TOWARDS OPTIMALITY IN ACTIVE PREFERENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning large language models (LLMs) depends on high-quality datasets of human preference labels, which are costly to collect. Although active learning has been studied to improve sample efficiency relative to passive collection, many existing approaches adopt classical experimental design criteria such as G- or D-optimality. These objectives are not tailored to the structure of preference learning, leaving open the design of problem-specific algorithms. In this work, we identify a simple intuition specific to preference learning that calls into question the suitability of these existing design objectives. Motivated by this insight, we propose two active learning algorithms. The first provides the first instance-dependent label complexity guarantee for this setting, and the second is a simple, practical greedy method. We show the optimality of our instance-dependent term up to a constant depending on the logit bound, compared with a lower-bound characteristic. We evaluate our algorithms on real-world preference datasets and synthetic data observing improved sample efficiency compared to existing methods.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) is a central post-training approach for aligning large language models (LLMs) with human preferences, addressing limitations of standard supervised fine-tuning. A core component of RLHF is learning a reward model that captures human judgments about the quality of model-generated responses. This reward model then guides optimization of the LLM, often via reinforcement learning algorithms such as Proximal Policy Optimization (PPO) (Ouyang et al., 2022).

The standard paradigm for learning a reward model is pairwise preference learning. In this setup, human labelers compare two LLM-generated responses to the same prompt and indicate which one they prefer. The resulting data are used to fit a reward model, commonly the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952). This paradigm is motivated by the cognitive observation that relative comparisons are typically easier and more reliable for humans than assigning absolute scores to individual items.

However, collecting human preference labels is expensive and time-consuming. This challenge is amplified when high-quality labels are required, such as in safety-critical applications like medical assistance, where each label may require expert knowledge and careful judgment. These constraints make data collection difficult to scale and can also compromise label quality (Zhang et al., 2023). Consequently, it is important to develop data-efficient methods that minimize the number of queries to human labelers while maintaining reward model quality.

To this end, we cast preference learning as an active learning problem, in which an algorithm adaptively selects informative pairs to query, since *not all pairs of responses are equally important*. Recent work applies active learning to this setting using well studied criteria from G-optimal and D-optimal experimental design (Das et al., 2025; Mehta et al., 2023; Scheid et al., 2024; Mukherjee et al., 2024). However, these approaches primarily provide worst-case guarantees and do not adapt to the instance-specific difficulty of the underlying preference learning problem. As a result, they may fail to exploit the structure present in a given problem instance, leading to suboptimal sampling strategies.

In this work, we argue for an instance-dependent approach. Our method is guided by a simple intuition: pairs of responses with near-ties in preference (i.e., with small reward differences) are the most ambiguous; therefore, querying them is informative. Building on this idea, we develop

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

two complementary algorithms. The first, based on a novel experimental design formulation, has an instance-dependent label-complexity bound whose leading term is near-optimal relative to a lower bound for logistic instances with bounded true logits, up to a constant depending on the logit bound. The second is a practical, easily implemented greedy method based on a new uncertainty sampling heuristic.

We summarize our contributions as follows:

- We propose a novel experimental design objective for active preference learning that, to the best of our knowledge, is the first to yield an instance-dependent label complexity guarantee for this active preference learning problem. We also show an information-theoretic lower bound and compare it with the leading term for bounded-logit logistic instances.
- We design a simple and practical greedy algorithm based on a new uncertainty sampling heuristic. The method is easy to implement and avoids solving any complex optimization problems.
- We validate our methods on real-world preference datasets and synthetic data. The results show that the proposed algorithms consistently improve upon existing baselines for serveral different type of experiments including reward model accuracy and best-of-N-sampling.

Notation: Throughout the paper, we use the following notation: $[n] = \{1, 2, \dots, n\}$; $\triangle(\mathcal{Z}) = \{\lambda \in \mathbb{R}^{|\mathcal{Z}|} \mid \lambda_{z \in \mathcal{Z}} \geq 0, \sum_{z \in \mathcal{Z}} \lambda_z = 1\}$ is the probability simplex over the set $\mathcal{Z}$; $\|z\|_A = \sqrt{z^\top A z}$ is the norm of z with respect to a positive definite matrix A; $H(\lambda, \theta) = \sum_{z \in \mathcal{Z}} \lambda_z \dot{\mu}\left(z^\top \theta\right) z z^\top$, where $\dot{\mu}$ is the derivative of a function μ.

2 Related Work

The RLHF pipeline was popularized by InstructGPT (Ouyang et al., 2022), although the underlying idea dates back to earlier work of Christiano et al. (2017). The now-standard optimization algorithm in this setting, Proximal Policy Optimization (PPO), was introduced in Schulman et al. (2017). Subsequently, alternative approaches such as Direct Preference Optimization (DPO) have been proposed Rafailov et al. (2023). Despite differences in methodology, these alignment techniques all require high-quality preference datasets, motivating the development of sample-efficient data collection methods.

A recent line of work studies active learning for preference collection to improve sample efficiency (Das et al., 2025; Mehta et al., 2023; Scheid et al., 2024; Zhang et al., 2025; Dwaracherla et al., 2024; Muldrew et al., 2024; Mukherjee et al., 2024; Liu et al., 2024; Schlaginhaufen et al., 2025; Kveton et al., 2025). Mehta et al. (2023) was among the first to formalize active learning for RLHF, proposing an active contextual dueling bandit approach with a worst-case suboptimality gap guarantee. Das et al. (2025) subsequently relaxed the assumptions to better match the BTL model, and Liu et al. (2024) further extended the framework to account for labelers with heterogeneous expertise. Other work adopts alternative problem formulations, including best-arm identification (Scheid et al., 2024) and online, rather than offline, alignment (Zhang et al., 2025). Notably, all these results are worst-case and do not capture instance-dependent difficulty. More recently, active learning has also been developed specifically for the DPO pipeline (Muldrew et al., 2024; Kveton et al., 2025).

Many of these methods are grounded in experimental design, with D-optimal design being particularly popular (Mukherjee et al., 2024; Liu et al., 2024; Schlaginhaufen et al., 2025). The foundational work in this method dates back to Kiefer & Wolfowitz (1960). Recent papers adapt D-optimal design to preference collection in several ways. Mukherjee et al. (2024) proposed D-optimal experimental design algorithms for various feedback models and evaluated them using ranking based loss functions rather than the more common classification accuracy; however, their procedures are non-adaptive and use fixed allocations. Liu et al. (2024) proposed a dual active learning approach based on D-optimal design that also models labeler expertise. Other work combines randomized exploration with D-optimal design (Schlaginhaufen et al., 2025). D-optimal design has also been adapted to DPO, with worst-case theoretical guarantees on logit estimation error (Kveton et al., 2025).

Beyond alignment-oriented preference learning, active learning and, more broadly, interactive machine learning are increasingly applied to other aspects of LLM use, including in-context learning (Margatina et al., 2023) and prompt engineering (Diao et al., 2024; Shi et al., 2024).

3 PRELIMINARIES

Consider a set of prompts $\mathcal{X}$ with $|\mathcal{X}| = n$, where each prompt $x \in \mathcal{X}$ is associated with two candidate responses, a_1 and a_2. The unlabeled dataset consists of triplets (x, a_1, a_2). Such response pairs are typically generated either by sampling the same LLM under different settings or by using two different LLMs. Following bandit terminology, we refer to each triplet as an arm, and let $z = \phi(x, a_1, a_2) \in \mathcal{Z} \subset \mathbb{R}^d$ be a feature mapping satisfying the antisymmetry property, i.e., $\phi(x, a_1, a_2) = -\phi(x, a_2, a_1)$ for any x, a_1, a_2. As in Das et al. (2025), we assume $\|z\| \leq 1$ for all $z \in \mathcal{Z}$. The preference label is $y \in \{1, 0\}$, where $y = 1$ indicates that the first response is preferred and $y = 0$ indicates that the second is preferred. We write $a_1 \succeq a_2$ to denote that a_1 is preferred over a_2.

We consider a realizable setting. There exists a true parameter $\theta^* \in \mathbb{R}^d$ such that the preference model satisfies, for any $z \in \mathcal{Z}$,

$$p(y = 1|x, a_1, a_2) = \psi\big(z^\top \theta^*\big), \tag{1}$$

where ψ is a monotone link function mapping the score $z^\top \theta^*$ to a probability. This framework includes several standard choices, including the logistic link $\psi(x) = \frac{1}{1+e^{-x}}$, the linear link $\psi(x) = \frac{x+1}{2}$ for $x \in [-1, 1]$, and the Gaussian CDF link $\psi(x) = \Phi(x)$, where Φ is the cumulative distribution function of the standard Gaussian distribution. In the preference learning literature, the logistic model is often referred to as the BTL model, and it is the primary focus of this paper. We assume that the minimum absolute margin is strictly positive, $\Delta = \min_{z \in \mathcal{Z}} |z^\top \theta^*| > 0$.

We formulate the pairwise preference learning task as an active classification problem for the sign of $z^\top \theta^*$. The goal is to learn a classifier that correctly predicts the preferred response for every pair by actively and adaptively selecting an arm at each round and querying human annotators for the corresponding preference label.

At every round t, based on all the data collected up to time t, the learner selects an arm and queries the human annotator for its preference label. The learner then receives the label that is generated by eq. (1). This sequential protocol can be seen as a special case of a batched setting, which is more realistic in large-scale label collection where the learner can select a batch of arms and query for their preference labels at once (Scheid et al., 2024). The labeled preference dataset at time t can thus be represented as $\mathcal{D}_t = \big\{(z_s, y_s) \mid s \in [t]\big\}$. A classifier $\hat{\theta}_t$ can be trained based on $\mathcal{D}_t$ by solving the following maximum likelihood estimation (MLE) problem,

$$\hat{\theta}_t = \arg\max_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} y_s \log \psi(z_s^\top \theta) + (1 - y_s) \log\big(1 - \psi(z_s^\top \theta)\big). \tag{2}$$

An arm z is said to be classified correctly by the classifier $\hat{\theta}_t$ if $\operatorname{sign}(z^\top \hat{\theta}_t) = \operatorname{sign}(z^\top \theta^*)$. The goal of the learner is to design a strategy that can classify all the arms correctly with high probability. We define the following notion of probably approximately correct (PAC) strategy.

Definition 1. *(δ-PAC) A strategy is said to be δ-PAC if, for any problem instance $(\mathcal{Z}, \theta^*)$ and any $\delta > 0$, it can be executed in a finite number of steps and with probability at least $1 - \delta$, it outputs a classifier that can classify all the arms correctly.*

Our goal is to design a δ-PAC strategy, while minimizing the number of queries made to human annotators.

Remark. *Previous works Das et al. (2025) and Mehta et al. (2023) use the notion of a suboptimality gap to evaluate the performance of the learned reward model, which is defined as the maximum difference between the true reward of the best response and the true reward of the response selected by the policy, often the* greedy *policy defined by $\hat{\theta}$. Formally, it can be expressed as*

$$\mathrm{SubOpt}(\pi) = \max_{x \in \mathcal{X}} \max_{a \in \mathcal{A}} \big(r^*(x, a) - r^*(x, \pi(x))\big),$$

where $\pi(x)$ is the policy defined by $\hat{\theta}$, $\mathcal{A}$ is a set of all responses, and $r^*(x, a)$ is a reward function. This measure has two key limitations. First, it focuses exclusively on the worst-case context, indeed all of the previous works with this measure so far are only able to show a worst-case guarantee in the end, and fail to capture instance-dependent behavior, ignoring the structural nuances of the problem instance. In contrast, we will show later in this paper that our approach is able to provide an instance-dependent guarantee. Second, it is insensitive to finer-grained preferences between responses, which is crucial for the alignment task. For example, consider a prompt x with three

different responses satisfying $a_1 \succeq a_2 \succeq a_3$, a reward model that correctly identifies a_1 as the best response (resulting in zero SubOpt*) may still fail to distinguish the preference between a_2 and a_3. Indeed, a reward model is not just to pick a winner, but to provide a nuanced understanding of the relative quality of all responses. Note that given any number of responses, we can always form pairs from them.*

3.1 Confidence Interval

Our approach relies on two key ingredients: a confidence width of the MLE estimator in eq. (2) and the relative position of the resulting confidence interval, as we clarify in the next section. For completeness, we state here a well-grounded confidence interval in the literature, used throughout the rest of the paper. However, our framework is general and does not depend on this particular choice.

Theorem 1. *(Theorem 1 of Jun et al. (2021)) Let $\delta \le e^{-1}$. Let $\hat{\theta}_t$ be the solution of eq. (2) where, for every $s \in [t]$, y_s is conditionally independent from $\{z_i\}_{i=1}^t \setminus \{z_s\}$ given z_s (i.e., the z_s's are a fixed design). Fix $z \in \mathbb{R}^d$ with $\|z\| \le 1$. Let t_{eff} be the number of distinct vectors in $\{z_s\}_{s=1}^t$. Define*

$$\gamma(d) = 64\left(d\log(6) + \log\left(\frac{2+t_{eff}}{\delta}\right)\right).$$

Define the event

$$\mathcal{E}_{var} = \left\{\forall z',\ \frac{1}{\sqrt{2.2}}\|z'\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \le \|z'\|_{H'(\{z_s, s\in[t]\},\theta_t)^{-1}} \le \sqrt{2.2}\,\|z'\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}}\right\}.$$

If

$$\xi_t^2 := \max_{s\in[t]} \|z_s\|^2_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \le \frac{1}{\gamma(d)},$$

then

$$\mathbb{P}\left(\left|z^\top(\hat{\theta}_t - \theta^*)\right| > D_{\delta,t}(z) := 2.4\|z\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} \cdot \sqrt{\log\frac{2(2+t_{eff})}{\delta}},\ \mathcal{E}_{var}\right) \le \delta,$$

where $H'(\mathcal{D},\theta) = \sum_{z\in\mathcal{D}} \dot{\mu}\left(z^\top\theta\right) zz^\top$ is the Fisher information matrix at θ.

3.2 Intuition

The core intuition of our approach is to leverage, for each arm $z \in \mathcal{Z}$, the *individual* location of the confidence interval for $z^\top\theta^*$. This contrasts with most existing works, which primarily aim to reduce *overall* uncertainty in the estimate of θ^*. For instance, D-optimal experimental design minimizes the volume of the confidence ellipsoid for θ^*, while G-optimal design minimizes the maximum prediction variance over arms. Equivalently, these criteria depend only on the widths of the confidence intervals for $z^\top\theta^*$ and ignore their locations.

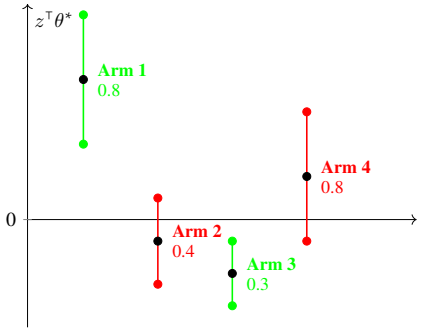


Figure 1: Intuition for our proposed method. An arm's sign is considered confidently determined (green) if its confidence interval does not overlap with zero, and uncertain (red) otherwise. The numbers indicate the widths of the confidence intervals.

We illustrate our intuition in fig. 1 using four arms with distinct confidence intervals on $z^\top\theta^*$. The varying widths of these intervals reflect different levels of uncertainty at a given time t during the learning process. Algorithms that are designed to minimize uncertainty of each individual arm, such as G-optimal experimental design (Das et al., 2025), would prioritize arms with larger uncertainty, i.e., wide confidence intervals, such as arms 1 and 4 in the figure. Although this strategy may seem reasonable from a general uncertainty reduction perspective and is often used in many heuristics, from a sign classification perspective, this strategy can be suboptimal. For example, for arm 1, the entire confidence interval is above zero, so we can predict its sign correctly with high probability despite the interval's large width. In contrast, the confidence interval for arm 2 includes zero. This indicates that under the current estimate $\hat{\theta}$, its sign remains ambiguous, even

though its associated uncertainty width is relatively small. Therefore, intuitively it is more beneficial to pull arm 2 to reduce its uncertainty until its confidence interval is fully separated from zero.

In the next two sections, we formalize the above intuition into two algorithm designs.

4 Active Preference Learning via Experimental Design

In this section, we develop a δ-PAC algorithm for the active preference learning problem based on the experimental design framework.

4.1 Optimal Allocation

To motivate our algorithm, we first analyze the optimal query allocation in an oracle setting where the true parameter θ^* is known. This allocation represents the ideal distribution of preference queries across pairs of arms. Subsequently, we will introduce an adaptive algorithm that approximates this optimal allocation without knowledge of θ^*.

We begin by defining $\mathcal{C}_\theta$ as the set of all parameters θ that correctly classify all arms,

$$\mathcal{C}_\theta := \left\{\theta \in \mathbb{R}^d \mid \forall z \in \mathcal{Z}, \operatorname{sign}\left(z^\top \theta\right) = \operatorname{sign}\left(z^\top \theta^*\right)\right\}.$$

Intuitively, the algorithm should be designed to return a classifier $\hat{\theta}$ that belongs to the set $\mathcal{C}_\theta$ with high probability and minimal label complexity. This requires an optimal allocation over the arms $\mathcal{Z}$ that efficiently exploits the structure of $\mathcal{C}_\theta$, ensuring the classifier $\hat{\theta}$ converges to this set as quickly as possible. We define

$$\mathcal{Z}^+_{\theta^*} := \left\{z \in \mathcal{Z} \mid z^\top \theta^* > 0\right\} \quad \text{and} \quad \mathcal{Z}^-_{\theta^*} := \left\{z \in \mathcal{Z} \mid z^\top \theta^* < 0\right\}.$$

By theorem 1, we have

$$z^\top \theta^* - z^\top \hat{\theta}_t \le \mathcal{C} \cdot \|z\|_{H'(\{z_s, s\in[t]\}, \theta^*)^{-1}} \sqrt{\log\left(t/\delta\right)}, \tag{3}$$

where $\mathcal{C}$ is a positive constant introduced to simplify notation. For any arm $z \in \mathcal{Z}^+_{\theta^*}$, the requirement $z^\top \hat{\theta}_t > 0$ is equivalent to:

$$z^\top \theta^* - z^\top \hat{\theta}_t < z^\top \theta^*, \tag{4}$$

Given the confidence bound in eq. (3), a sufficient condition for eq. (4) to hold is

$$\mathcal{C} \cdot \|z\|_{H'(\{z_s, s\in[t]\}, \theta^*)^{-1}} \sqrt{\log\left(t/\delta\right)} < z^\top \theta^*. \tag{5}$$

To ensure $z^\top \hat{\theta}_t > 0$ for all $z \in \mathcal{Z}^+_{\theta^*}$, this must hold for the worst case, leading to the sufficient condition:

$$\frac{1}{\mathcal{C} \cdot \sqrt{\log\left(t/\delta\right)}} \ge \max_{z\in\mathcal{Z}} \frac{\|z\|_{H'(\{z_s, s\in[t]\}, \theta^*)^{-1}}}{z^\top \theta^*} = \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(t \sum_{z\in\mathcal{Z}} \lambda_{z,t}\, \dot{\mu}(z^\top \theta^*)\, z z^\top\right)^{-1}}}{z^\top \theta^*},$$

where $\lambda_{z,t} = \sum_{s=1}^t \mathbf{1}\{z_s = z\}/t$ is the empirical frequency of arm z till time t. This motivates choosing a design that satisfies this sufficient condition as quickly as possible,

$$\lambda^*_+ = \arg \min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(\sum_{z\in\mathcal{Z}} \lambda_z \dot{\mu}(z^\top \theta^*) z z^\top\right)^{-1}}}{z^\top \theta^*}. \tag{6}$$

Similarly, for an arm $z \in \mathcal{Z}^-_{\theta^*}$, a parallel argument leads to the design:

$$\lambda^*_- = \arg \min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(\sum_{z\in\mathcal{Z}} \lambda_z \dot{\mu}(z^\top \theta^*) z z^\top\right)^{-1}}}{-z^\top \theta^*}. \tag{7}$$

Combining the objectives from eq. (6) and eq. (7) yields a unified design that ensures the parameter estimate converges to the true parameter θ^* at a fast rate,

$$\lambda^* = \arg \min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|_{\left(\sum_{z\in\mathcal{Z}} \lambda_z \dot{\mu}(z^\top \theta^*) z z^\top\right)^{-1}}}{\left|z^\top \theta^*\right|}.$$

Comparison to G-Optimal and XY-Optimal Designs. Readers familiar with experimental design methods, particularly in multi-armed bandits, may have noticed a similarity between our design objective and the G-optimal or XY-optimal design objective as frequently used in best arm identification problems (Soare et al., 2014; Zhao et al., 2024; Fiez et al., 2019; Yang & Tan, 2022). However, there are crucial differences. The G-optimal design aims to minimize the maximum prediction variance,

$$\arg \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z \in \mathcal{Z}} \|z\|_{(\sum_{z \in \mathcal{Z}} \lambda_z z z^\top)^{-1}}.$$

The core distinction in our objective is the instance-dependent denominator $|z^\top \theta^*|$, which is absent in the G-optimal design objective. As explained in section 3.2, this term augments the G-optimal design objective by considering the confidence interval's location relative to the decision boundary. This augmentation is the key to achieving our instance-dependent results in the next section, moving beyond worst-case guarantees (Das et al., 2025; Mukherjee et al., 2024; Liu et al., 2024).

Our objective also differs from the XY-optimal design (Fiez et al., 2019; Zhao et al., 2024), which is defined as,

$$\arg \min_{\lambda \in \Delta(\mathcal{Z})} \max_{z_1, z_2 \in \mathcal{Z}} \frac{\|z_1 - z_2\|_{(\sum_{z \in \mathcal{Z}} \lambda_z z z^\top)^{-1}}}{|(z_1 - z_2)^\top \theta^*|}.$$

While the XY-optimal design also has an instance-dependent denominator, it is tailored specifically for best arm identification, as the allocation focuses on reducing uncertainty in the direction of $(z_1 - z_2)$ for all $z_1, z_2 \in \mathcal{Z}$. This is suitable for identifying a single winner among all arms. In contrast, our objective is designed to classify all arms, focusing on resolving the uncertainty of each individual arm z with respect to the decision boundary. There are also other differences between our problem and the best arm identification problem, which we explain in the remark of next section.

4.2 Algorithm Design with Unknown θ^*

Guided by the oracle allocation from the previous section, we now present an adaptive algorithm that approximates this allocation without access to the true parameter θ^*. Our approach is inspired by the sequential experimental design framework Fiez et al. (2019); Jun et al. (2021); Zhao et al. (2024), which has been applied successfully to various multi-armed bandit problems with linear structure.

The key idea is that, to classify all arms with high confidence, it suffices to shrink the confidence interval for each $z^\top \theta^*$ until it no longer overlaps zero. Once an interval is separated from zero, the sign of the corresponding $z^\top \theta^*$ is determined with high confidence by eq. (4)-eq. (5). The number of samples required for a given arm $z \in \mathcal{Z}$ depends on the magnitude of $z^\top \theta^*$. The oracle design from the previous section provides an allocation that achieves this separation efficiently by adapting to the structure of the problem instance.

The full algorithm is presented in algorithm 2. It proceeds in two phases: a warm-up procedure to obtain an initial estimate of θ^*, and a main sequential experimental design procedure. The warm-up phase, detailed in algorithm 1, collects an initial dataset by sampling according to a G-optimal design.

The main phase of the algorithm operates in rounds. At each round ℓ, we maintain a set of 'active' arms, $\mathcal{Z}_\ell$, which includes all arms whose signs have not yet been confidently determined by the previous estimate, $\hat{\theta}_{\ell-1}$. We start with $\mathcal{Z}_1 = \mathcal{Z}$. The algorithm also uses an exponentially decreasing precision parameter, ϵ_ℓ, that controls the effective distance to the decision boundary for the active set. In each round, we compute a new design, λ_ℓ, based on the estimate $\hat{\theta}_{\ell-1}$ and the active set $\mathcal{Z}_\ell$. Arms that are far from the decision boundary, i.e., where $\left|z^\top \theta^*\right|$ is large, are expected to be classified with high confidence in early rounds and are thus removed from the active set.

We note that the warm-up procedure requires knowledge of a lower bound on the minimum derivative of the link function at the true parameter.

Remark. *The final, non-adaptive sampling phase in algorithm 2 (lines 16-17) is a critical step. It is necessitated by a technical requirement of the confidence bound in theorem 1 and same for many other similar confidence bounds (Abbasi-Yadkori et al., 2011), which requires that data is collected non-adaptively. That is, the selection of arms cannot depend on previous observations. Consequently, the confidence guarantee would not hold if we were to compute the final MLE using only the data gathered adaptively during the main rounds. This requirement distinguishes our setting from many standard sequential design problems (Fiez et al., 2019; Jun et al., 2021; Zhao et al., 2024), where the goal is to identify a single best arm. In contrast, our goal is to return a reward model, which necessitates a valid confidence bound on the final parameter estimate itself.*

Algorithm 1 Warm-up procedure

Input: $\mathcal{Z}, \delta, \omega, \kappa_0 = \min_{z\in\mathcal{Z}} \dot{\mu}\big(z^\top\theta^*\big)$
Initialize:

$$\lambda_0 = \arg\min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \|z\|^2_{(\sum_{z\in\mathcal{Z}}\lambda_z zz^\top)^{-1}}$$

Set: $N_0 = \left\lceil 3(1+\omega)\kappa_0^{-1} d\gamma(d)\log\Big(2|\mathcal{Z}|\big(2+|\mathcal{Z}|\big)/\delta\Big)\right\rceil$
Pull arms in $Z_{N_0} = \text{ROUND}(\lambda_0, N_0)$ and observe Y_{N_0}
return: MLE $\hat{\theta}_0$ by eq. (2)

Algorithm 2 Experimental design algorithm

Input $\mathcal{Z}, \delta, \omega$
Initialize: $\ell = 1, \mathcal{Z}_1 = \mathcal{Z}, \epsilon_\ell = 2^{-\ell+1}, Q = \varnothing$
$\hat{\theta}_0$ = algorithm 1 (warm-up procedure)
Set

$$f_1(\hat{\theta}_{\ell-1}, \lambda) := \gamma(d)\max_{z\in\mathcal{Z}} \|z\|^2_{H(\lambda,\hat{\theta}_{\ell-1})^{-1}}$$
$$f_2(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda) := 2.4^2\epsilon_\ell^{-2}\max_{z\in\mathcal{Z}_\ell} \|z\|^2_{H(\lambda,\hat{\theta}_{\ell-1})^{-1}}$$
$$f(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda) := f_1(\hat{\theta}_{\ell-1}, \lambda) \vee f_2(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda)$$

while $|\mathcal{Z}_\ell| \geq 1$ **do**
  $\lambda_\ell = \arg\min_{\lambda\in\Delta(\mathcal{Z})} f(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda)$
  $\rho(\mathcal{Z}_\ell) = \min_{\lambda\in\Delta(\mathcal{Z})} f(\mathcal{Z}_\ell, \hat{\theta}_{\ell-1}, \lambda)$
  $N_\ell = \left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell)\log\left(\frac{2\ell^2|\mathcal{Z}|(2+|\mathcal{Z}|)}{\delta}\right)\right\rceil \vee r(\omega)$
  $Z_\ell = \text{ROUND}(\lambda_\ell, N_\ell)$
  $Q \leftarrow Q \cup Z_\ell$
  Pull arms in Z_ℓ and observe Y_ℓ
  Compute MLE $\hat{\theta}_\ell$ with data $\{Z_\ell, Y_\ell\}$ by eq. (2)
  $\mathcal{Z}_{\ell+1} = \mathcal{Z}_\ell \backslash \left\{z \in \mathcal{Z}_\ell \middle| \left|z^\top\hat{\theta}_\ell\right| > \epsilon_\ell\right\}$
  $\ell \leftarrow \ell + 1$
end while
Pull arms in Q and observe Y
Compute MLE $\hat{\theta}$ with data $\{Q, Y\}$ by eq. (2)
Output: $\hat{\theta}$

We show the following two instance dependent quantities that essentially capture the complexity of the problem instance,

$$\rho^* = \min_{\lambda\in\Delta(\mathcal{Z})}\max_{z\in\mathcal{Z}} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{|z^\top\theta^*|^2} \quad \text{and} \quad \rho_0 = \min_{\lambda\in\Delta(\mathcal{Z})}\max_{z\in\mathcal{Z}} 3\gamma(d)\|z\|^2_{H(\lambda,\theta^*)^{-1}}.$$

We define the following shorthand for simplicity before we are ready to present an upper bound for the label complexity of algorithm 2, $\ell^* = \left\lceil\log\big(4\Delta^{-1}\big)\right\rceil$, $\Delta = \min_{z\in\mathcal{Z}}\left|z^\top\theta^*\right|$, $\overline{\log} = \log\left(\frac{2\ell^{*2}|\mathcal{Z}|(2+|\mathcal{Z}|)}{\delta}\right)$.

Theorem 2. *(Label complexity) With a probability of at least $1-\delta$, algorithm 2 will stop after at most $c(1+\omega)\ell^*\overline{\log}\rho^* + c(1+\omega)\ell^*\overline{\log}\rho_0 + c(1+\omega)\kappa_0^{-1}d\gamma(d)\overline{\log} + c\ell^* r(\omega)$ rounds, where c is an absolute constant.*

We defer the proof to section A.1 of the Appendix.

Theorem 3. *(Correctness) With a probability of at least $1-\delta$, the reward model $\hat{\theta}$ will classify all arms in $\mathcal{Z}$ correctly, i.e.,*

$$\text{sign}(z^\top\hat{\theta}) = \text{sign}(z^\top\theta^*) \quad \textit{for } \forall z \in \mathcal{Z}.$$

We defer the proof to section A.2 of the Appendix.

Our result stands in stark contrast to existing work on active learning for RLHF. Prior works Das et al. (2025); Kveton et al. (2025); Mehta et al. (2023); Scheid et al. (2024), for instance, have analyzed various performance metrics, such as suboptimality gap, simple regret, and logit error, but have only provided worst-case guarantees. Their convergence rates, typically on the order of $O(d/\sqrt{T})$, lack dependence on the problem instance and therefore fail to capture the intrinsic difficulty of a given problem instance.

In contrast, our results are instance-dependent. This is because our approach identifies an experimental design objective that is intrinsically tied to the problem structure. Previous works have relied on standard, off-the-shelf objectives like G-optimality and D-optimality. While these are powerful general purpose tools, they are not specifically tailored to the active learning problem in RLHF, which limits the tightness of their theoretical guarantees. We provide an information-theoretic lower bound in section B of the Appendix and compare its instance-dependent characteristic with ρ^* below for the logistic link. We also provide a concrete instance in the next section, showing a qualitative gap between a G-optimality method and a greedy method based on the same intuition in section 3.2.

4.3 Near-optimality under bounded logits

We now compare the leading instance-dependent term in theorem 2 with a lower bound. Let $\mathcal{Z} = \{z_1, \dots, z_n\}$, and let $\mathcal{C}_{\theta^*,\mathcal{Z}}$ be the set of parameters that change at least one arm's sign, as defined in theorem 6 of the Appendix. Define the lower-bound characteristic

$$T^\star_{\mathrm{KL}} := \inf_{\lambda \in \Delta(\mathcal{Z})} \sup_{\theta \in \mathcal{C}_{\theta^*,\mathcal{Z}}} \frac{1}{\sum_{i=1}^n \lambda_i D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big)},$$

where λ_i is the sampling mass on z_i and $P_{\theta,i} = \mathrm{Bern}\big(\psi(z_i^\top \theta)\big)$. Then theorem 6 shows that $T^\star_{\mathrm{KL}} \log\big(1/(2.4\delta)\big)$ lower-bounds the expected label complexity of strategies that are uniformly δ-PAC over logistic instances. The lower-bound statement and its proof are given in section B of the Appendix.

Theorem 4 (Leading-term comparison for bounded logits)**.** *Consider the logistic link* $\psi(u) = 1/(1+e^{-u})$*. Let* $\mathcal{Z} = \{z_1, \dots, z_n\}$ *span* $\mathbb{R}^d$*, and suppose*

$$\max_{z \in \mathcal{Z}} |z^\top \theta^*| \le B, \qquad \kappa_B := \psi(B)(1-\psi(B)).$$

Then

$$T^\star_{\mathrm{KL}} \ge 8 \kappa_B \rho^*. \tag{8}$$

Thus, the leading ρ^* term in theorem 2 is upper bounded by the lower bound characteristic up to the constant $1/(8\kappa_B)$. We defer the proof to section B.1 of the Appendix.

5 Greedy Algorithm

In this section, we propose a simple greedy algorithm for the active preference learning problem that is directly motivated by the intuition from section 3.2. In contrast to the method in section 4, this approach avoids solving a complex optimization problem and does not require any knowledge of the minimum derivative of the link function. The key idea is to use a greedy objective that incorporates the location of each confidence interval relative to the decision boundary. As argued in section 3.2, relying solely on confidence width can lead to inefficient sampling. Our objective therefore uses the interval's location to augment its width.

Specifically, for each arm z, we define its *remaining uncertainty* as the minimum amount by which its confidence interval must shrink to no longer overlap with the decision boundary (zero). This value captures the uncertainty that must be resolved to classify the arm

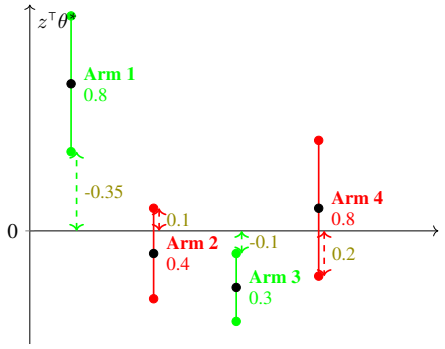


Figure 2: Illustration of remaining uncertainty (in olive). The remaining uncertainty is positive only when a confidence interval contains the decision boundary (zero), indicating that the arm's sign is not yet confidently determined.

Algorithm 3 Greedy Algorithm for Active Preference Learning

Input $\mathcal{Z}$, δ, labeling budget T
Initialize $\mathcal{D} \leftarrow \varnothing$
for $t = 1, \ldots, T$ **do**
  Compute $\hat{\theta}_t$ and $(\mathsf{LCB}_t(z), \mathsf{UCB}_t(z))$ for all $z \in \mathcal{Z}$
  $z_t \leftarrow \arg\max_{z \in \mathcal{Z}} \min\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\}$ with ties broken arbitrarily.
  Query z_t to observe preference y_t.
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(z_t, y_t)\}$.
end for
return: MLE $\hat{\theta}$ on $\mathcal{D}$ using eq. (2).

confidently. Formally, the remaining uncertainty of arm z at round t is defined as:

$$\mathsf{RU}_t(z) = \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\},$$

where $\mathsf{LCB}_t(z)$ and $\mathsf{UCB}_t(z)$ are the lower and upper confidence bounds of $z^\top \theta^*$ at round t respectively, i.e., $\mathsf{LCB}_t(z) = z^\top \hat{\theta}_t - D_{\delta,t}(z)$ and $\mathsf{UCB}_t(z) = z^\top \hat{\theta}_t + D_{\delta,t}(z)$ with $D_{\delta,t}(z)$ being the confidence width defined in theorem 1. This concept is illustrated in fig. 2, which depicts four possible scenarios for the confidence interval locations.

Based on this concept, we present our algorithm in algorithm 3. The algorithm follows a simple greedy strategy: in each round, it pulls the arm with the maximum remaining uncertainty.

The greedy objective in algorithm 3 can be interpreted as a refined measure of uncertainty for this problem. It refines the standard confidence width by also incorporating the confidence interval's location relative to the decision boundary, as shown in the following lemma.

Lemma 4.1. *The greedy objective in algorithm 3 can be equivalently written as*

$$\max_z \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\} = \max_z \left\{D_{\delta,t}(z) - \left|\hat{\theta}_t^\top z\right|\right\}. \tag{9}$$

We defer the proof to section A.3 of the Appendix. This form balances the uncertainty reduction of G-/D-optimal design with Uncertainty Sampling's preference for small estimated margins.

To demonstrate the advantage of our algorithm over a G-optimal based method, e.g., Das et al. (2025) pulling arm based on $\arg\max_{i \in [d]} \|e_i\|_{H_t'^{-1}}$, we show a simple canonical instance where their respective label complexities differ significantly. Consider the following instance. Let ϵ be a small enough constant, e.g., $\epsilon \in (0, 1/4)$. Using the stopping rule from lemma 5.5 in the Appendix, we have the following result.

Theorem 5. *For the canonical instance $z_i = e_i$ for $i = 1, \ldots, d$ and $\theta^* = (1, \ldots, 1, \epsilon)^\top$ for $\epsilon \in (0, 1/4)$, the greedy selection objective $\arg\max_{i \in [d]} \|e_i\|_{H_t'^{-1}}$ has label complexity scaling as $\tilde{O}\left(\frac{d}{\epsilon^2}\right)$, while with algorithm 3, the label complexity scales as $\tilde{O}\left(d + \frac{1}{\epsilon^2}\right)$.*

We defer the proof to section A.4 of the Appendix.

6 Experiments

Due to space limitations, experimental details are discussed in section C of the Appendix.

7 Conclusion

In this work, we addressed the problem of active preference learning for RLHF by proposing a novel, tailored experimental design objective. Based on this objective, we developed an efficient algorithm that achieves an instance-dependent label complexity bound, improving upon existing worst-case only results, and showing the instance-dependent near optimality for a leading term with explicit dependence on the logit bound. We also introduced a simple and practical greedy algorithm as an alternative. However, another question remains open: how does the reward model's performance eventually impact the performance of the LLM alignment? Investigating this connection and developing a more precise experimental design objective that directly targets the overall RLHF goal are compelling directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to polish wording, correct grammar and typos, adjust LaTeX spacing, and refine figure presentation. We also used these tools for mathematical verification. We take responsibility for the final content of this work.

REFERENCES

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

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

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 96–112. Springer, 2025.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. Active prompting with chain-of-thought for large language models. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pp. 1330–1350. Association for Computational Linguistics (ACL), 2024.

Vikranth Dwaracherla, Seyed Mohammad Asghari, Botao Hao, and Benjamin Van Roy. Efficient exploration for LLMs. In *International Conference on Machine Learning*, pp. 12215–12227. PMLR, 2024.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pp. 5148–5157. PMLR, 2021.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Branislav Kveton, Xintong Li, Julian McAuley, Ryan Rossi, Jingbo Shang, Junda Wu, and Tong Yu. Active learning for direct preference optimization. *arXiv preprint arXiv:2503.01076*, 2025.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pp. 13–19. ACM New York, NY, USA, 1995.

Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. Provably efficient online rlhf with one-pass reward modeling. *arXiv preprint arXiv:2502.07193*, 2025.

Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. Active learning principles for in-context learning with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint arXiv:2312.00267*, 2023.

Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Anand Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human preference elicitation. *Advances in Neural Information Processing Systems*, 37:90132–90159, 2024.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *International Conference on Machine Learning*, pp. 36577–36590. PMLR, 2024.

Stephen Mussmann and Percy S Liang. Uncertainty sampling is preconditioned stochastic gradient descent on zero-one loss. *Advances in Neural Information Processing Systems*, 31, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Anant Raj and Francis Bach. Convergence of uncertainty sampling for active learning. In *International Conference on Machine Learning*, pp. 18310–18331. PMLR, 2022.

Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I Jordan, Pierre Ménard, Eric Moulines, and Michal Valko. Optimal design for reward modeling in rlhf. *arXiv preprint arXiv:2410.17055*, 2024.

Andreas Schlaginhaufen, Reda Ouhamma, and Maryam Kamgarpour. Efficient preference-based reinforcement learning: Randomized exploration meets experimental design. *arXiv preprint arXiv:2506.09508*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in neural information processing systems*, 27, 2014.

Gemma Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL `https://www.kaggle.com/m/3301`.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10859–10885, 2023.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Han Zhong, Zhihan Liu, Ziyi Yang, Shuohang Wang, Hany Hassan Awadalla, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *Transactions on Machine Learning Research*, 2025.

Yao Zhao, Kwang-Sung Jun, Tanner Fiez, and Lalit Jain. Adaptive experimentation when you can't experiment. *Advances in neural information processing systems*, 37, 2024.

Appendix

A Proofs

A.1 Proof of Theorem 2

Theorem 2. *(Label complexity) With a probability of at least $1-\delta$, algorithm 2 will stop after at most $c(1+\omega)\ell^*\overline{\log}\rho^* + c(1+\omega)\ell^*\overline{\log}\rho_0 + c(1+\omega)\kappa_0^{-1}d\gamma(d)\overline{\log} + c\ell^* r(\omega)$ rounds, where c is an absolute constant.*

Proof. We prove an upper bound for the label complexity in the following.

Define $\mathcal{S}_\ell := \left\{z \in \mathcal{Z} \middle| |z^\top \theta^*| \le 2\epsilon_\ell\right\}$. Thus, by lemma 5.1, with a probability at least $1-\delta$, we have $\bigcap_\ell \{\mathcal{Z}_\ell \subseteq S_\ell\}$ hold. This implies the following is true with a probability at least $1-\delta$ for all ℓ

$$
\begin{aligned}
\rho(\mathcal{Z}_\ell) &= \min_{\lambda\in\Delta(\mathcal{Z})} \max\left\{\gamma(d)\max_{z\in\mathcal{Z}}\|z\|^2_{H(\lambda,\hat\theta_{\ell-1})^{-1}}, 2.4^2\epsilon_\ell^{-2}\max_{z\in\mathcal{Z}_\ell}\|z\|^2_{H(\lambda,\hat\theta_{\ell-1})^{-1}}\right\} \\
&\le \min_{\lambda\in\Delta(\mathcal{Z})} \max\left\{\gamma(d)\max_{z\in\mathcal{Z}}\|z\|^2_{H(\lambda,\hat\theta_{\ell-1})^{-1}}, 2.4^2\epsilon_\ell^{-2}\max_{z\in\mathcal{S}_\ell}\|z\|^2_{H(\lambda,\hat\theta_{\ell-1})^{-1}}\right\} && (\mathcal{Z}_\ell \subseteq S_\ell) \\
&\le \min_{\lambda\in\Delta(\mathcal{Z})} \max\left\{3\gamma(d)\max_{z\in\mathcal{Z}}\|z\|^2_{H(\lambda,\theta^*)^{-1}}, 3\cdot 2.4^2\epsilon_\ell^{-2}\max_{z\in\mathcal{S}_\ell}\|z\|^2_{H(\lambda,\theta^*)^{-1}}\right\} && (\text{lemma } 5.3) \\
&\le \min_{\lambda\in\Delta(\mathcal{Z})}\max_{z\in\mathcal{Z}} 3\gamma(d)\|z\|^2_{H(\lambda,\theta^*)^{-1}} + \min_{\lambda\in\Delta(\mathcal{Z})}\max_{z\in\mathcal{S}_\ell} 3\cdot 2.4^2\epsilon_\ell^{-2}\|z\|^2_{H(\lambda,\theta^*)^{-1}} \\
&=: \rho_0 + \rho_\ell. && (10)
\end{aligned}
$$

By lemma 5.1, with a probability at least $1-\delta$, we have $S_\ell = \varnothing$ for $\ell \ge \ell^* = \left\lceil \log\left(4\Delta^{-1}\right)\right\rceil$. The total label complexity of algorithm 2 is the summation of the number of samples in each round, which can be upper bounded as

$$
\begin{aligned}
&\sum_{\ell=1}^{\ell^*} N_\ell \\
&= \sum_{\ell=1}^{\ell^*}\left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell)\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)\right\rceil \vee r(\omega) \\
&\le \sum_{\ell=1}^{\ell^*} 4(1+\omega)\rho(\mathcal{Z}_\ell)\overline{\log} \vee r(\omega) \\
&\le \sum_{\ell=1}^{\ell^*} 4(1+\omega)(\rho_0+\rho_\ell)\overline{\log} \vee r(\omega) && (\text{due to } (10)) \\
&\le \sum_{\ell=1}^{\ell^*} 4(1+\omega)\overline{\log}\rho_\ell + 4(1+\omega)\ell^*\overline{\log}\rho_0 + \ell^* r(\omega). && (11)
\end{aligned}
$$

On the other hand, we have

$$\begin{aligned}
\rho^* &= \min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in\mathcal{Z}} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{\left|z^\top\theta^*\right|^2} \\
&= \min_{\lambda\in\Delta(\mathcal{Z})} \max_{\ell} \max_{z\in S_\ell} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{\left|z^\top\theta^*\right|^2} \\
&\geq \frac{1}{\ell^*} \min_{\lambda\in\Delta(\mathcal{Z})} \sum_{\ell=1}^{\ell^*} \max_{z\in S_\ell} \frac{\|z\|^2_{H(\lambda,\theta^*)^{-1}}}{\left|z^\top\theta^*\right|^2} \\
&\geq \frac{1}{16\ell^*} \sum_{\ell=1}^{\ell^*} 2^{2\ell} \min_{\lambda\in\Delta(\mathcal{Z})} \max_{z\in S_\ell} \|z\|^2_{H(\lambda,\theta^*)^{-1}}. \qquad (12)
\end{aligned}$$

Combining (11), (12), and N_0 for the warm-up, we have

$$\sum_{\ell=0}^{\ell^*} N_\ell \leq c(1+\omega)\ell^*\overline{\log}\rho^* + c(1+\omega)\ell^*\overline{\log}\rho_0 + c(1+\omega)\kappa_0^{-1}d\gamma(d)\overline{\log} + c\ell^* r(\omega),$$

where c is an absolute constant. □

A.2 Proof of Theorem 3

Theorem 3. *(Correctness) With a probability of at least $1-\delta$, the reward model $\hat{\theta}$ will classify all arms in $\mathcal{Z}$ correctly, i.e.,*

$$\operatorname{sign}(z^\top\hat{\theta}) = \operatorname{sign}(z^\top\theta^*) \quad \textit{for } \forall z \in \mathcal{Z}.$$

Proof. The proof is similar to lemma 5.1 and lemma 5.2. Readers are recommended to read the proof of lemma 5.1 and lemma 5.2 first. The difference is that here we need to show that the final reward model $\hat{\theta}$ will classify all arms in $\mathcal{Z}$ correctly, while in lemma 5.2, each reward model $\hat{\theta}_\ell$ is only shown to classify the arms in $\mathcal{Z}_\ell\backslash\mathcal{Z}_{\ell+1}$ correctly. Note that since $\mathcal{Z}_\ell$ will be eventually empty, for any $z \in \mathcal{Z}$, there exists ℓ such that $z \in \mathcal{Z}_\ell\backslash\mathcal{Z}_{\ell+1}$. So what we need to show is that for all $z \in \mathcal{Z}_\ell\backslash\mathcal{Z}_{\ell+1}$, the reward model $\hat{\theta}$ will classify z correctly.

With a probability of at least $1-\frac{\delta}{2\ell^2|\mathcal{Z}|}$, we have for all $z \in \mathcal{Z}_\ell$,

$$\begin{aligned}
\left|\left(\hat{\theta}-\theta^*\right)^\top z\right| &\overset{b_1}{\leq} 2.4\|z\|_{H'(Q,\theta^*)^{-1}} \sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
&\leq 2.4\|z\|_{H'(\mathcal{Z}_\ell,\theta^*)^{-1}} \sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \qquad (H'(\mathcal{Z}_\ell,\theta^*) \preceq H'(Q,\theta^*)) \\
&\overset{b_2}{\leq} \frac{2.4\sqrt{1+\omega}\|z\|_{H(\lambda_\ell,\theta^*)^{-1}}}{\sqrt{N_\ell}} \sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
&\overset{b_3}{\leq} \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{\sqrt{N_\ell}} \sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
&= \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{2.4\sqrt{\left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell)\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)\right\rceil \vee r(\omega)}} \sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
&\leq \epsilon_\ell, \qquad (13)
\end{aligned}$$

where b_1 is due to theorem 1, b_2 is due to lemma 5.4, b_3 is due to lemma 5.3. Consider an arm $z \in \mathcal{Z}_\ell\backslash\mathcal{Z}_{\ell+1}$. We have $\left|z^\top\hat{\theta}_\ell\right| > \epsilon_\ell$ by the elimination condition in algorithm 2. Thus by eq. (13) and

lemma 5.5 we know that $\hat{\theta}_\ell$ will classify z correctly. By a union bound same as eq. (15), we know that with a probability at least $1-\delta$, all of arms in $\mathcal{Z}$ will be classified correctly by $\hat{\theta}$. □

Lemma 5.1. *With a probability of at least $1-\delta$, for all $z \in \mathcal{Z}_\ell$, we have $\left|z^\top \theta^*\right| \le 2\epsilon_\ell$, for all $\ell \ge 1$.*

Proof. With a probability of at least $1 - \frac{\delta}{2\ell^2|\mathcal{Z}|}$, we have for all $z \in \mathcal{Z}_\ell$,

$$\begin{aligned}
\left|z^\top\left(\hat{\theta}_\ell - \theta^*\right)\right| \overset{b_1}{\le} & 2.4\|z\|_{H'(\mathcal{Z}_\ell,\theta^*)^{-1}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\overset{b_2}{\le} & \frac{2.4\sqrt{1+\omega}\|z\|_{H(\lambda_\ell,\theta^*)^{-1}}}{\sqrt{N_\ell}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\overset{b_3}{\le} & \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{\sqrt{N_\ell}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
= & \frac{2.4\sqrt{3(1+\omega)}\|z\|_{H\left(\lambda_\ell,\hat{\theta}_{\ell-1}\right)^{-1}}}{2.4\sqrt{\left\lceil 3(1+\omega)\rho(\mathcal{Z}_\ell)\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)\right\rceil \vee r(\omega)}}\sqrt{\log\left(\frac{2\ell^2|\mathcal{Z}|\left(2+|\mathcal{Z}|\right)}{\delta}\right)} \\
\le & \epsilon_\ell,
\end{aligned} \tag{14}$$

where b_1 is due to theorem 1, b_2 is due to lemma 5.4, b_3 is due to lemma 5.3.

Define a good event $\mathcal{E}_{\ell,z}$ for each ℓ and $z \in \mathcal{Z}_\ell$ as

$$\mathcal{E}_{\ell,z} = \left\{\left|z^\top\left(\hat{\theta}_\ell - \theta^*\right)\right| \le \epsilon_\ell\right\}.$$

We claim that with a probability at least $1-\delta$, the event $\bigcap_{\ell=1}^\infty \bigcap_{z\in\mathcal{Z}_\ell} \mathcal{E}_{\ell,z}$ holds. It can be proved by a union bound as follows,

$$\begin{aligned}
\mathbb{P}\left(\left(\bigcap_{\ell=1}^\infty \bigcap_{z\in\mathcal{Z}_\ell} \mathcal{E}_{\ell,z}\right)^c\right) \le & \sum_{\ell=1}^\infty \sum_{z\in\mathcal{Z}_\ell} \mathbb{P}\left(\mathcal{E}_{\ell,z}^c\right) \\
\le & \sum_{\ell=1}^\infty \sum_{z\in\mathcal{Z}_\ell} \frac{\delta}{2\ell^2|\mathcal{Z}|} \\
\le & \frac{\delta}{2}\sum_{\ell=1}^\infty \frac{1}{\ell^2} \\
\le & \delta,
\end{aligned} \tag{15}$$

where the last step is by the fact that $\sum_{\ell=1}^\infty \frac{1}{\ell^2} = \frac{\pi^2}{6} < 2$. Under the event $\bigcap_{\ell=1}^\infty \bigcap_{z\in\mathcal{Z}_\ell} \mathcal{E}_{\ell,z}$, we show the following.

For all $z \in \mathcal{Z}_\ell$ satisfying $\left|z^\top\theta^*\right| > 2\epsilon_\ell$, we have, if $z^\top\theta^* > 2\epsilon_\ell$,

$$\begin{aligned}
z^\top\hat{\theta}_\ell = & z^\top\left(\hat{\theta}_\ell - \theta^*\right) + z^\top\theta^* \\
> & -\epsilon_\ell + 2\epsilon_\ell \\
> & \epsilon_\ell.
\end{aligned} \tag{16}$$

If $z^\top\theta^* < -2\epsilon_\ell$,

$$\begin{aligned}
z^\top\hat{\theta}_\ell = & z^\top\left(\hat{\theta}_\ell - \theta^*\right) + z^\top\theta^* \\
< & \epsilon_\ell - 2\epsilon_\ell \\
< & -\epsilon_\ell.
\end{aligned} \tag{17}$$

Combining eq. (16) and eq. (17), we obtain $\left|z^\top \hat{\theta}_\ell\right| > \epsilon_\ell$, which means z satisfying $\left|z^\top \theta^*\right| > 2\epsilon_\ell$ will be removed from $\mathcal{Z}_\ell$ at the end of round ℓ according to the elimination condition in algorithm 2. □

Lemma 5.2. *For all $\ell \geq 1$, with a probability at least $1-\delta$, the arms in $\mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$ are all classified correctly by $\hat{\theta}_\ell$.*

Proof. This result directly follows from eq. (14), lemma 5.5, and the elimination condition in algorithm 2.

By eq. (14), we know that the confidence width of $z^\top \theta^*$ is at most ϵ_ℓ for all $z \in \mathcal{Z}_\ell$. By the elimination condition in algorithm 2, we have $\left|z^\top \hat{\theta}_\ell\right| > \epsilon_\ell$ for all $z \in \mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$, thus the sufficient condition in lemma 5.5 is satisfied, which means $\hat{\theta}_\ell$ will classify all arms in $\mathcal{Z}_\ell \backslash \mathcal{Z}_{\ell+1}$ correctly with a probability at least $1-\delta$. □

Lemma 5.3. *(Lemma 7 of Jun et al. (2021)) For each $\ell \geq 0$, define the event,*

$$\mathcal{R}_\ell = \left\{ \frac{1}{3} H(\lambda, \theta^*) \leq H(\lambda, \hat{\theta}_\ell) \leq 3H(\lambda, \theta^*), \forall \lambda \in \Delta(\mathcal{Z}) \right\}.$$

We have that

- $\mathbb{P}(\mathcal{R}_0) \geq 1 - 2\delta$ *and* $\mathbb{P}\big(\mathcal{R}_\ell \mid \mathcal{R}_{\ell-1}, \cdots, \mathcal{R}_0\big) \geq 1 - 2\delta$ *for all* $\ell \geq 1$*;*
- $\max_{z \in \mathcal{Z}} \|z\|^2_{H'(\mathcal{Z}_\ell, \theta^*)^{-1}} \leq 1/\gamma(d)$.

Lemma 5.4. *(Lemma 13 of Jun et al. (2021)) Assume that $\lambda \in \Delta_{\mathcal{X}}$, and that we have sampled $x_1, \ldots, x_n \sim$ ROUND(λ, n, ϵ) with $n \geq r(\epsilon) = \frac{(d(d+1)+2)}{\epsilon}$, and $\epsilon \leq 1$. Then, for any θ,*

$$\sum_{s=1}^{n} \dot{\mu}(x_s^\top \theta) x_s x_s^\top \geq \frac{n}{1+\epsilon} \sum_{x \in \mathcal{X}} \lambda_x \dot{\mu}(x^\top \theta) x x^\top.$$

This in particular implies

- *For any z,*
$$\|z\|^2_{\left(\sum_{s=1}^n \dot{\mu}(x_s^\top \theta) x_s x_s^\top\right)^{-1}} \leq \frac{(1+\epsilon)}{n} \|z\|^2_{\left(\sum_{x \in \mathcal{X}} \lambda_x \dot{\mu}(x^\top \theta) x x^\top\right)^{-1}}$$
- $$\lambda_{\min}\left(\sum_{s=1}^{n} \dot{\mu}(x_s^\top \theta) x_s x_s^\top\right) \geq \frac{n}{1+\epsilon} \lambda_{\min}\left(\sum_{x \in \mathcal{X}} \lambda_x \dot{\mu}(x^\top \theta) x x^\top\right)$$

Lemma 5.5. *Let $D_\delta(z)$ be any valid confidence width for the estimated model $\hat{\theta}$ on arm z with a probability of at least $1-\delta$, i.e., with a probability of at least $1-\delta$, we have*

$$z^\top \hat{\theta} - D_\delta(z) < z^\top \theta < z^\top \hat{\theta} + D_\delta(z).$$

Then the following condition

$$D_\delta(z) < \left|z^\top \hat{\theta}\right|$$

is a sufficient condition to guarantee that the reward model $\hat{\theta}$ makes a correct classification on arm z with probability at least $1-\delta$, i.e.,

$$\text{sign}(z^\top \hat{\theta}) = \text{sign}(z^\top \theta).$$

Proof. Suppose that $D_\delta(z) < \left|z^\top \hat{\theta}\right|$. We now show that $z^\top \hat{\theta}$ and $z^\top \theta$ share the same sign.

Case 1: $z^\top \hat{\theta} > 0$.
Since $D_\delta(z) < z^\top \hat{\theta}$, we have

$$z^\top \hat{\theta} - D_\delta(z) > 0.$$

Hence, with a probability of at least $1-\delta$,

$$z^\top\theta > z^\top\hat{\theta} - D_\delta(z) > 0,$$

so $z^\top\theta > 0$ and the classification is correct.

Case 2: $z^\top\hat{\theta} < 0$.
Here $-z^\top\hat{\theta} > 0$ and $D_\delta(z) < -z^\top\hat{\theta}$, so

$$z^\top\hat{\theta} + D_\delta(z) < 0.$$

Hence, with a probability of at least $1-\delta$,

$$z^\top\theta < z^\top\hat{\theta} + D_\delta(z) < 0,$$

so $z^\top\theta < 0$ and again the classification is correct. In both cases the sign of $z^\top\hat{\theta}$ matches that of $z^\top\theta$. □

A.3 Proof of Lemma 4.1

Lemma 4.1. *The greedy objective in algorithm 3 can be equivalently written as*

$$\max_z \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\} = \max_z \left\{D_{\delta,t}(z) - \left|\hat{\theta}_t^\top z\right|\right\}.$$

Proof.

$$\begin{aligned}
&\max_z \min\left\{-\mathsf{LCB}_t(z), \mathsf{UCB}_t(z)\right\} \\
&= \max_z \min\left\{-(\hat{\theta}_t^\top z - D_{\delta,t}(z)), \hat{\theta}_t^\top z + D_{\delta,t}(z)\right\} \\
&= \max_z \min\left\{-\hat{\theta}_t^\top z + D_{\delta,t}(z), \hat{\theta}_t^\top z + D_{\delta,t}(z)\right\} \\
&= \max_z \left\{\min\left\{-\hat{\theta}_t^\top z, \hat{\theta}_t^\top z\right\} + D_{\delta,t}(z)\right\} \\
&= \max_z \left\{D_{\delta,t}(z) - \left|\hat{\theta}_t^\top z\right|\right\}.
\end{aligned}$$

□

A.4 Proof of Theorem 5

Theorem 5. *For the canonical instance $z_i = e_i$ for $i = 1, \ldots, d$ and $\theta^* = (1, \ldots, 1, \epsilon)^\top$ for $\epsilon \in (0, 1/4)$, the greedy selection objective $\arg\max_{i\in[d]} \|e_i\|_{H_t'^{-1}}$ has label complexity scaling as $\tilde{O}\left(\frac{d}{\epsilon^2}\right)$, while with algorithm 3, the label complexity scales as $\tilde{O}\left(d + \frac{1}{\epsilon^2}\right)$.*

Proof. We begin by establishing a sufficient condition for the algorithms to stop. Our analysis considers the canonical instance where the arm set $\mathcal{Z}$ consists of the standard basis vectors $\{e_1, \ldots, e_d\}$. In this setting, the two algorithms greedily pull arms based on their respective selection rules:

- **G-optimal based method Das et al. (2025):** At each round t, select arm $\arg\max_{i\in[d]} \|e_i\|_{H_t'^{-1}}$, where $H_t' = \sum_{s=1}^t \dot{\mu}(z_s^\top\hat{\theta}_t) z_s z_s^\top$.
- **Our method:** At each round t, select arm $\arg\max_{i\in[d]}\{D_{\delta,t}(e_i) - |e_i^\top\hat{\theta}_t|\}$.

From lemma 5.5, the suggested stopping rule is

$$D_{\delta,t}(e_i) - |e_i^\top\hat{\theta}_t| < 0, \quad \text{for all } i \in [d], \tag{18}$$

where $D_{\delta,t}(e_i)$ is the confidence width for arm e_i at round t. Since this rule involves the empirical estimate $\hat{\theta}_t$, we first derive a simpler sufficient condition that depends only on the true parameter θ^*.

We show that the condition

$$D_{\delta,t}(e_i) < \frac{1}{2}|e_i^\top \theta^*|, \quad \text{for all } i \in [d], \tag{19}$$

is sufficient for the stopping rule in eq. (18) to be satisfied. By definition of the confidence width, we know that with probability at least $1-\delta$, $|e_i^\top \hat{\theta}_t - e_i^\top \theta^*| \le D_{\delta,t}(e_i)$. Using the triangle inequality, we have:

$$\begin{aligned} |e_i^\top \hat{\theta}_t| &\ge |e_i^\top \theta^*| - |e_i^\top \hat{\theta}_t - e_i^\top \theta^*| \\ &\ge |e_i^\top \theta^*| - D_{\delta,t}(e_i). \end{aligned}$$

If we assume eq. (19) holds, then $|e_i^\top \theta^*| > 2D_{\delta,t}(e_i)$. Substituting this into the inequality above gives:

$$|e_i^\top \hat{\theta}_t| > 2D_{\delta,t}(e_i) - D_{\delta,t}(e_i) = D_{\delta,t}(e_i).$$

This directly implies that $D_{\delta,t}(e_i) - |e_i^\top \hat{\theta}_t| < 0$, satisfying the stopping rule.

Now, we use this sufficient condition to determine the number of pulls t_i required for each arm e_i. For the canonical instance, lemma 5.6 allows us to express the condition in eq. (19) as:

$$2.4\sqrt{\frac{1}{\dot{\mu}(\theta_i^*)t_i}}\sqrt{\log\frac{2(2+|\mathcal{Z}|)}{\delta}} < \frac{1}{2}|\theta_i^*|, \quad \text{for all } i \in [d]. \tag{20}$$

Solving for t_i, we find that each arm i must be pulled a number of times satisfying:

$$t_i > \frac{\mathcal{C}}{\dot{\mu}(\theta_i^*)(\theta_i^*)^2}\log\frac{2(2+|\mathcal{Z}|)}{\delta},$$

where $\mathcal{C}$ is a numerical constant.

We can now analyze the total sample complexity for each algorithm.

G-optimal Method As established in lemma 5.7, the G-optimal method behaves as a round-robin for the canonical instance, meaning $t_i \approx T/d$ for all i after T rounds. To satisfy the stopping condition for all arms, the number of pulls for the arm with the smallest $|\theta_d^*| = \varepsilon$ must meet the requirement. This leads to a total sample complexity T of:

$$T \approx d \cdot t_d > d \cdot \frac{\mathcal{C}}{\varepsilon^2}\log\left(\frac{|\mathcal{Z}|}{\delta}\right) \implies T = \tilde{O}\left(\frac{d}{\varepsilon^2}\right).$$

The algorithm continues to pull all arms in a round-robin fashion until the condition is met for every arm, including those that may have satisfied it much earlier.

Our Method A fundamental difference in our algorithm is that its selection rule, $\arg\max_i\{D_{\delta,t}(e_i) - |e_i^\top \hat{\theta}_t|\}$, is directly aligned with the stopping rule. Once an arm e_i satisfies the stopping condition $D_{\delta,t}(e_i) - |e_i^\top \hat{\theta}_t| < 0$, its selection metric becomes negative and it will not be pulled again, as the algorithm prioritizes arms with larger, positive values. The algorithm thus focuses its budget on arms that have not yet met the condition.

The total sample complexity is the sum of the required pulls for each arm. Let $|\theta_d^*| = \varepsilon$. The complexity is dominated by the pulls required for this arm, while other arms require fewer pulls.

$$T = \sum_{i=1}^{d} t_i = \sum_{i=1}^{d} \tilde{O}\left(\frac{1}{(\theta_i^*)^2}\right) = \tilde{O}\left(\sum_{i=1}^{d-1}\frac{1}{(\theta_i^*)^2} + \frac{1}{\varepsilon^2}\right) = \tilde{O}\left(d + \frac{1}{\varepsilon^2}\right).$$

This demonstrates a significant improvement over the G-optimal method when ε is small. □

Lemma 5.6. *For the canonical instance where $z_i = e_i$ for $i = 1, \cdots, d$, the confidence interval in theorem 1 can reduce to*

$$\left|\hat{\theta}_{t,i} - \theta_i^*\right| < 2.4\sqrt{\frac{1}{\dot{\mu}(\theta_i^*)t_i}}\sqrt{\log\frac{2(2+|\mathcal{Z}|)}{\delta}}.$$

Proof. The result is straightforward by noting the canonical instance and the confidence interval in theorem 1 □

Lemma 5.7. *For the canonical instance where $z_i = e_i$ for $i = 1, \cdots, d$, and $\theta^* = (1, \ldots, 1, \epsilon)^\top$, the algorithm in Das et al. (2025) pulls arms almost uniformly, i.e.,*

$$t_1 = \cdots = t_{d-1} \quad \textit{and} \quad t_d < t_1 < 2t_d.$$

Proof. For the algorithm of Das et al. (2025), the selection rule is

$$\arg\max_i \|z_i\|_{H'(\{z_s, s\in[t]\},\theta^*)^{-1}} = \arg\max_i \frac{1}{\dot\mu\big(\theta_i^*\big)t_i} = \arg\min_i \dot\mu\big(\theta_i^*\big)t_i.$$

Note that we have

$$\dot\mu\big(\theta_i^*\big) = \dot\mu(1) \text{ for } i = 1, \cdots, d-1, \text{ and } \dot\mu\big(\theta_d^*\big) = \dot\mu(\epsilon).$$

Also note that for a small ϵ, e.g., $\epsilon \in (0, 1/4)$,

$$\dot\mu(1) < \dot\mu(\epsilon) < 2\dot\mu(1).$$

This implies that the algorithm will pull the arms almost uniformly. Specifically, if t_i is the number of times arm i is pulled, then we have

$$t_1 = \cdots = t_{d-1} \quad \text{and} \quad t_d < t_1 < 2t_d.$$

□

B LOWER BOUND

We provide an information-theoretic lower bound for this problem.

Theorem 6. *Let $\mathcal{Z} = \{z_1, \ldots, z_n\}$ and consider the logistic link $\psi(u) = 1/(1 + e^{-u})$. Suppose $\min_{z\in\mathcal{Z}} |z^\top\theta^*| > 0$, and let $0 < \delta \le e^{-1}$. Define the sign-changing alternative set*

$$\mathcal{C}_{\theta^*,\mathcal{Z}} := \Big\{\theta \in \mathbb{R}^d : \exists\, z \in \mathcal{Z},\ \operatorname{sign}(z^\top\theta^*) \neq \operatorname{sign}(\theta^\top z)\Big\}.$$

For any strategy that is uniformly δ-PAC over logistic parameters with no zero arm margins, the expected label complexity satisfies

$$\mathbb{E}_{\theta^*}[\tau] \ge \log\frac{1}{2.4\,\delta} \inf_{\lambda\in\Delta(\mathcal{Z})} \sup_{\theta\in\mathcal{C}_{\theta^*,\mathcal{Z}}} \frac{1}{\sum_{i=1}^n \lambda_i\, D_{\mathrm{KL}}\big(P_{\theta^*,i}\,\|\,P_{\theta,i}\big)},$$

where $P_{\theta,i} = \mathrm{Bern}\big(\psi(z_i^\top\theta)\big)$. Alternative parameters are unrestricted apart from the sign change; zero-margin boundary points are interpreted by continuity.

Proof. Let T_i be the number of queries to arm z_i, so $\tau = \sum_{i=1}^n T_i$. If $\mathbb{E}_{\theta^*}[\tau] = \infty$, the result is immediate. Otherwise, the transportation lemma of Kaufmann et al. (2016), applied to the event that the returned classifier has the same arm signs as θ^*, gives

$$\sum_{i=1}^n \mathbb{E}_{\theta^*}[T_i]\, D_{\mathrm{KL}}\big(P_{\theta^*,i}\,\|\,P_{\theta,i}\big) \ \ge\ \log\frac{1}{2.4\,\delta}$$

for every sign-changing alternative θ with no zero arm margins. Since $\mathcal{Z}$ is finite and the true margins are nonzero, the boundary points in $\mathcal{C}_{\theta^*,\mathcal{Z}}$ can be approached by such alternatives. Continuity of the Bernoulli KL divergence therefore extends the inequality to the whole set.

The preceding inequality shows that $(\mathbb{E}_{\theta^*}[T_1], \ldots, \mathbb{E}_{\theta^*}[T_n])$ is a feasible solution to the following optimization problem,

$$\begin{aligned} &\min_{\mathbf{t}\in[0,\infty)^n} \sum_{i=1}^n t_i \\ \text{s.t.}\quad &\inf_{\theta\in\mathcal{C}_{\theta^*,\mathcal{Z}}} \sum_{i=1}^n t_i\, D_{\mathrm{KL}}\big(P_{\theta^*,i}\,\|\,P_{\theta,i}\big) \ \ge\ \log\frac{1}{2.4\,\delta}. \end{aligned}$$

Let $\mathbf{t}^*$ be an optimal solution, which exists since the feasible set is closed and has a nonempty compact objective sublevel set. Since the expected-count vector is feasible, we have

$$\sum_{i=1}^{n} \mathbb{E}_{\theta^*}[T_i] \;\geq\; \sum_{i=1}^{n} t_i^*.$$

Normalizing $\mathbf{t}^*$ by its positive total gives

$$\inf_{\theta \in \mathcal{C}_{\theta^*,\mathcal{Z}}} \sum_{i=1}^{n} \frac{t_i^*}{\sum_{j=1}^{n} t_j^*} D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big) \;\geq\; \frac{\log\big(1/(2.4\delta)\big)}{\sum_{j=1}^{n} t_j^*} \;\geq\; \frac{\log\big(1/(2.4\delta)\big)}{\sum_{j=1}^{n} \mathbb{E}_{\theta^*}[T_j]}.$$

Since $\mathbf{t}^* / \sum_{j=1}^{n} t_j^*$ belongs to $\Delta(\mathcal{Z})$, we conclude

$$\sup_{\lambda \in \Delta(\mathcal{Z})} \inf_{\theta \in \mathcal{C}_{\theta^*,\mathcal{Z}}} \sum_{i=1}^{n} \lambda_i \, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big) \;\geq\; \frac{\log\big(1/(2.4\delta)\big)}{\sum_{i=1}^{n} \mathbb{E}_{\theta^*}[T_i]}.$$

Taking reciprocals and rearranging the terms, we obtain

$$\mathbb{E}_{\theta^*}[\tau] = \sum_{i=1}^{n} \mathbb{E}_{\theta^*}[T_i] \geq \log \frac{1}{2.4\,\delta} \inf_{\lambda \in \Delta(\mathcal{Z})} \sup_{\theta \in \mathcal{C}_{\theta^*,\mathcal{Z}}} \frac{1}{\sum_{i=1}^{n} \lambda_i \, D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta,i}\big)}.$$

□

The instance-dependent quantity ρ^* is analogous to complexity terms that arise in some best-arm identification problems with linear structures (Jun et al., 2021; Fiez et al., 2019; Zhao et al., 2024). This term captures the problem's complexity through the geometry of the arm set and the true parameter θ^*. For linear bandits, Fiez et al. (2019); Zhao et al. (2024) have shown that such a term characterizes the optimal sample complexity via a minimax lower bound. Here we make the comparison precise for the logistic link. Its Fisher information depends on the true logits, so the comparison with the KL-based lower bound must account for this curvature. Under the bounded-logit conditions in theorem 4, the following proof gives this comparison with an explicit dependence on the logit upper bound.

B.1 Proof of Theorem 4

Theorem 4 (Leading-term comparison for bounded logits)**.** *Consider the logistic link $\psi(u) = 1/(1+e^{-u})$. Let $\mathcal{Z} = \{z_1, \ldots, z_n\}$ span $\mathbb{R}^d$, and suppose*

$$\max_{z \in \mathcal{Z}} |z^\top \theta^*| \leq B, \qquad \kappa_B := \psi(B)(1-\psi(B)).$$

Then

$$T^\star_{\mathrm{KL}} \geq 8 \kappa_B \rho^*.$$

Proof. Fix sampling weights λ, and write $H_\lambda := H(\lambda, \theta^*)$. First suppose that H_λ has an inverse, and define

$$V_\lambda := \sum_{i=1}^{n} \lambda_i z_i z_i^\top, \qquad R_\lambda := \max_{z \in \mathcal{Z}} \frac{z^\top H_\lambda^{-1} z}{|z^\top \theta^*|^2}.$$

When H_λ has no inverse, we set $R_\lambda = +\infty$ and use the same convention in the objective defining ρ^*. Thus $\rho^* = \inf_\lambda R_\lambda$.

Choose an arm z_λ with the largest ratio in the definition of R_λ. Let $g = z_\lambda^\top \theta^*$ be its true logit and write $q = z_\lambda^\top H_\lambda^{-1} z_\lambda$, so $R_\lambda = q/g^2$. For $\eta > 0$, we change the parameter by v_η as follows:

$$v_\eta := -(1+\eta)\frac{g}{q} H_\lambda^{-1} z_\lambda, \qquad \theta_\eta := \theta^* + v_\eta.$$

This gives $z_\lambda^\top \theta_\eta = -\eta g$, so the preference for z_λ is reversed and $\theta_\eta \in \mathcal{C}_{\theta^*,\mathcal{Z}}$. Each new logit is its nonzero true logit plus a constant times $1+\eta$, so it can be zero for at most one value of η. There

are finitely many arms, so we can choose positive η arbitrarily close to zero without any ties. The changed logits do not need to satisfy the bound B.

To bound the KL divergence, define $A(u) = \log(1 + e^u)$. Its derivatives satisfy $A'(u) = \psi(u)$ and $A''(u) = \psi'(u) \le 1/4$. By Taylor's theorem, for any two logits $a, b \in \mathbb{R}$,

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(\psi(a)) \,\|\, \mathrm{Bern}(\psi(b))\big) = A(b) - A(a) - A'(a)(b-a) \le \frac{(a-b)^2}{8}.$$

The bound on the true logits gives $\psi'(z_i^\top \theta^*) \ge \kappa_B$ for every arm. Hence $H_\lambda \succeq \kappa_B V_\lambda$, so $v_\eta^\top V_\lambda v_\eta \le v_\eta^\top H_\lambda v_\eta / \kappa_B$. The logit of arm z_i changes by $z_i^\top v_\eta$. Applying the KL bound above to each arm and multiplying by its sampling weight λ_i, we obtain

$$\begin{aligned} I_\lambda(\theta_\eta) &:= \sum_{i=1}^{n} \lambda_i D_{\mathrm{KL}}\big(P_{\theta^*,i} \,\|\, P_{\theta_\eta,i}\big) \\ &\le \frac{1}{8} v_\eta^\top V_\lambda v_\eta \le \frac{1}{8\kappa_B} v_\eta^\top H_\lambda v_\eta = \frac{(1+\eta)^2}{8\kappa_B} \frac{g^2}{q} = \frac{(1+\eta)^2}{8\kappa_B R_\lambda}. \end{aligned}$$

For the last two equalities, we substitute v_η and use $R_\lambda = q/g^2$. Since $\theta_\eta \in \mathcal{C}_{\theta^*,\mathcal{Z}}$, the supremum below is at least $1/I_\lambda(\theta_\eta)$. Taking the reciprocal of the bound above and letting η approach zero gives

$$\sup_{\theta \in \mathcal{C}_{\theta^*,\mathcal{Z}}} \frac{1}{I_\lambda(\theta)} \ge 8\kappa_B R_\lambda.$$

Finally, suppose H_λ has no inverse. This means that some changes to the parameter leave the logits of all arms with positive sampling weight unchanged. More precisely, there is a nonzero vector u such that $z_i^\top u = 0$ whenever $\lambda_i > 0$. For any t, replacing θ^* by $\theta^* + tu$ therefore leaves the label probabilities of all these arms unchanged.

However, the full arm set $\mathcal{Z}$ spans $\mathbb{R}^d$, so some arm must have $z_i^\top u \ne 0$. Its new logit is $z_i^\top \theta^* + t z_i^\top u$, and we can choose t to reverse its sign. We can also avoid ties by excluding the finitely many values of t that make any arm's logit zero. The resulting parameter changes a preference but gives exactly the same label probabilities on every arm with positive sampling weight. Its total KL divergence under λ is therefore zero. With $1/0 = +\infty$, the left-hand side above is infinite, as is R_λ by our convention.

We have proved the comparison for every choice of sampling weights. Taking the infimum over λ gives

$$T^\star_{\mathrm{KL}} \ge 8\kappa_B \inf_\lambda R_\lambda = 8\kappa_B \rho^*.$$

□

C Experiments

We evaluate both of our proposed algorithms against five baselines using real-world and synthetic preference data. We evaluate the greedy algorithm in algorithm 3 on the Anthropic helpful and harmless (Anthropic HH) dataset (Bai et al., 2022), using classification accuracy on held-out preference pairs and best-of-N response selection with two reward-model judges. We also evaluate the experimental design algorithm in algorithm 2 on synthetic instances with known true preferences. Specifically, we consider the following five algorithms, including existing works on preference learning and established active learning methods adapted to this setting, for comparison with our proposed algorithms:

- **Random Sampling:** Arms are selected uniformly at random at each round.
- **Uncertainty Sampling:** It selects arms closest to the decision boundary based on the current estimate, i.e., those with the smallest absolute estimated reward: $\arg\min_z \left| z^\top \hat{\theta}_t \right|$. A simple and widely adopted heuristic in active learning (Lewis, 1995; Tong & Koller, 2001; Mussmann & Liang, 2018; Raj & Bach, 2022). It has proven effective in various settings.

- **Selective Sampling:** A popular active learning method for the streaming setting where the learner decides whether to query a human labeler as each arm arrives. We adapt it to our setting by iterating through the arms in a shuffled order and selecting the first arm z whose lower confidence bound is below a threshold, e.g., $\mathsf{LCB}_t(z) < 0.1$. Note that all true rewards $z^\top \theta^*$ are positive in our setup, since we construct each z as the difference between the chosen and rejected responses.

- **APO (Das et al., 2025):** It pulls the arm that is most uncertain under the current model, measured by the width of the confidence ellipsoid in its direction: $\arg\max_z \|z\|_{H'(\{z_s, s\in[t]\}, \hat{\theta}_t)^{-1}}$. The same approach is also used in Li et al. (2025).

- **D-optimal design:** This common experimental design criterion, used in Liu et al. (2024); Kveton et al. (2025), selects the arm that greedily maximizes the determinant of the information matrix, i.e., $\arg\max_z \det\Big(\sum_{s=1}^{t} \dot{\mu}(z_s^\top \hat{\theta}_t) z_s z_s^\top\Big)$. Note that Kveton et al. (2025) propose an equivalent formulation that avoids computing the determinant directly.

For the Anthropic HH evaluations, we use Gemma 2 (`gemma-2-2b-it`) (Team, 2024) to generate text embeddings. We clarify that our setup does not involve full parameter fine-tuning, rather, it is equivalent to using a frozen backbone while the learning process focuses on the reward head, as suggested by Kumar et al. (2022). The comparison focuses on the performance of different preference data selection methods. Freezing the LLM backbone and isolating the linear reward head enables us to rigorously bound the sample complexity of our active learning algorithm, ensuring that these theoretical bounds are not confounded by the unstable feature distortion often associated with underlying network updates.

C.1 Preference Classification Accuracy

Following practical considerations for real-world applications, we implement all algorithms in a batched setting, as a fully sequential, round-by-round implementation is often infeasible due to labeling latency (Scheid et al., 2024). In each epoch, we sample a batch of 50 arms to query, collect their labels, and then update the model. All reported results are averaged over 50 independent runs. Note that any given dataset represents a single realization of the underlying label noise, meaning the chosen and rejected labels for each pair are fixed. The randomness in our experimental results, therefore, comes from the random train-test split performed at the start of each independent run.

We report the classification accuracy of the trained reward models on the Anthropic HH test set in fig. 3. Our proposed method consistently outperforms the baselines. One interesting observation from our experiments is that methods like D-optimal design perform well early on, but then struggle to keep improving near the end. This is expected because D-optimal design is non-adaptive here and does not use any of the label information it collects. In contrast, Uncertainty Sampling is adaptive, so it eventually reaches performance comparable to ours, but it tends to lag at the beginning. Our method achieves the best of both behaviors. In fact, this is reflected in eq. (9): one term matches the signal used by Uncertainty Sampling, while the other mirrors the criterion behind D-optimal design.

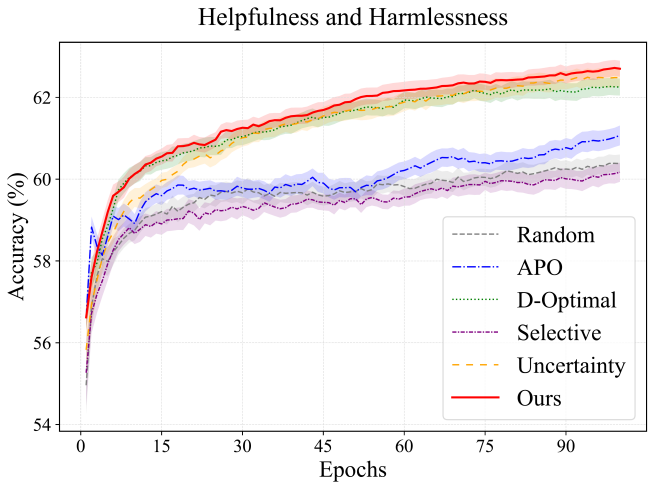


Figure 3: Preference classification accuracy on Anthropic HH, averaged over 50 independent runs.

C.2 Best-of-N Response Selection

We also evaluate whether the trained reward models select better responses from a set of candidates for the same prompt. We reserve 1000 held-out prompts that are disjoint from the selection pool of preference pairs. For each of 50 random seeds, we redraw an 80/20 split of this pool and train the reward head using batches of 50 queries.

For each held-out prompt x, we independently sample $N = 8$ responses from `Qwen2.5-7B-Instruct`, with temperature 0.8 and top-p = 0.95. Each candidate pool is generated once and reused across all methods, seeds, and labeling budgets. Let $h(x, a)$ denote the embedding produced by the frozen Gemma 2 backbone for prompt x and response a, so the trained reward head assigns score $\hat{\theta}^\top h(x, a)$ to each response. It selects

$$\hat{a}(x) = \arg \max_{a \in \{a_1, \ldots, a_N\}} \hat{\theta}^\top h(x, a).$$

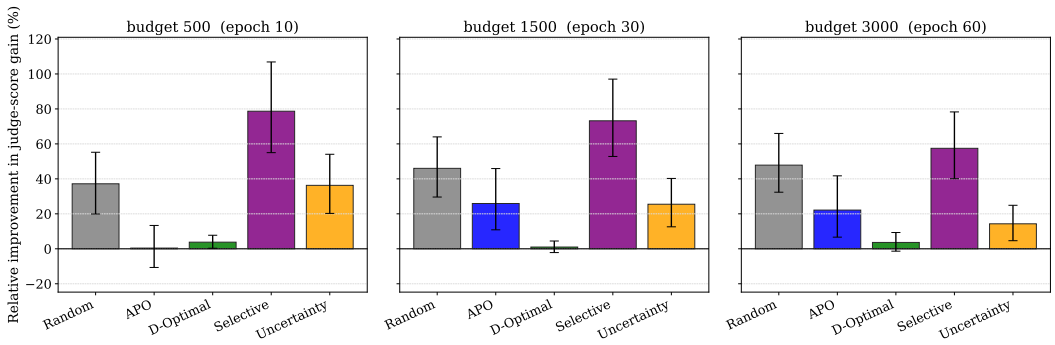
We score the same responses with two open source reward models as judges: `Skywork-Reward-V2-Llama-3.1-8B` (Skywork 8B) and `INF-ORM-Llama3.1-70B` (INF-ORM 70B, int8 quantization was applied to the INF-ORM 70B judge). For each judge, scores are computed once for each candidate and reused in every comparison.

We compute the relative improvement for each judge. Let $\bar{r}_{\text{ours}}$ and $\bar{r}_{\text{base}}$ be the mean judge scores of the responses selected by our method and a baseline at the same labeling budget, averaged over the 1000 held-out prompts and 50 runs. Let $\bar{r}_{\text{rand}}$ be the corresponding mean score when a response is chosen uniformly at random from each candidate pool; for each prompt, this is the average judge score of its N candidates. We report

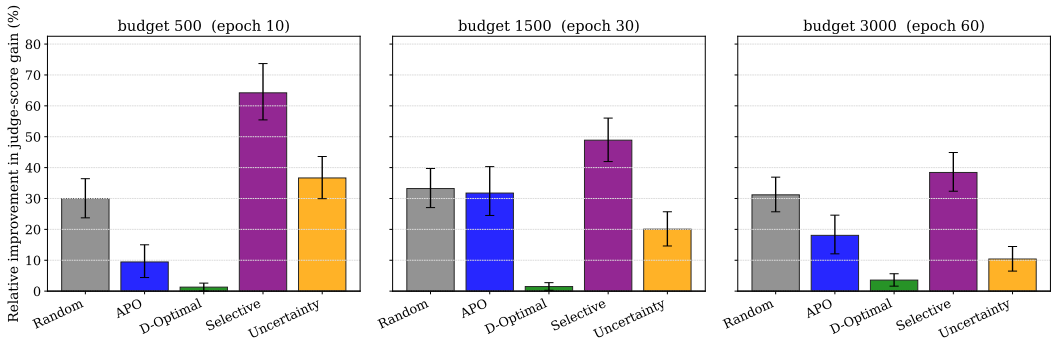
$$\text{Relative improvement (\%)} = 100 \, \frac{\bar{r}_{\text{ours}} - \bar{r}_{\text{base}}}{\bar{r}_{\text{base}} - \bar{r}_{\text{rand}}}.$$

This measures the improvement over a baseline relative to what that baseline gains from using a reward model instead of choosing a response at random. Random response selection is distinct from the Random Sampling baseline, which trains a reward model on randomly queried preference pairs.

The relative improvement of our method over each baseline is reported in fig. 4. For both judges, the results favor our method against all five baselines across the three labeling budgets. The gains are largest over Selective Sampling and smaller over D-optimal design. Some confidence intervals overlap zero, but the overall trend is consistent across the two judges.



(a) Skywork 8B judge.



(b) INF-ORM 70B judge.

Figure 4: Best-of-N response selection on Anthropic HH with $N = 8$, evaluated by two reward-model judges. Each bar shows the percentage improvement of our method over the indicated baseline in judge-score gain above a uniformly random response from the same candidate pool. Scores are averaged over 1000 held-out prompts and 50 runs. Positive values favor our method.

C.3 Synthetic Preference Data

We also evaluate the experimental design algorithm in algorithm 2 against the same five baselines on synthetic preference data.

We produce a pool of $n = 100$ arms in $d = 15$ dimensions. In each run, we sample a random unit vector θ^* and construct the arms as

$$z = m\theta^* + \sqrt{1 - m^2}\, v, \qquad \|v\| = 1, \quad v^\top \theta^* = 0,$$

so that $z^\top \theta^* = m$, where m may be positive or negative. The pool contains two groups:

- **60 easy arms:** Unit norm and large absolute margins (0.61–1.00), with random positive or negative signs. Each arm has its own randomly sampled direction v orthogonal to θ^*.
- **40 hard arms:** Random positive or negative signs, with all arms sharing one orthogonal direction up to sign, v or $-v$. We use pre-scaling absolute margins 0.30–0.60 and multiply each entire vector by 0.15, giving each arm a norm of 0.15 and absolute margins of approximately 0.045–0.09.

We use a practical version of algorithm 2. It uses $5d$ warm-up labels. The precision starts at $\epsilon_1 = 2 \max_z |z^\top \hat{\theta}_0|$ and is multiplied by 0.7 when the round's target is met. The target for each arm is $\max\{|z^\top \hat{\theta}|, \epsilon_\ell\}$. We fit the regularized logistic MLE using all labels collected so far, and account for them when computing the next design and round size.

At each budget t, we report the percentage of arms satisfying $\widetilde{D}_t(z) < |z^\top \hat{\theta}_t|$, where

$$\widehat{H}_t = \sum_{s=1}^{t} \psi'(z_s^\top \hat{\theta}_t) z_s z_s^\top + \alpha I_d, \qquad \widetilde{D}_t(z) = \sqrt{\log \frac{2n(2+t)}{\delta}} \, \|z\|_{\widehat{H}_t^{-1}}.$$

Here $\hat{\theta}_t$ is the regularized MLE on all collected labels. The criterion checks whether an arm's estimated sign is separated from zero by more than its estimated uncertainty. The plot refers to these arms as certified.

In fig. 5, our method outperforms all five baselines under the empirical criterion at both 10048 and 20000 labels. Uncertainty Sampling concentrates on the small-margin hard arms, which share an orthogonal direction. Other directions therefore receive few queries, leaving the estimated confidence intervals for many easy arms wide. Consequently, its performance is significantly worse than that of the other methods.

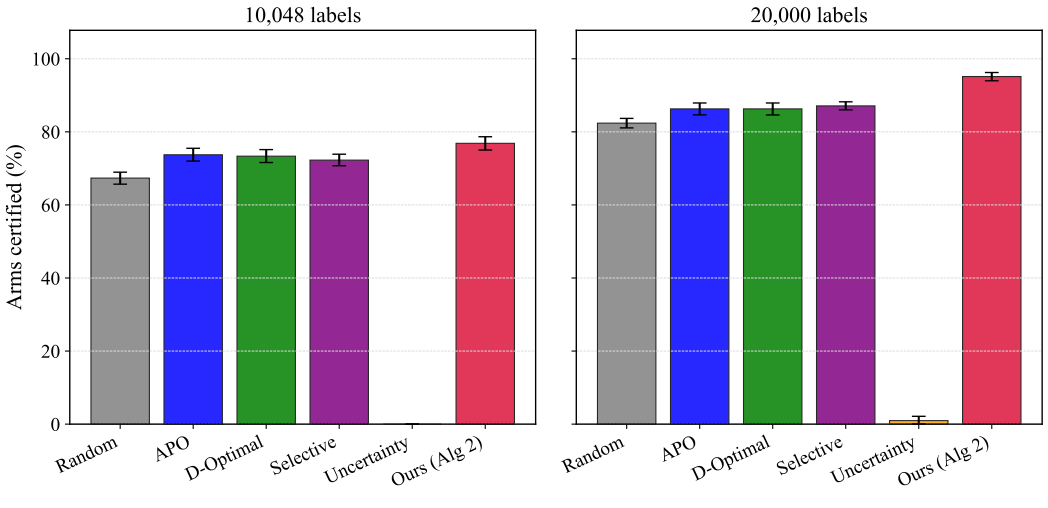


Figure 5: Synthetic preference learning with $n = 100$ arms and $d = 15$. Bars show the percentage of arms satisfying the empirical confidence-width criterion, averaged over 50 runs.