Preference is More than Comparisons: Rethinking Dueling Bandits with Augmented Human Feedback

Shengbo Wang[1], **Hong Sun**[1], **Ke Li**[2]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2]Department of Computer Science, University of Exeter, EX4 4RN, Exeter, UK
shnbo.wang@foxmail.com; k.li@exeter.ac.uk

Abstract

Interactive preference elicitation (IPE) aims to substantially reduce human effort while acquiring human preferences in wide personalization systems. Dueling bandit (DB) algorithms enable optimal decision-making in IPE building on pairwise comparisons. However, they remain inefficient when human feedback is sparse. Existing methods address sparsity by heavily relying on parametric reward models, whose rigid assumptions are vulnerable to misspecification. In contrast, we explore an alternative perspective based on feedback augmentation, and introduce critical improvements to the model-free DB framework. Specifically, we introduce augmented confidence bounds to integrate augmented human feedback under generalized concentration properties, and analyze the multi-factored performance trade-off via regret analysis. Our prototype algorithm achieves competitive performance across several IPE benchmarks, including recommendation, multi-objective optimization, and response optimization for large language models, demonstrating the potential of our approach for provably efficient IPE in broader applications.

Code — https://github.com/COLA-Laboratory/IPEA-HF
Extended version — https://arxiv.org/abs/2511.09047

1 Introduction

In personalization systems ranging from recommendation (Austin et al. 2024) and multi-objective optimization (Huang, Wang, and Li 2024) to large language models (LLMs) (Rafailov et al. 2023), acquiring user preferences is essential but often incurs great human effort. Interactive preference elicitation (IPE) has the potential to substantially reduce this burden by selectively querying users through iterative strategies (Xiong et al. 2024). Building on pairwise comparisons, the dueling bandit (DB) framework has evolved into a strong theoretical foundation for IPE over the past decade (Yue et al. 2012; Zoghi et al. 2014; Saha 2021). However, even under optimal strategies, the DB framework struggles to maintain efficiency in the presence of sparse feedback, leading to waning research attention and limited practical adoption in recent years. As a result, while personalization systems continue to demand data-efficient interaction, achieving *provably efficient* IPE remains a pressing challenge.

Sparse feedback in practice often reveals a deeper insight: human preferences are more than isolated pairwise comparisons. They are also shaped by cues such as contextual information and latent dependencies (Sun et al. 2024). The independent treatment of human feedback in the DB framework may be a key source of its inefficiency. Augmenting human feedback with these cues could pave the way for a more effective DB framework that better facilitates IPE. Preliminary efforts in this direction have primarily relied on parametric preference models, most notably the Bradley–Terry (BT) model (Bradley and Terry 1952). However, while such model assumptions are convenient and analytically tractable, they have been increasingly criticized for their susceptibility to model misspecification (Heckel et al. 2016; Verma et al. 2025) and suboptimal performance in the presence of non-transitive preferences (Munos et al. 2024). Alternative methods that exploit contextual information and dependencies do exist but have yet to form a widely applicable framework (Sui et al. 2017; Xiao et al. 2025). Furthermore, several fundamental questions about the DB framework with augmented human feedback remain unaddressed. In this work, we focus on the following three research questions (RQs):

RQ1: *How can the DB framework integrate and interpret the role of augmented human feedback, while reducing reliance on rigid, predefined model assumptions?*

In the broader framework of contextual bandits (Lattimore and Szepesvári 2020), side information can be exploited through various approaches, including structured reward estimation, candidate partitioning, or similarity-based methods. Algorithms based on parametric BT models, for instance, typically fall under the reward estimation category. As discussed earlier, reward estimation risks model misspecification and struggles to accommodate non-transitive preferences. Alternatively, candidate partitioning offers favorable regret bounds that scale with the number of partition groups (Huang, Wang, and Li 2024). However, it relies on a strong assumption that candidates can be cleanly divided into distinguishable subsets, an assumption often unverifiable in real-world settings. In this work, we explore similarity-based methods to integrate augmented human feedback. While closely related to other approaches (Slivkins 2014), such methods remain largely underexplored with respect to their efficiency in IPE (Sui et al. 2017). Our methodology is motivated by extending the concentration results of the context-free DB framework

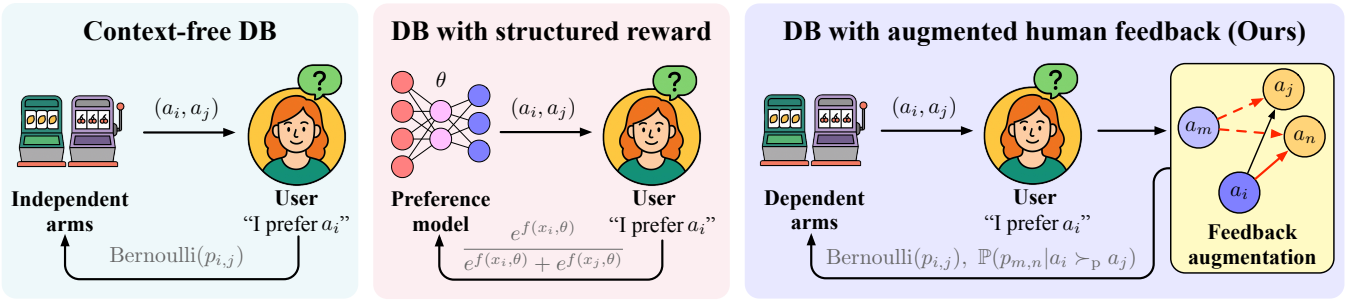


Figure 1: Comparison of three DB approaches: context-free DB, structured reward estimation, and DB with augmented feedback.

(Zoghi et al. 2014). We conclude that augmenting human feedback serves as a mechanism for uncertainty calibration, offering a unifying perspective across various DB approaches. A comparison of DB approaches is presented in Figure 1.

RQ2: *Does the incorporation of augmented human feedback consistently improve the efficiency of DB algorithms, or can it sometimes introduce performance degradation?*

Existing studies on augmented human feedback, typically grounded in parametric or distinguishable assumptions, have shown considerable promise in improving sample efficiency. However, they often struggle to explain common pitfalls such as overfitting (Azar et al. 2024) and optimization inefficiencies (Razin et al. 2025). By discarding these predefined assumptions, we open the black box of feedback augmentation, enabling a deeper investigation into how augmented human feedback influences decision-making within the DB framework. Following the upper confidence bound (UCB) method (Srinivas et al. 2010), we provide theoretical analysis of sample complexity and regret bounds, revealing an explicit trade-off governed by multiple factors. These findings provide practical guidance on when to leverage, how to calibrate, and what to explore regarding human feedback augmentation, positioning the DB framework as a stronger foundation for subsequent IPE algorithm design and evaluation.

RQ3: *Can the DB framework be extended beyond pairwise comparisons to incorporate richer forms of human feedback, thereby establishing a more general foundation for IPE?*

Most IPE approaches, including augmented variants of the DB framework, are constrained to receiving human preferences exclusively through pairwise comparisons. However, incorporating richer forms of human feedback, such as feature-level comparisons (Austin et al. 2024), expert demonstrations (Sun et al. 2024), and explanations of human choice (Ghazimatin et al. 2021), has been shown to positively influence decision-making efficiency. As diverse feedback sources become increasingly available, the standard DB framework remains limited in its ability to directly utilize such information. Fortunately, our DB framework offers the potential to integrate heterogeneous forms of human feedback beyond pairwise comparisons. This is achieved by treating all feedback as a unified signals that characterize contextual similarity and dependencies among candidates. Ultimately, this would provide a more flexible foundation for IPE tasks.

We highlight the following key advances in this work:

- We develop a model-free DB framework with a generalized concentration property, enabling the integration of augmented human feedback and the quantification of its influence. Our framework connects human feedback to the calibration of confidence bounds, aligning in mechanism with a broad class of DB approaches.
- We study the sample complexity and cumulative regret bounds in the proposed DB framework, revealing an explicit trade-off between the amount of augmented human feedback and the strength of contextual dependencies. In addition, our analysis includes partition-based approaches (Huang, Wang, and Li 2024) as a special case.
- We present prototype designs that incorporate a similarity-based graph structure and an auxiliary annotation process for capturing contextual dependencies. Our algorithms demonstrate competitive performance across several benchmarks, including recommendation, multi-objective optimization, and LLM response optimization.

2 Problem Formulation

We consider an IPE task involving K ($K \geq 2$) candidates, where the goal is to identify the candidate that best aligns with the user's preference, typically inferred through pairwise comparisons. Mathematically, the task can be formulated as a K-arms DB problem, where the set of arms is indexed by $\mathcal{A} = \{1, 2, \ldots, K\}$. When contextual information is available, each arm a_i is associated with a context vector $x_i \in \mathcal{X}$. In the t-th round, the user is asked to evaluate a pair of arms (a_i, a_j), where $i, j \in \mathcal{A}$ and $i \neq j$. The user should decide, based on her/his preferences, whether a_i is better, worse, or equivalent to a_j, denoted as $a_i \succ_{\mathrm{p}} a_j$, $a_i \prec_{\mathrm{p}} a_j$, or $a_i \simeq_{\mathrm{p}} a_j$. For stochastic preferences, we assume a fixed preference matrix $\mathbf{P} = [p_{i,j}]_{K\times K}$, where $p_{i,j}$ denotes the probability that arm a_i is preferred over arm a_j (Yue et al. 2012). Without loss of generality, we have $p_{i,j} + p_{j,i} = 1$ and $p_{i,i} = 0.5$. An arm a_i is said to be superior to the a_j if $p_{i,j} > 0.5$. We also denote the best candidate as the winner a_*, defined by criteria such as the Condorcet or Copeland winners (Urvoy et al. 2013), among all arms.

The *context-free* DB framework maintains a winning matrix $\mathbf{B} = [b_{i,j}]_{K\times K}$ to record the pairwise comparison labels, where $b_{i,j}$ denotes the number of times when arm a_i is preferred over arm a_j. The estimated preference probability with mean $\tilde{p}_{i,j} = \frac{b_{i,j}}{b_{i,j}+b_{j,i}}$, and the upper confidence bound (UCB) $\mathbf{U} = [\tilde{u}_{i,j}]_{K\times K}$ and lower confidence bound (LCB)

$\mathbf{L} = [\tilde{l}_{i,j}]_{K\times K}$ are given by (Zoghi et al. 2014):

$$\tilde{u}_{i,j} = \tilde{p}_{i,j} + \sqrt{\frac{\alpha \log t}{b_{i,j} + b_{j,i}}},\ \tilde{l}_{i,j} = \tilde{p}_{i,j} - \sqrt{\frac{\alpha \log t}{b_{i,j} + b_{j,i}}}, \quad (1)$$

where $\alpha > 0$ controls the confidence interval, and t is the total number of comparisons so far. Consequently, all arms are treated independently according to their own labels.

To encode contextual information and dependencies (Lattimore and Szepesvári 2020), a common DB approach is based on structured reward estimation. This approach assumes a structured form for $p_{i,j}$, such as the BT model, and a parametric reward function $f(x_i, x_j) = \theta^\top (x_i - x_j)$, with the unknown parameter θ to be determined:

$$\tilde{p}_{i,j}(x_i, x_j) = \frac{1}{1 + \exp(-f(x_i, x_j))}, \quad x_i, x_j \in \mathcal{X}. \quad (2)$$

This formulation transforms the problem of identifying the best arm into one of learning the unknown parameter θ. Despite its favorable sample complexity, both the probability structure and parametric reward assumptions have been called into question (Heckel et al. 2016; Verma et al. 2025; Munos et al. 2024; Azar et al. 2024; Razin et al. 2025).

Another way to leveraging contextual information is through candidate partitioning, such as clustering (Huang, Wang, and Li 2024), merging (Li et al. 2020), or team grouping (Cohen, Schmidt-Kraepelin, and Mansour 2021). These methods partition the K arms into C subsets, where $C < K$ in order to improve sample efficiency. The partitioning process is often based on the assumption that all arms within the winner subset are strictly superior to those in other subsets (Jedor, Perchet, and Louëdec 2019), a property we referred to as *distinguishability*. However, candidate partitioning in real-world IPE may not always guarantee such property. A more robust alternative is to model dependencies through similarity (Sui et al. 2017). Let $s : \mathcal{X} \times \mathcal{X} \to [0, 1]$ denote a similarity function. The special case where $s \in \{0, 1\}$ corresponds to settings satisfying the distinguishable property. Nevertheless, the sample efficiency based on similarity remains unclear. We will explore this gap in the subsequent sections.

In this work, we consider the Condorcet winner (Yue et al. 2012) for both analytical convenience and consistent benchmarking across different DB approaches. Specifically, the probability of a_* satisfies $p_{*,j} > 1/2$, $\forall j \in \mathcal{A} \setminus \{*\}$. If a_i and a_j were chosen for comparison at time t, the instantaneous regret is defined as $r_t = \frac{\Delta_i + \Delta_j}{2}$, where $\Delta_k = p_{*,k} - \frac{1}{2}$, $\forall k \in \mathcal{A}$. Accordingly, in the context of DB, the objective of IPE is to minimize the *cumulative regret up to time* T, defined as $R(T) = \sum_{t=1}^{T} r_t$. The equivalence of regret measures for parametric DB approach is demonstrated in (Saha 2021).

3 Method

In this section, we delineate the DB framework for IPE with augmented human feedback, referred to as `IPEA-HF`. The overall structure is outlined in Algorithm 1, which compromises four key components. It begins by introducing augmented confidence bounds that incorporate augmented human feedback in a model-free setting. This is followed by pair selection criteria, which depend on the specific DB algorithm employed, such as the relative upper confidence bound (RUCB) (Zoghi et al. 2014) or double Thompson sampling (DTS) (Wu and Liu 2016). Upon observing a user response, `IPEA-HF` augments the feedback based on contextual similarity and extracted dependencies through computational algorithm designs and additional annotations.

Algorithm 1: Pseudo-code of `IPEA-HF`

Input: Candidate number K, context set $\mathcal{X}$, $\alpha > 0$
Init: $\mathbf{B} = [0]_{K\times K}$, graph $\mathcal{G}(\mathcal{X}, K)$, $W = [0]$.
repeat
  $\mathbf{U}, \mathbf{L} \leftarrow$ `AugConfidanceBound`$(\mathbf{B}, \mathcal{G}, W, \alpha)$
  Select pairs $(a_i, a_j) \leftarrow$ `DuelingBanditAlgo`$(\mathbf{U}, \mathbf{L})$
  Observe a_{win}, a_{lose} and update $b_{\text{win,lose}} \leftarrow b_{\text{win,lose}} + 1$
  $W \leftarrow$ `DependencyExtract`$(\mathcal{X}, \mathcal{G}, W)$
  $\mathcal{G} \leftarrow$ `FeedbackAug`$(a_{\text{win}}, a_{\text{lose}}, W, \mathcal{G})$
until IPE task finished or budget exhausted

3.1 `AugConfidanceBound`: Integrating Confidence Bounds with Augmented Feedback

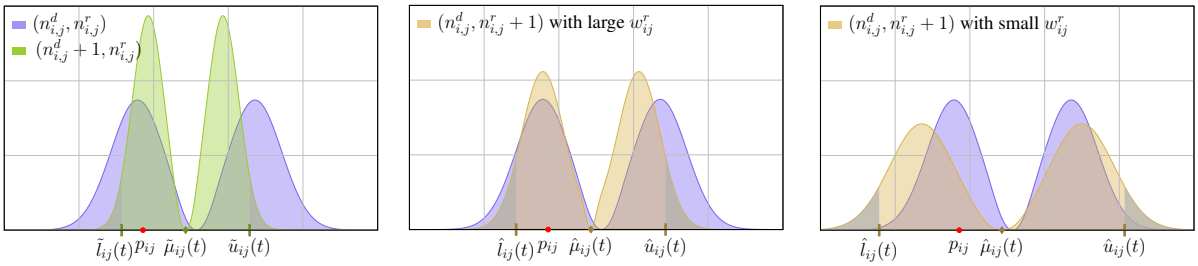
The confidence bounds in the DB framework play a pivotal role in determining which pairs are selected for comparison, and ultimately shape the overall query process. As formalized in equation (1), bounds $\tilde{u}_{i,j}$ and $\tilde{l}_{i,j}$ in the context-free DB approach rely solely on *direct* observations from comparisons between the pair (a_i, a_j), specifically the counts $b_{i,j}$ and $b_{j,i}$. In contrast, we explore whether incorporating additional *related* observations from augmented human feedback can enhance the estimation of confidence bounds. For $i, j \in \mathcal{A}$, let $n^d_{i,j}(t) = b_{i,j} + b_{j,i}$ denote the number of direct comparisons between a_i and a_j up to time t, and let $n^r_{i,j}(t)$ denote the number of related observations inferred from other pairwise comparisons up to time t. We then define the total observation count as $n_{i,j}(t) = n^d_{i,j}(t) + n^r_{i,j}(t)$. The augmented mean becomes $\hat{p}_{i,j} = \frac{1}{\eta n_{i,j}(t)} \left(b_{i,j} + \sum_{k=1}^{n^r_{i,j}} X^k_{i,j} \right)$, and the UCB and LCB take the following form:

$$\hat{u}_{i,j} = \hat{p}_{i,j} + \frac{1}{\eta}\sqrt{\frac{\alpha \ln t}{n_{i,j}(t)}},\ \hat{l}_{i,j} = \hat{p}_{i,j} - \frac{1}{\eta}\sqrt{\frac{\alpha \ln t}{n_{i,j}(t)}}. \quad (3)$$

where $\eta = \left(n^d_{i,j} + \sum_{k=1}^{n^r_{i,j}} w^k_{i,j} \right) / n_{i,j}$ with $w^k_{i,j} \in [0, 1]$ to be determined, and $X^k_{i,j}$ denotes a random variable in $[0, 1]$ determined by related observations. By definition, without augmented feedback, i.e., $n^r_{i,j} = 0$, the augmented confidence bounds reduce to model-free ones in equation (1).

Concentration property We provide the following concentration property of the augmented confidence bounds as a generalized result from the context-free DB framework (Zoghi et al. 2014; Wu and Liu 2016).

Theorem 3.1. *Assume* $X^k_{i,j} \sim \text{Bernoulli}(w^k_{i,j} p_{i,j})$ *and let* $C(\delta) = \left(\frac{(4\alpha-1)K^2}{(2\alpha-1)\delta} \right)^{\frac{1}{2\alpha-1}}$. *Given the preference matrix* $\mathbf{P}$



(a) Direct observation (b) Related observation above threshold (c) Related observation under threshold

Figure 2: Comparison of confidence intervals adding a direct observation or a related observation. The curve shows the derivative of the bound in equation (4), where the shaded area is the probability that the confidence interval fails to contain $p_{i,j}$.

with K arms, then, for any $\alpha > 0.5$ and $\delta \in (0,1)$, we have:

$$P\left(\forall t > C(\delta),\ i,j \in \mathcal{A},\ p_{i,j} \in \left[\hat{l}_{i,j}(t),\ \hat{u}_{i,j}(t)\right]\right) > 1-\delta. \tag{4}$$

The proof is given in Appendix 10.1. This result suggests that additional observations can be leveraged to inform the estimation of confidence bounds in conjunction with the latent dependency weight $w^k_{i,j}$. We will later discuss how to obtain the weights in Section 3.3.

We now study the property of concentration. Consider the case where $n^r_{i,j}(t) = 0$. Given a direct observation, the confidence interval shrinks from $\sqrt{\alpha \ln t / n^d_{i,j}(t)}$ to $\sqrt{\alpha \ln t/(n^d_{i,j}(t)+1)}$. The confidence intervals progressively narrow as the number of direct comparisons increases (Zoghi et al. 2014). By contrast, for a related observation, the confidence bound takes the form $\frac{1}{\eta}\sqrt{\alpha \ln t/(n^d_{i,j}(t)+1)}$. Notably, the weighting term $w^k_{i,j}$ influences the behavior of the confidence bound. When $w^r_{i,j} = 1$, we have $\eta = 1$, and the confidence bound shrinks in the same way as with a direct observation. Conversely, when $w^r_{i,j} = 0$, the bound becomes $\sqrt{\alpha(n^d_{i,j}(t)+1)\ln t / {n^d_{i,j}}^2(t)}$, which is strictly larger than the original bound $\sqrt{\alpha \ln t / n^d_{i,j}(t)}$. Therefore, both direct and relative observations contribute to the *uncertainty calibration* for the confidence bounds, with the effect of relative observations depending on the strengths of dependencies.

Calibration threshold When estimating $p_{i,j}$, a smaller confidence interval is desirable for making more informed decisions. We can quantify the impact of augmented feedback by the ratio of confidence intervals with and without a relative observation:

$$\begin{aligned}\frac{\hat{u}_{i,j}-\hat{l}_{i,j}}{\tilde{u}_{i,j}-\tilde{l}_{i,j}} &= \frac{1}{\eta}\sqrt{\frac{\alpha \ln t}{n^d_{i,j}(t)+1}} \Big/ \sqrt{\frac{\alpha \ln t}{n^d_{i,j}(t)}} \\ &\in \left[\sqrt{1-\frac{1}{n^d_{i,j}(t)+1}},\ \sqrt{1+\frac{1}{n^d_{i,j}(t)}}\right]. \end{aligned} \tag{5}$$

We can derive a threshold condition for identifying *good* feedback augmentation that facilitates uncertainty calibration:

$$w^r_{i,j} > \eta n_{i,j}(t)\left(\sqrt{1+\frac{1}{n_{i.j}(t)}}-1\right). \tag{6}$$

This condition holds for all $n_r \geq 0$. The detailed derivation is provided in Appendix 10.2. The result indicates that the contribution of a relative observation increases monotonically with the strength of $w^r_{i,j}$, becomes equivalent to a direct observation when $w^r_{i,j} = 1$, and diminishes as the number of direct observations increases, as illustrated in Figure 2.

Connections to other DB approaches The augmented confidence bounds naturally encompass partition-based approaches as a special case. Specifically, for all arms within a distinguishable group, related observations correspond to complete dependencies, i.e., $w^r_{i,j} = 1$. By considering only related observations within the same group, the comparison between individual arms effectively reduces to a comparison between groups, mirroring the operations adopted in partition-based approaches. For DB approaches with structured reward, they maintain a pairwise score of the form $u_f(x_i, x_j) = \hat{\theta}(x_i - x_j) + \sigma\|x_i - x_j\|_{V^{-1}}$ where $\hat{\theta}$ is the estimated parameter, $\sigma > 0$, and V is the sum of outer products of compared pairs (Verma et al. 2025; Saha 2021). The second term of $u_f(x_i, x_j)$ represents the confidence bound expressed as a Mahalanobis norm (Mahalanobis and 1936). For a pair (a_i, a_j), more frequent comparisons, i.e., more direct observations in our setting, assign smaller weight to the direction of $x_i - x_j$. Given a related observation that is close in distance to (a_i, a_j), representing stronger dependencies, the weight is further reduced along that direction, indicating a narrowed uncertainty calibration for the score of (a_i, a_j). This mechanism that lowers the direction weight in the presence of high-correlated observations aligns with our proposed method. However, the confidence bound in structured reward approaches remains heuristic and lacks formal guarantees on the concentration property (i.e., $p_{i,j}$ may not be included within the bound with high probability as t increases beyond a certain threshold), making it prone to overfitting due to model misspecification.

Insights on RQ1: *By leveraging the generalized concentration property, our method fundamentally strengthens the model-free DB framework by incorporating related observations into the estimation of winning probabilities and confidence bounds. We identify the key conditions under which contextual correlations enhance uncertainty calibration and establish clear connections to structured and partition-based approaches, highlighting the broader applicability.*

3.2 `DuelingBanditAlgo`: Decision-Making with Augmented Feedback

The estimation of winning probabilities and their confidence bounds is integrated into DB decision-making algorithms to strike a balance between *exploitation* and *exploration*. These algorithms select candidate pairs through a two-stage process:

1. (Exploitation) Select the most promising arm as the first candidate. This is achieved by identifying the arm whose UCBs outperform the largest number of other arms.
2. (Exploration) Select the most competitive arm against the first candidate. This aims to identify the arm likely to beat the first selected arm according to their UCB and LCB.

Existing DB algorithms differ in their selection mechanism, including both deterministic and stochastic strategies, such as RUCB (Zoghi et al. 2014) and DTS (Wu and Liu 2016). Our augmented confidence bounds can be seamlessly integrated into both types. We propose two DB algorithms: `IPEA-RUCB` and `IPEA-DTS` integrated with augmented confidence bounds, with detailed descriptions in Appendix 7.

Sample Efficiency Following the analysis in (Zoghi et al. 2014), we investigate the RUCB variant integrated with augmented confidence bounds. Since our framework incorporates both direct observations (real samples) and related observations (virtual samples), the following result establishes a high-probability bound on the total sample complexity of each suboptimal arm pair.

Theorem 3.2. *Given the setup in Theorem 3.1, and let* $D^w_{i,j} = \frac{4\alpha}{\min_r {w^r_{i,j}}^2 \min\{\Delta_i^2, \Delta_j^2\}}$. *For the* `IPEA-RUCB` *algorithm and any suboptimal pair* $(a_i, a_j) \neq (a_*, a_*)$, $n_{i,j}(t)$ *satisfies*

$$P(\exists t,\ i, j \in \mathcal{A},\ n_{i,j}(t) > C(\delta) \vee D^w_{i,j} \ln t) < \delta. \quad (7)$$

The proof is given in Appendix 10.3. Accordingly, the multi-factored sample complexity reveals a fundamental **trade-off** in integrating augmented feedback. On the positive side, for a pair (a_i, a_j) that utilizes related observations of (a_m, a_n), it holds that $n^d_{i,j}(t) + n^d_{m,n}(t) \leq n_{i,j}(t)$. That is, the number of direct observations of one pair that provide augmented feedback to other pairs is already included in the total sample count of the latter. If more pairs are related, the total sampling count satisfies $n^d_{i,j}(t) + \sum_{m,n} n^d_{m,n}(t) \leq n_{i,j}(t)$. This indicates that incorporating related observations has the potential to improve overall sample efficiency. However, for less dependent observations, e.g., when $\min_r w^r_{i,j} \to 0$, $D^w_{i,j}$ can increase significantly, introducing a large coefficient into the overall sample complexity. According to this trade-off, augmented feedback should be integrated selectively, prioritizing observations with strong contextual dependencies.

Regret Analysis To conduct regret analysis, we assume bidirectional dependency: if a pair (a_i, a_j) utilizes related observations from another pair (a_m, a_n), the latter also utilizes the observations from (a_i, a_j). A direct consequence of this assumption is the soft-clustering, grouping candidates into C subsets. Notably, existing methods based on distinguishability (Li et al. 2020; Cohen, Schmidt-Kraepelin, and Mansour 2021) or arm dependencies (Sui et al. 2017) satisfy this assumption. Furthermore, our method generalizes beyond these settings, as illustrated by the diverse cases in Figure 3.

Theorem 3.3. *Follow the setup in Theorems 3.1 and 3.2 and assume bidirectional dependencies. If we have* C *soft clusters and each cluster contains* K_i *candidates, the cumulated regret bound of* `IPEA-RUCB` *is* $\mathcal{O}\left(\frac{1}{\min_{i,j,r} w^r_{i,j}} \hat{K}^2 \log T\right)$, *where* $\hat{K} = \max\{C, K_1, \ldots, K_C\}$.

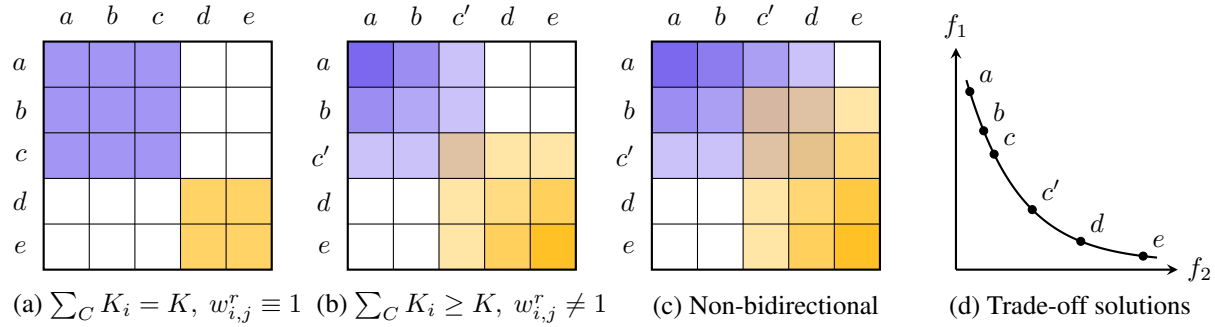
The proof is provided in Appendix 10.4. This result subsumes the partition-based DB approach as a special case. For example, the cluster-based DB method achieves a regret bound of $\mathcal{O}\left(C^2 \log T\right)$ (Huang, Wang, and Li 2024), where the analysis is conducted at the cluster level under the assumption of distinguishability ($w^r_{i,j} \equiv 1$). In contrast, our result operates at the candidate level and explicitly accounts for varying degree of dependency, such that the candidate count in each subset is also reflected in the effective parameter $\hat{K}$. In conclusion, our regret bound generalizes the partition-based approach (Huang, Wang, and Li 2024) and offers promising support for DB algorithms that incorporate correlations or dependent arms (Sui et al. 2017).

Insights on RQ2: *Our analysis of sample complexity and regret bounds reveals a fundamental trade-off in incorporating augmented human feedback into DB algorithms. Theoretical results consistently show that related observations with strong dependencies can substantially reduce the number of required interactions, thereby improving efficiency. This finding aligns with existing conclusions and offers generalized support to model-free, dependency-aware DB approaches.*

3.3 `DependencyExtract & FeedbackAug`: Computational Design

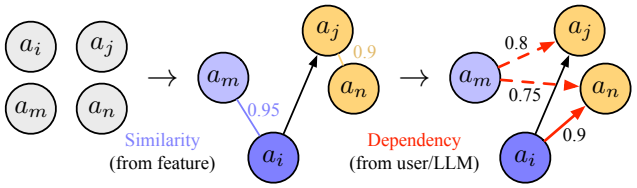
Our algorithm consists of two key components: extracting dependencies and augmenting human feedback. Our analysis (Theorem 3.1) characterizes latent dependencies as the joint distribution of direct and indirect observations, formalized as $\text{Bernoulli}(w^k_{i,j} p_{i,j})$. The weighting term $w^k_{i,j}$ captures the conditional dependency between pairs.

To estimate $w^r_{i,j}$, we follow a two-step procedure, as illustrated in Figure 4. First, we construct a prior structural skeleton by building a similarity graph over the candidates, where the edges are determined by similarity distances in the context space $\mathcal{X}$. We filter out dissimilar pairs using graph partitioning or soft clustering, thereby concentrating on the most meaningful dependencies. As a result, each candidate



(a) $\sum_C K_i = K,\ w^r_{i,j} \equiv 1$ (b) $\sum_C K_i \geq K,\ w^r_{i,j} \neq 1$ (c) Non-bidirectional (d) Trade-off solutions

Figure 3: Illustrative cases for bidirectional dependency. (a) Candidate partitioning. (b) Dependent arms with symmetric correlations. (c) General case without bidirectional dependency. (d) Two-objective trade-off solutions whose dependencies are determined by their distances. Given $[a, b, c, d, e]$, the candidates can be safely grouped into two subsets. When c shifts to c', the dependencies weaken, making soft clustering a more suitable choice.



(a) Candidates (b) Direct feedback (c) Augmented feedback

Figure 4: Illustration of computational design of dependency extraction and feedback augmentation.

is associated with one or more groups, depending on the employed partitioning method. At each round of pairwise comparisons, we observe human feedback on a pair, e.g., a_i is preferred over a_j. Then, we should determine the dependencies. Through additional annotations, we selectively query user or LLMs to annotate conditional dependencies over local pairwise relations, and incorporate related observations with high dependency scores into the DB loop. Specifically, we iterate over and combine all pairs from the groups associated with a_i and a_j. When a_i and a_j belong to the same group, no feedback augmentation is performed.

While the annotation process for dependency estimation provides valuable signals for feedback augmentation, it may introduce errors, particularly when relying on simulated annotators such as LLMs. The robustness of our framework stems from the fact that the influence of a related observation is strictly weaker than that of a direct observation, and further diminishes as the number of direct observations increases (see equation (5)). In addition, the impact of biased estimation in DB has been analyzed in (Yi, Kang, and Li 2024). As interactions accumulate, the estimated dependencies are either refined or discarded, enabling the algorithm to recover from initial noisy annotations over time.

Richer Forms of Human Feedback In the modelling of conditional dependencies, multiple forms of feedback can be leveraged. Beyond pairwise comparisons and contextual similarities, our method also incorporates reasoning signals derived from LLMs or domain experts on latent contextual dependencies. Additional types of data, such as human explanations and rationales (Ghazimatin et al. 2021), can also be readily integrated, all contributing to the discovery and refinement of informative conditional dependencies. This enables the incorporation of richer feedback signals within a principled framework, supported by the theoretical foundations of our DB framework. As more dependency data becomes available, probabilistic graphical models can be employed to capture joint distributions and conditional structures, thereby further denoising sparse human feedback and estimating missing relationships.

4 Empirical Study

We evaluate our algorithms (`IPEA-RUCB` and `IPEA-DTS`) across a diverse set of benchmarks:

- **Item Recommendation.** We consider the Sushi dataset containing full rankings over 10 types of sushi collected from 5,000 customer orders (Kamishima and Akaho 2010), and a Car Preference dataset with full rankings over 10 cars provided by 60 U.S. users via Amazon Mechanical Turk (Abbasnejad et al. 2013).
- **Multi-Objective Optimization.** We adopt the *a posteriori* setting of preference-based evolutionary multi-objective optimization (PBEMO) (Huang, Wang, and Li 2024), considering the celebrated synthetic test problems (DTLZ2 and DTLZ7) with different landscapes (Deb et al. 2005).
- **LLM Response Optimization**. We follow the experimental settings in (Verma et al. 2025; Dwaracherla et al. 2024) to conduct *active exploration* given a pool of responses for each prompt sampled from Anthropic Helpfulness and Harmlessness (H-H) dataset (Bai et al. 2022).

We consider five state-of-the-art DB methods spanning context-free and parametric frameworks.

- **Context-free algorithms.** We include two prominent DB algorithms: Relative Upper Confidence Bound (`RUCB`) (Zoghi et al. 2014) and Double Thompson Sampling (`DTS`) (Wu and Liu 2016).
- **Parametric algorithms.** We select three state-of-the-art methods based on the BT models: Maximum Informative

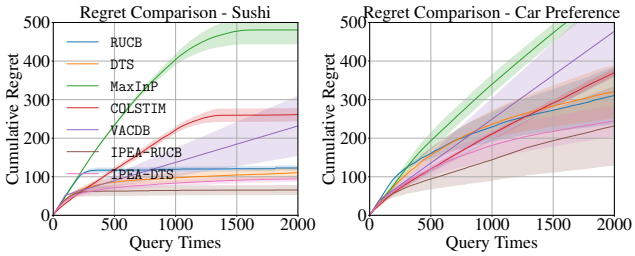


Figure 5: Regret trajectories on Sushi and Car Preference.

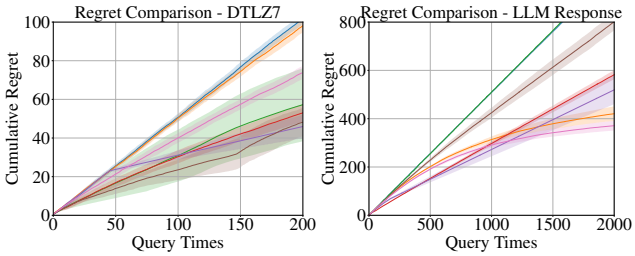


Figure 6: Regret on DTLZ7 and LLM Exploration.

Pair (MaxInP) (Saha 2021), COLSTIM (Bengs, Saha, and Hüllermeier 2022), and VACDB (Di et al. 2024).

Detailed experimental settings are given in Appendix 8.

Item Recommendation We simulate 2,000 interaction rounds to reflect the aggregated preferences of a user group. The cumulative regret of all algorithms is shown in Figure 5. Under a moderate number of candidate items, context-free DB algorithms consistently outperform parametric baselines. Building on this foundation, our proposed algorithms achieve further improvements in sample efficiency. The inferior performance of parametric methods is largely due to model misspecification, which is exacerbated when handling real-world, mixed-type feature spaces. Moreover, some parametric algorithms suffer from premature convergence. For example, VACDB exhibits a steadily increasing regret curve, indicating that it fails to maintain effective exploration over time.

Multi-Objective Optimization We simulate 200 interaction rounds to evaluate algorithm performance under sparse human feedback. The cumulative regret trajectories are shown in the left panel of Figure 6. As expected, the standard DB algorithms RUCB and DTS exhibit poor efficiency when facing a large candidate set (100^2 pairs) and a limited interaction budget. Although parametric DB algorithms demonstrate relatively better performance, their limitations are evident. Regarding the query frequency as given in Figure 7, both COLSTIM and VACDB repeatedly select a small subset of pairs, reflecting strong exploitation and a lack of sufficient exploration. In contrast, our proposed algorithms achieve a better trade-off between regret minimization and query diversity. Notably, IPEA-RUCB performs more favorably than IPEA-DTS under these conditions, suggesting that deterministic strategies may be more robust to sparse feedback.

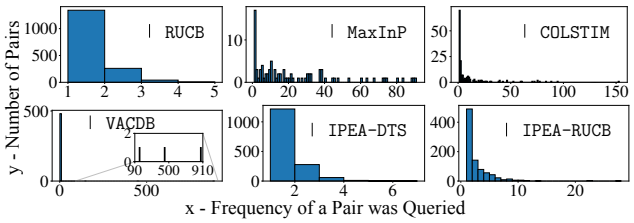


Figure 7: Analysis of Query frequency on the DTLZ7 task. In the histogram, $(2, 1000)$ denotes 1,000 unique pairs each queried twice. A broader spread at low frequencies indicates greater exploration.

Response Optimization The regret trajectories are shown in the right panel of Figure 6. Most algorithms exhibit a degree of inefficiency, with the exception of DTS and IPEA-DTS, both employ a stochastic strategy. This aligns with the empirical findings of (Dwaracherla et al. 2024), where DTS was shown to be the most effective algorithm for active exploration for LLMs. Notably, by integrating augmented human feedback, our algorithms consistently outperform their context-free counterparts, RUCB and DTS. In contrast, algorithms based on parametric models require significantly more computational resource, due to the high dimensionality of the embedded feature space (768 dimensions from the MPNet (Song et al. 2020)). As a result, without a dedicated context-aware mechanism (Verma et al. 2025), the effectiveness using parametric DB for response optimization remains questionable.

In response optimization, our algorithms leverage not only related observations from the same prompt but also inter-prompt observations with high contextual similarity and dependency. This pilot demonstration shows the effectiveness of feedback augmentation beyond pairwise comparisons.

Insights on RQ3: *By treating human feedback as unified signals characterizing contextual similarity and dependencies, our prototype computation designs are readily extensible to incorporate richer forms of feedback. In our experiments, we integrate contextual similarity, LLM-powered dependency annotations, and inter-prompt observations, each serving as an additional form of human feedback. This expands the applicability of the DB framework, offering a promising pathway for adapting to diverse feedback modalities in future IPE tasks.*

5 Conclusions

We introduced a novel DB framework that integrates augmented human feedback to enable provably efficient IPE. By analyzing the concentration properties and performance trade-offs, we demonstrated the effectiveness of feedback augmentation and established connections to a broad class of existing DB approaches. In addition, we proposed prototype computational designs that incorporate richer forms of human feedback, and showed competitive performance across diverse benchmarks. This work offers a principled, extensible foundation for the development of personalization systems.

Acknowledgments

We sincerely thank all the reviewers for their encouraging and constructive feedback. This work was supported by the UKRI Future Leaders Fellowship under Grant MR/S017062/1 and MR/X011135/1; in part by NSFC under Grant 62376056 and 62076056; in part by the Royal Society Faraday Discovery Fellowship (FDF/S2/251014), BBSRC Transformative Research Technologies (UKRI1875), Royal Society International Exchanges Award (IES/R3/243136), Kan Tong Po Fellowship (KTP/R1/231017); and the Amazon Research Award and Alan Turing Fellowship.

References

Abbasnejad, E.; Sanner, S.; Bonilla, E. V.; and Poupart, P. 2013. Learning Community-Based Preferences via Dirichlet Process Mixtures of Gaussian Processes. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 1213–1219. IJCAI/AAAI.

Austin, D. E.; Korikov, A.; Toroghi, A.; and Sanner, S. 2024. Bayesian Optimization with LLM-Based Acquisition Functions for Natural Language Preference Elicitation. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, 74–83.

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A General Theoretical Paradigm to Understand Learning from Human Preferences. In *International Conference on Artificial Intelligence and Statistics*, volume 238, 4447–4455. PMLR.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; and et al., T. H. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.

Bengs, V.; Saha, A.; and Hüllermeier, E. 2022. Stochastic Contextual Dueling Bandits under Linear Stochastic Transitivity Models. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 1764–1786.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4): 324–345.

Cohen, L.; Schmidt-Kraepelin, U.; and Mansour, Y. 2021. Dueling Bandits with Team Comparisons. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, 20633–20644.

Deb, K.; Thiele, L.; Laumanns, M.; and Zitzler, E. 2005. Scalable Test Problems for Evolutionary Multiobjective Optimization. In *Evolutionary Multiobjective Optimization*, Advanced Information and Knowledge Processing, 105–145. Springer.

Di, Q.; Jin, T.; Wu, Y.; Zhao, H.; Farnoud, F.; and Gu, Q. 2024. Variance-aware Regret Bounds for Stochastic Contextual Dueling Bandits. In *The Twelfth International Conference on Learning Representations*.

Ding, B.; Qin, C.; Zhao, R.; Luo, T.; Li, X.; Chen, G.; Xia, W.; Hu, J.; Luu, A. T.; and Joty, S. 2024. Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges. In *Findings of the Association for Computational Linguistics, ACL 2024*, 1679–1705.

Dudík, M.; Hofmann, K.; Schapire, R. E.; Slivkins, A.; and Zoghi, M. 2015. Contextual Dueling Bandits. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, 563–587.

Dwaracherla, V.; Asghari, S. M.; Hao, B.; and Roy, B. V. 2024. Efficient Exploration for LLMs. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Ghazimatin, A.; Pramanik, S.; Roy, R. S.; and Weikum, G. 2021. ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Models. In *WWW '21: The Web Conference 2021*, 3850–3860. ACM / IW3C2.

González, J.; Dai, Z.; Damianou, A.; and Lawrence, N. D. 2017. Preferential Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1282–1291. PMLR.

Heckel, R.; Shah, N. B.; Ramchandran, K.; and Wainwright, M. J. 2016. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*.

Huang, T.; Wang, S.; and Li, K. 2024. Direct Preference-Based Evolutionary Multi-Objective Optimization with Dueling Bandits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jamieson, K. G.; Katariya, S.; Deshpande, A.; and Nowak, R. D. 2015. Sparse Dueling Bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org.

Jedor, M.; Perchet, V.; and Louëdec, J. 2019. Categorized Bandits. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 14399–14409.

Kamishima, T.; and Akaho, S. 2010. Nantonac collaborative filtering: A model-based approach. In *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010*, 273–276. ACM.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Li, C.; Markov, I.; de Rijke, M.; and Zoghi, M. 2020. MergeDTS: A Method for Effective Large-Scale Online Ranker Evaluation. *ACM Trans. Inf. Syst.*, 38(4): 40:1–40:28.

Li, X.; Zhao, H.; and Gu, Q. 2024. Feel-Good Thompson Sampling for Contextual Dueling Bandits. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Lin, X.; Dai, Z.; Verma, A.; Ng, S.-K.; Jaillet, P.; and Low, B. K. H. 2024. Prompt Optimization with Human Feedback. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Mahalanobis; and (1936), P. 2018. On the Generalised Distance in Statistics. *Sankhya A*, 80: 1 – 7.

Munos, R.; Valko, M.; Calandriello, D.; Azar, M. G.; Rowland, M.; Guo, Z. D.; Tang, Y.; Geist, M.; Mesnard, T.; and Côme Fiegel, e. a. 2024. Nash Learning from Human Feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7): 3580–3599.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Razin, N.; Wang, Z.; Strauss, H.; Wei, S.; Lee, J. D.; and Arora, S. 2025. What Makes a Reward Model a Good Teacher? An Optimization Perspective. *CoRR*, abs/2503.15477.

Saha, A. 2021. Optimal Algorithms for Stochastic Contextual Preference Bandits. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, 30050–30062.

Saha, A.; and Gaillard, P. 2022. Versatile Dueling Bandits: Best-of-both World Analyses for Learning from Relative Preferences. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 19011–19026.

Slivkins, A. 2014. Contextual bandits with similarity information. *J. Mach. Learn. Res.*, 15(1): 2533–2568.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*.

Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M. W. 2010. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, 1015–1022. Omnipress.

Sui, Y.; Zhuang, V.; Burdick, J. W.; and Yue, Y. 2017. Multi-dueling Bandits with Dependent Arms. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*.

Sun, H.; Pouplin, T.; Astorga, N.; Liu, T.; and van der Schaar, M. 2024. Improving LLM Generation with Inverse and Forward Alignment: Reward Modeling, Prompting, Fine-Tuning, and Inference-Time Optimization. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Urvoy, T.; Clerot, F.; Féraud, R.; and Naamane, S. 2013. Generic exploration and K-armed voting bandits. In *ICML'13: Proc. of the 30th International Conference on Machine Learning*, 91–99. PMLR.

Verma, A.; Dai, Z.; Lin, X.; Jaillet, P.; and Low, B. K. H. 2025. Neural Dueling Bandits: Preference-Based Optimization with Human Feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2025*. OpenReview.net.

Wang, S.; and Li, K. 2024. Constrained Bayesian Optimization under Partial Observations: Balanced Improvements and Provable Convergence. In *Thirty-Eighth AAAI Conference on Artificial Intelligence 2024*, 15607–15615. AAAI Press.

Wu, H.; and Liu, X. 2016. Double Thompson Sampling for Dueling Bandits. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, 649–657.

Xiao, T.; Ge, Z.; Sanghavi, S.; Wang, T.; Katz-Samuels, J.; Versage, M.; Cui, Q.; and Chilimbi, T. 2025. InfoPO: On Mutual Information Maximization for Large Language Model Alignment. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xiong, W.; Dong, H.; Ye, C.; Wang, Z.; Zhong, H.; Ji, H.; Jiang, N.; and Zhang, T. 2024. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-constraint. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Yang, X.; Lu, J.; and Yu, E. 2025. Walking the Tightrope: Disentangling Beneficial and Detrimental Drifts in Non-Stationary Custom-Tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yi, B.; Kang, Y.; and Li, Y. 2024. Biased Dueling Bandits with Stochastic Delayed Feedback. *Transactions on Machine Learning Research*.

Yu, E.; Lu, J.; Yang, X.; Zhang, G.; and Fang, Z. 2025. Learning Robust Spectral Dynamics for Temporal Domain Generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yue, Y.; Broder, J.; Kleinberg, R.; and Joachims, T. 2012. The K-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5): 1538–1556.

Zhao, C.; Yu, T.; Xie, Z.; and Li, S. 2022. Knowledge-aware Conversational Preference Elicitation with Bandit Feedback. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 483–492. ACM.

Zoghi, M.; Karnin, Z. S.; Whiteson, S.; and de Rijke, M. 2015. Copeland Dueling Bandits. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 307–315.

Zoghi, M.; Whiteson, S.; Munos, R.; and de Rijke, M. 2014. Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32, 10–18.