^* - \theta_{0,a}\|_2^2 \right)^{1/2}.$$

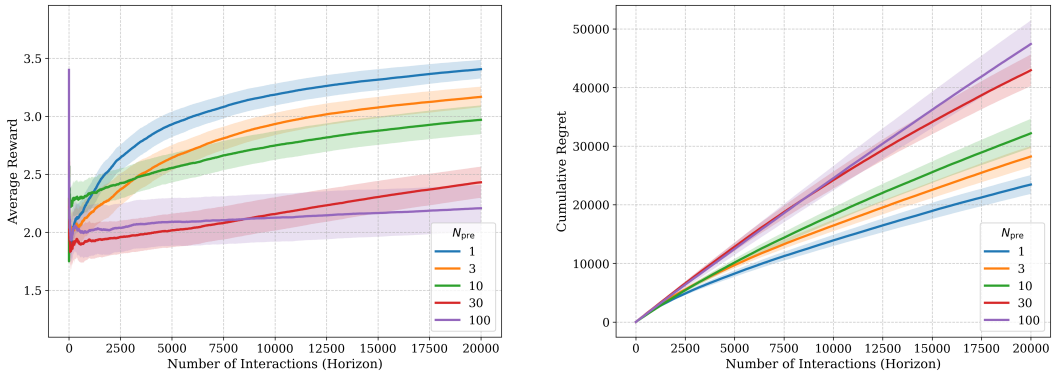
Substituting these identities into Theorem 5.1 gives

$$R_T \leq 2\beta_T^{\mathrm{dis}}(\delta) \sqrt{2T c_\lambda \log \frac{\prod_{a=1}^{K} \det(A_{T,a})}{\lambda^{Kd}}},$$

which proves the claim. □

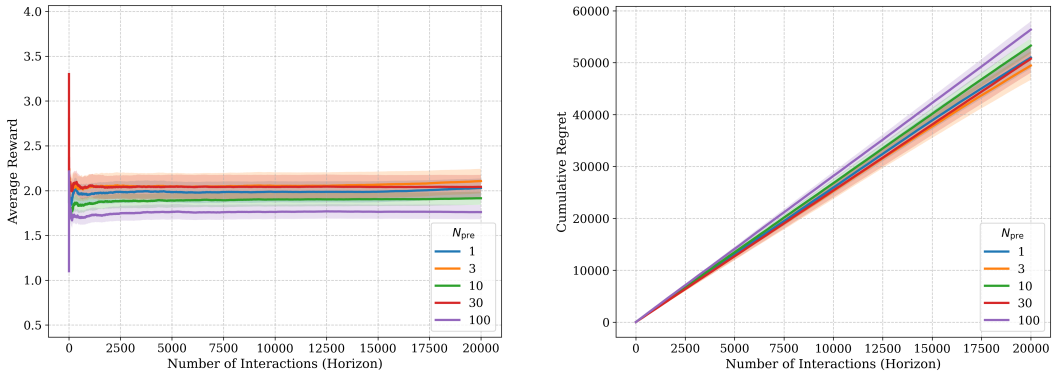
B Ablation Studies on MovieLens

B.1 CBLI



(a) Average reward (b) Cumulative regret

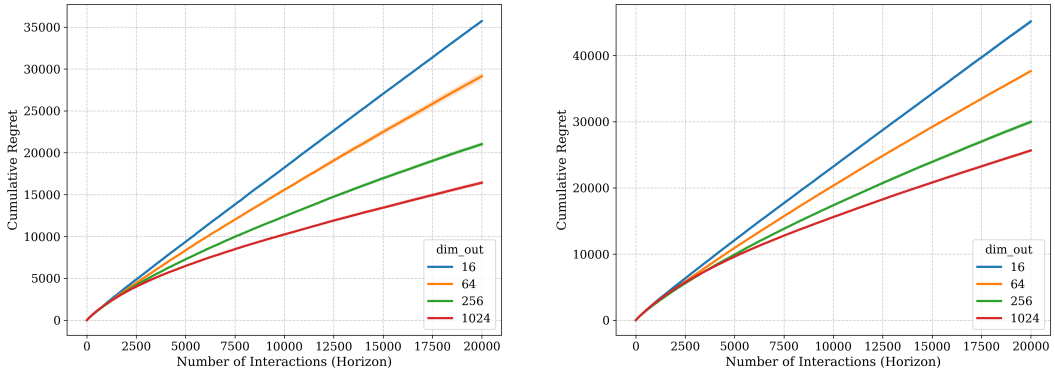
Figure 5: Average reward and cumulative regret of CBLI under oracle regime with varying numbers of pretraining steps. Here, the total pretraining budget scales as $N_{\text{pre}} \times |\mathcal{C}_0| \times |\mathcal{A}_0|$. Interestingly, increasing N_{pre} leads to degraded online performance. This is because excessive pretraining induces an overconfident prior, causing the UCB confidence intervals to collapse and the agent to overly exploit estimated rewards as if they were ground truth, thereby suppressing exploration.



(a) Average reward (b) Cumulative regret

Figure 6: Average reward and cumulative regret of CBLI under text-only regime with varying numbers of pretraining steps. Here, the total pretraining budget scales as $N_{\text{pre}} \times |\mathcal{C}_0| \times |\mathcal{A}_0|$. Interestingly, increasing N_{pre} cannot lead to improve online performance. This is because excessive pretraining induces an overconfident prior, causing the UCB confidence intervals to collapse and the agent to overly exploit estimated rewards as if they were ground truth, thereby suppressing exploration.

B.2 Output Dimension

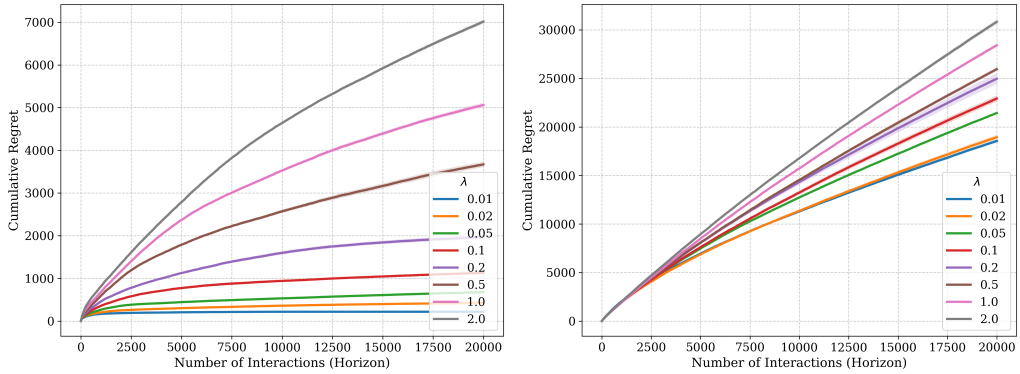


(a) LLM-PriorCB with $K = 10$

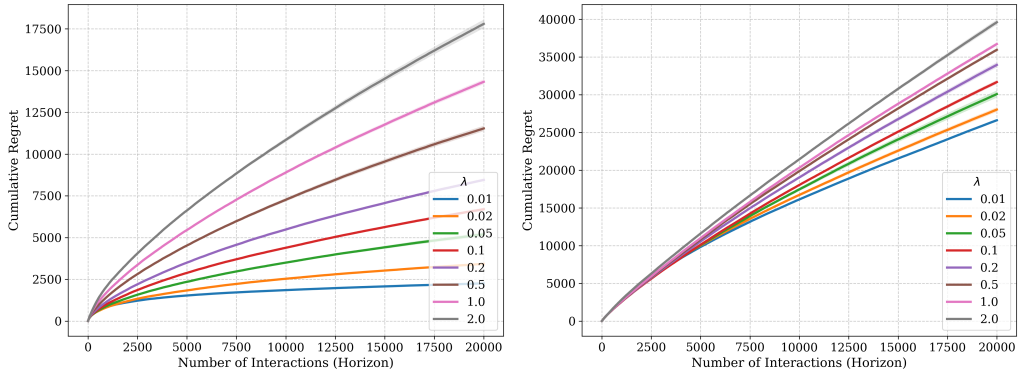
(b) LLM-PriorCB with $K = 30$

Figure 7: Cumulative regret of our algorithm with varying feature vector dimensions. We empirically find that higher-dimensional representations achieve lower cumulative regret across the horizon regardless of the number of arms. We conjecture that increased dimensionality provides greater representational capacity to capture relevant contextual information, thereby improving reward estimation during online learning.

B.3 λ



(a) Our algorithm with $K = 10$

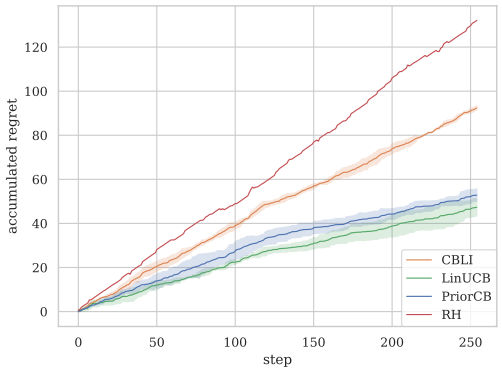


(b) Our algorithm with $K = 30$

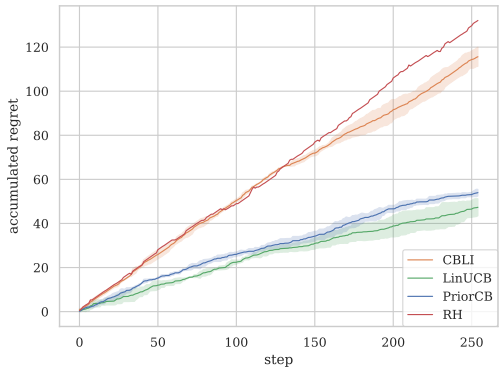
Figure 8: We report cumulative regret of LLM-PriorCB for $\lambda \in [0.01, 2.0]$ under (a) the oracle regime and (b) the text-only regime. Smaller λ consistently yields lower regret, while larger λ degrades performance, indicating that overly confident priors can suppress exploration and hinder effective online adaptation.

C Ablation Studies on GAIA

C.1 Cumulative regret



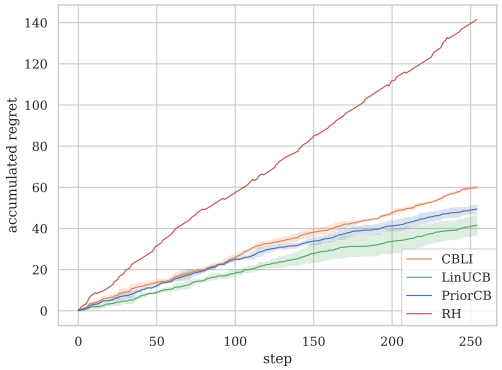
(a) Cumulative regret of Prior from different eval contexts



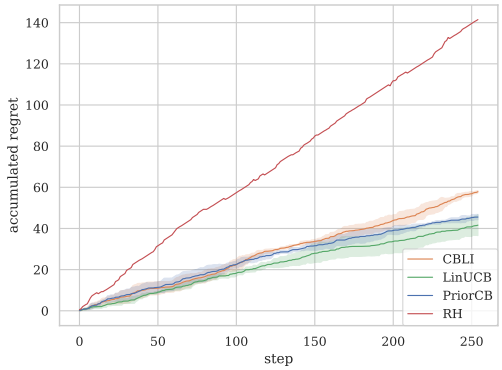
(b) Cumulative regret of Prior from same eval contexts

Figure 9: Performance as a function of the interaction horizon when $K = 3$ using gpt-5.2. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels.

Especially, Figure 9 presents cumulative regret in the early stage of online learning for PriorCB and baselines-including RH, where RH results are shown only for the first three epochs due to cost and time constraints. The near-linear growth of the cumulative regret of LLM clearly indicates that an LLM alone struggles to effectively solve the contextual bandit problem. In contrast, PriorCB—leveraging the LLM as a prior—achieves cumulative regret in the early stage that is comparable to LinUCB, indicating effective learning. Additional results using GPT-5-mini are provided in the Appendix C.1. The average reward in the early stage of online learning is reported in the Appendix C.2.

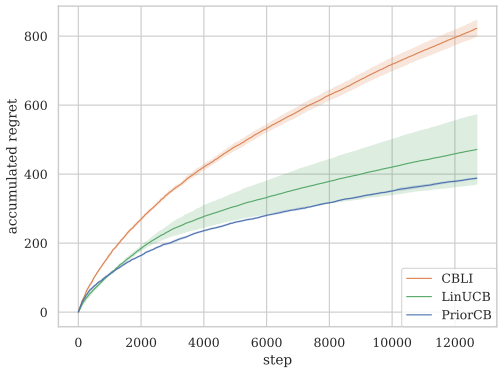


(a) Cumulative regret of Prior from different eval contexts

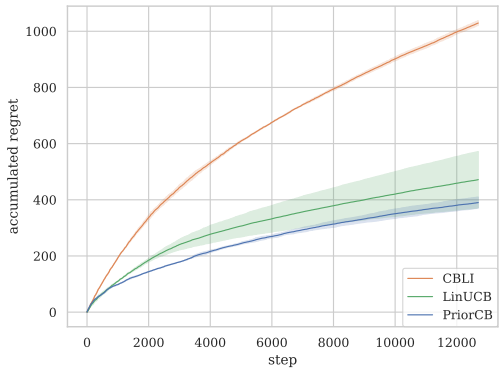


(b) Cumulative regret of Prior from same eval contexts

Figure 10: Performance as a function of the interaction horizon when $K = 3$ using gpt-5-mini. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels.



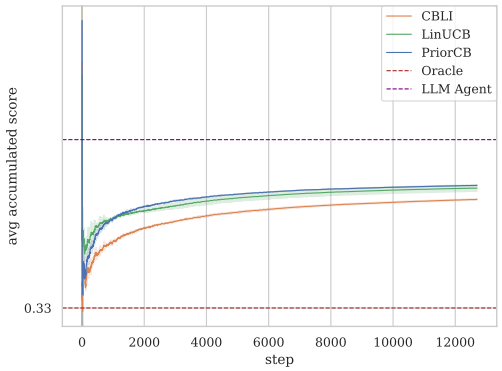
(a) Cumulative regret of Prior from different eval contexts



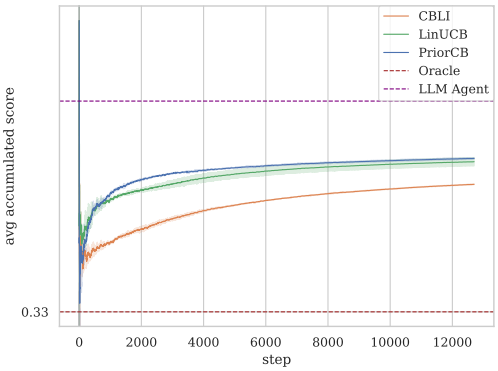
(b) Cumulative regret of Prior from same eval contexts

Figure 11: Performance as a function of the interaction horizon when $K = 3$ using gpt-5-mini. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels.

C.2 Average reward

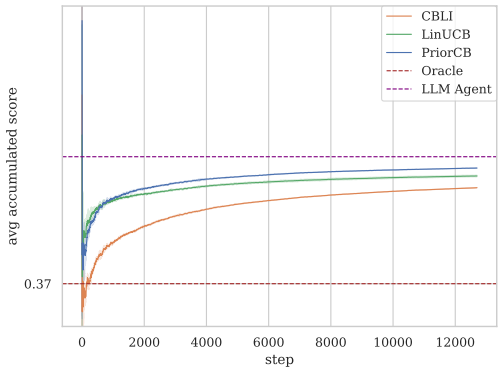


(a) Average reward of Prior from different eval contexts

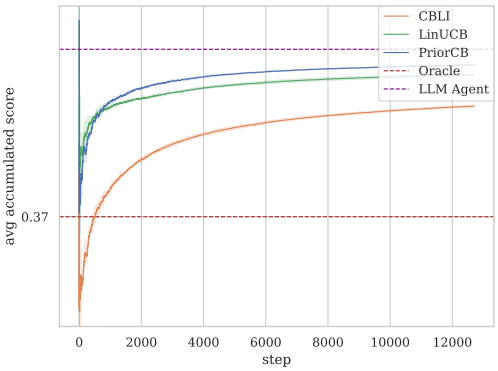


(b) Average reward of Prior from same eval contexts

Figure 12: Performance as a function of the interaction horizon when $K = 3$ using gpt-5-mini with medium reasoning effort. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels. Oracle denotes the tool choice that achieves best reward, whereas LLM Agent denotes the tool choice selected by the LLM agent.

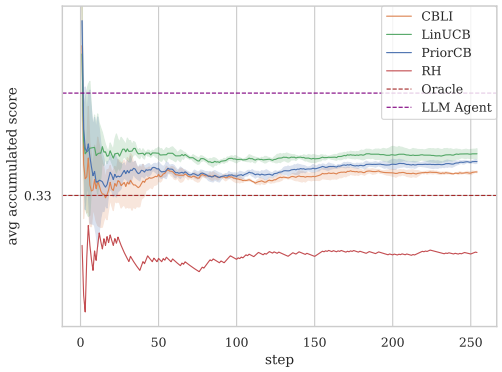


(a) Average reward of Prior from different eval contexts

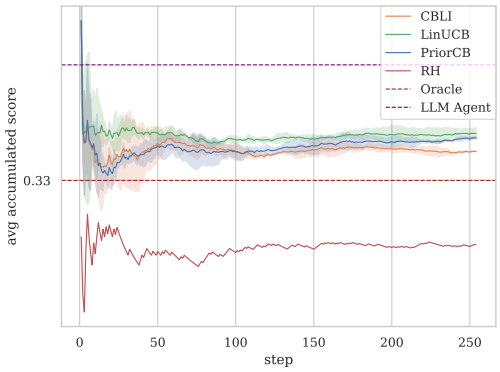


(b) Average reward of Prior from same eval contexts

Figure 13: Performance as a function of the interaction horizon when $K = 3$ using gpt-5.2 with medium reasoning effort. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels. Oracle denotes the tool choice that achieves best reward, whereas LLM agent denotes the tool choice selected by the LLM agent.

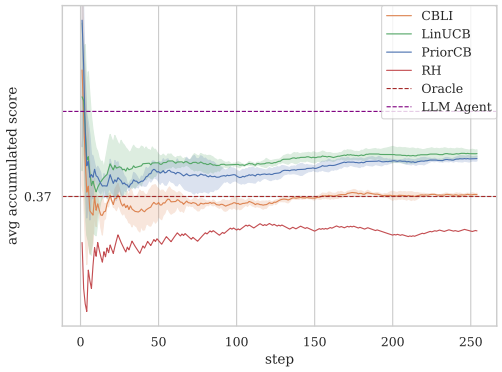


(a) Average reward of Prior from different eval contexts

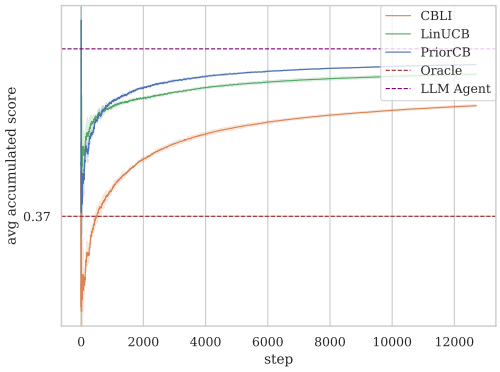


(b) Average reward of Prior from same eval contexts

Figure 14: Performance as a function of the interaction horizon when $K = 3$ using gpt-5-mini with medium reasoning effort. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels. Oracle denotes the tool choice that achieves best reward, whereas LLM agent denotes the tool choice selected by the LLM agent.



(a) Average reward of Prior from different eval contexts



(b) Average reward of Prior from same eval contexts

Figure 15: Performance as a function of the interaction horizon when $K = 3$ using gpt-5.2 with medium reasoning effort. Shaded regions indicate $\pm$ one standard error over 3 independent runs. We use the validation split for evaluation due to answer-label availability. Priors are constructed either from the validation set (same eval contexts) or from the test set (different eval contexts), using no labels. Oracle denotes the tool choice that achieves best reward, whereas LLM Agent denotes the tool choice selected by the LLM agent.

D Experimental Details

D.1 Movie Lens

Tool Description: search_deep

Purpose: Run a deeper web search to answer questions that require multiple independent lookups and cross-checking.

What it does:

- Generates **three distinct, high-precision search queries**.
- Retrieves results for **all three** queries.
- **Synthesizes** the evidence to produce a final answer.

Use when (triggers):

- The task is **multi-hop** (needs 2+ facts or subquestions).
- The answer benefits from **cross-validation** across sources.
- Entities are **ambiguous**, niche, or likely to have changed.
- You need **citations** from multiple sources for reliability.

Do not use when (anti-triggers):

- A **single lookup** is enough (use `search_light`).
- The question can be answered reliably from **parametric knowledge** alone.

Cost: High (3 web searches).

D.2 GAIA

D.2.1 Tool Descriptions

Tool Description: reason_only

Purpose: answer using only internal (parametric) knowledge without any web search.
What it does:

- Produces an answer using only internal knowledge and reasoning.
- Performs no web search.
- Lowest cost.

Use when (triggers):

- The question is reasoning-heavy (math/logic/proof/code reading/conceptual explanation).
- Facts are stable and unlikely to have changed (basic definitions, well-known principles).
- The task is to transform/rewrite/summarize provided text (no external facts needed).
- You are confident the answer does not depend on the latest information or specific citations.

Do not use when (anti-triggers):

- The question requires up-to-date facts (current events, latest versions, prices, active roles/titles).
- The question requests sources, citations, or "according to" style evidence.
- High risk of hallucination: niche proper nouns, many numbers, exact dates/titles.

Output expectation:

- Provide the best possible answer from internal knowledge.
- If uncertainty is high or factuality is critical, recommend escalation to `search_lite`/`search_heavy`.

Common failure modes:

- Outdated facts.
- Overconfidence on niche or recent details.

Tool Description: search_light

Purpose: quickly verify or look up a single key fact via the web.
What it does:

- Runs EXACTLY ONE web search with ONE query.
- Uses the retrieved results as evidence for the final answer.
- Medium cost (1 search).

Use when (triggers):

- The question can be resolved by ONE primary lookup (one fact / definition / spec / date).
- Internal knowledge may be outdated or uncertain, but only a single check is needed.
- The task benefits from a source but does not require multiple independent sources.
- A high-precision query is obvious from the question.

Do not use when (anti-triggers):

- The question is multi-hop (requires 2+ independent facts or subquestions).
- You need cross-validation across sources (high risk of misinformation or ambiguous entities).
- The answer requires aggregation/comparison across multiple entities (list/compare/rank).
- The problem is pure reasoning/math/logic that does not benefit from web facts.

Output expectation:

- The final answer must be supported by evidence from the search results.
- If evidence is insufficient, state so and escalate to `search_heavy` if needed.

Common failure modes:

- Wrong or overly broad query $\rightarrow$ irrelevant results (one-shot failure).
- A single source may be insufficient for disambiguation.

Query writing hint (1-shot):

- Include unique identifiers (proper nouns, version/year) and intent words such as "official", "documentation", "paper pdf".

Tool Description: search_deep

Purpose: solve questions that require multiple lookups or stronger evidence via multi-query web search.
What it does:

- Generates THREE distinct web-search queries.
- Retrieves results for all three queries.
- Combines and cross-checks the evidence to support the final answer.
- Highest cost (3 searches).

Use when (triggers):

- The question is multi-hop (requires 2+ subfacts; e.g., find a value, verify a condition, then produce a final result).
- The question demands high precision (exact number/date/quote/title) where small errors cause failure.
- The question implies freshness or current status (e.g., "latest", "current", "updated", "as of <date>").
- The answer benefits from cross-validation (multiple independent sources or official + secondary confirmation).
- Ambiguity is likely (multiple entities with the same name) and disambiguation is needed.

Do not use when (anti-triggers):

- A single lookup is sufficient (use `search_lite`).
- The task is pure reasoning with stable facts (use `knowledge_only`).
- You have no clear way to form three meaningfully distinct queries (risk of wasted cost).

Output expectation:

- Synthesize evidence across queries.
- If sources conflict, explicitly note the conflict and prefer more authoritative/primary sources.
- The final answer must be traceable to the retrieved evidence.

Suggested 3-query decomposition:

1. Core fact/definition (what is X?).
2. Exact detail (number/date/spec/quote).
3. Verification (official document or independent confirmation).

Common failure modes:

- Query diversity too low $\rightarrow$ redundant evidence.
- Over-searching simple questions $\rightarrow$ unnecessary cost.

D.2.2 Reward Generation Prompt

Reward Generation Prompt

```
REWARD_W_COST_SYSTEM_PROMPT = (
    "You are an evaluation engine for cost-aware tool selection.\n"
    "Your objective is to maximize expected utility: higher correctness with
    lower cost.\n"
    "Choose the single tool with the best cost--accuracy tradeoff.\n"
    "Think carefully, but do NOT output any reasoning.\n"
    "Output MUST be strict JSON only."
)

REWARD_W_COST_USER_PROMPT = (
    "Select exactly ONE tool to answer the question.\n"
    "Objective: maximize cost-aware correctness (expected utility).\n"
    "Decision rule:\n"
    "- Prefer the lowest-cost tool that is likely to answer correctly.\n"
    "- Choose a higher-cost tool only if it meaningfully increases the
    probability of correctness.\n\n"
    "Question:\n"
    "{question}\n\n"
    "Available tools (name: description; cost is implied by the tool
    choice):\n"
    "{tool_descriptions_in_text}\n\n"
    "Strict output format (JSON only; exactly one key):\n"
    "{\"answer\":\"<tool_name>\"}"
)
```

D.2.3 Hyperparameters

Method	Setting
CBLI	epoch = 10
PriorCB	epoch = 10, learning_rate = 1×10^{-3}, weight_decay = 1×10^{-4}

Table 1: Training settings.