CONTEXT-AWARE PREFERENCE BANDITS WITH HETEROGENEOUS HUMAN FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

In context-aware preference learning, an online learner infers user preferences from sequential feedback. We study this problem under the contextual bandit framework, where each arm is associated with an unknown feature vector. However, context dictates not only the optimal arm but also alters the evaluation scale itself, in that user feedback across scenarios is inherently uncalibrated and highly heterogeneous. To capture these effects, we model the linear utility of each arm via a context-dependent affine calibration, parameterized by a multiplicative scale and an additive offset. We propose a phased elimination algorithm that combines pilot-estimation sample splitting with a one-step Newton correction. For pointwise rating feedback, we group contexts using score binning and learn each group separately, achieving a near-optimal $\tilde{\mathcal{O}}(\sqrt{T} + T^{\frac{1}{1+\rho}})$ regret, where ρ parameterizes the tail decay rate of the score bounds. For pairwise comparison feedback, the additive offset cancels out exactly, yielding a near-optimal $\tilde{\mathcal{O}}(\sqrt{T})$ regret for any ρ. We further establish lower bounds for both feedback protocols, matching the common estimation term in our upper bounds. Experiments on synthetic data and benchmark datasets validate our theoretical findings.

1 INTRODUCTION

Preference learning uses feedback to identify preferred user actions (Bengs et al., 2021; Christiano et al., 2017). As a core technology for language model alignment and adaptive agents (Stiennon et al., 2020; Ouyang et al., 2022; Liang et al., 2026), it is especially vital in domains like writing assistance and personalized recommendation (Li et al., 2010), where optimal choices depend heavily on task requirements and user contexts. Contextual bandits model this sequential decision problem: at each round, a learner observes the context of its interaction with the user, chooses an action representing a strategy for generating a response, and uses the resulting feedback to improve later choices (Chu et al., 2011; Saha & Krishnamurthy, 2022). In human-facing applications, feedback may take several forms, including a pointwise rating of one candidate and a pairwise comparison between two candidates (Wang et al., 2024; 2025). Therefore, to understand when and how these feedback mechanisms effectively support learning, it is essential to consider both the quality of the decision-making process and the information conveyed during each observation.

Users' evaluation standards may vary with the context in which feedback is given. A user may tolerate imperfections in an early draft, scrutinize a recommendation for an important decision, or apply stricter standards to messages containing confidential information. For example, a user may approve two financial summaries containing the same confidential information for internal review but reject both for external sharing, while preferring the clearer summary in either setting (Nissenbaum, 2004; Shao et al., 2024). Context may also affect the user's sensitivity to quality differences, making the same differences between the summaries easy to distinguish in some settings but difficult to discern in others. Learning preferences from heterogeneous human feedback therefore requires accounting for both the user's evaluation baseline and sensitivity.

To capture these variations in evaluation standards, we study context-aware preference bandits with K arms, d-dimensional context representations, and a known affine calibration shared across arms. In this setting, each arm represents a different response strategy and is characterized by an unknown parameter vector in $\mathbb{R}^d$, which determines its linear utility in the current context. Within our model, the calibration represents feedback sensitivity through a positive multiplicative scale and the

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.

evaluation baseline through an additive offset. We allow both the scale and the offset to be unbounded, since some users may be highly sensitive to quality differences or apply extreme acceptance standards in certain contexts. Furthermore, recognizing that users may carry evaluation habits from earlier contexts into the current one, we allow the calibration to depend on the entire context history. We measure regret in calibrated utility gaps, ensuring that differences critical to decision quality are preserved even when the observed user feedback becomes nearly indistinguishable under a logistic link (Zhang et al., 2016; Bradley & Terry, 1952).

In this work, we provide theoretical guarantees for regret minimization in context-aware preference bandits with unbounded calibration, addressing both pointwise ratings and pairwise comparisons. A natural baseline would concatenate all arm parameters into a single dK-dimensional vector and apply EXP4.P for ratings or Sparring EXP4.P for comparisons over a finite policy set (Beygelzimer et al., 2011; Dudík et al., 2015). However, this straightforward approach still faces two major challenges under our model. First, because of the unbounded calibration, ratings or comparisons on extreme rounds become nearly deterministic and reveal little about the arm parameters, even though the underlying utility gaps remain large. Applying this baseline directly allows a few such rounds with extreme calibration to dominate the overall regret. Second, the regret of this approach may depend heavily on the size of d or K. Because the concatenated parameter vector has dimension dK and its norm scales with $\sqrt{K}$, any policy set covering this space must grow accordingly, yielding regret that depends linearly or even superlinearly on K. To overcome these hurdles, we propose BINPILOT, a unified phased elimination framework designed for calibrated ratings and comparisons that maintains a separate learner for each bin. Our main contributions are summarized as follows.

- **Novel Algorithmic Framework (BINPILOT):** We group interaction rounds into bins based on the range containing their current scores, employing an adaptive startup mechanism to control initialization costs in rarely visited bins. Within each active bin, we construct a preliminary *pilot* estimate, refine it via a one-step Newton correction on independent samples, and center matrix regularization at the current pilot to form tight confidence bounds for arm elimination.
- **Pointwise Rating and Pairwise Comparison:** For pointwise ratings, under empirical exponential tail conditions on sensitivity and offset, we establish a near-optimal regret bound of $\widetilde{O}\big(\sqrt{T} + T^{1/(1+\rho_{\mathrm{pt}})}\big)$. Here $\rho_{\mathrm{pt}} > 0$ reflects the tail decay rates relative to context and parameter norm bounds. Notably, whenever $\rho_{\mathrm{pt}} \geq 1$, this guarantee simplifies to the optimal $\widetilde{O}(\sqrt{T})$ rate. For pairwise comparisons, we prove a near-optimal regret bound of $\widetilde{O}(\sqrt{T})$ under any positive sensitivity tail rate. For pairwise comparisons, we prove a near-optimal regret bound of $\widetilde{O}(\sqrt{T})$ under any positive sensitivity tail rate The initialization overhead grows as $O(T^{1/(2+\rho_{\mathrm{pr}})})$, which is strictly sub-$\sqrt{T}$. This indicates that pairwise comparisons are more robust to extreme calibration than pointwise ratings for large T, as the shared offset cancels out in comparisons.
- **Fine-Grained Regret:** We establish an optimal regret $\widetilde{O}(\sqrt{dKT})$ when no calibration is present. Furthermore, we show pointwise admits a regret $\widetilde{O}\big(\sqrt{dKT} + (dK)^{\frac{5\rho_{\mathrm{pt}}}{2(1+\rho_{\mathrm{pt}})}} T^{\frac{1}{1+\rho_{\mathrm{pt}}}}\big)$ and pairwise admits a regret $\widetilde{O}\big(\sqrt{dKT} + d^{\frac{5\rho_{\mathrm{pr}}+2}{2(2+\rho_{\mathrm{pr}})}} K^{\frac{5\rho_{\mathrm{pr}}+4}{2(2+\rho_{\mathrm{pr}})}} T^{\frac{1}{2+\rho_{\mathrm{pr}}}}\big)$, which highlights the respective advantages of the two feedback protocols.

Section 4 empirically validates our theoretical insights across both controlled and real-world settings. In controlled synthetic experiments, we isolate the effect of the offset tail: slower decay increases pointwise regret while pairwise regret remains nearly unaffected, directly illustrating the cancellation of shared offsets in comparisons. Furthermore, on PrivacyLens contexts (Shao et al., 2024), an online Newton learner adapted to our calibrated formulation consistently achieves lower regret than an uncalibrated baseline, demonstrating the practical value of accounting for context-dependent evaluation standards.

Comparison with related work. Contextual preference bandits traditionally focus on sequential pairwise comparisons (Dudík et al., 2015; Saha, 2021), whereas Ji et al. (2023) analyze ratings and preferences strictly offline, and Verma et al. (2025) study neural dueling bandits with extensions to online binary feedback. In contrast, we unify both online protocols under a known, context-dependent affine calibration. Crucially, our shared offset operates before the logistic link and can reduce

feedback information through link saturation even when fully known—a mechanism distinct from additive reward confounding (Krishnamurthy et al., 2018) and comparison-label corruption (Di et al., 2025). Furthermore, while GLM-UCB (Filippi et al., 2010) measures regret in expected rewards, our objective targets calibrated utility gaps, which remain large even when feedback probabilities become nearly indistinguishable; linear utility regret under uncalibrated binary feedback is also studied by Zhang et al. (2016). To control this objective, we build on curvature-sensitive logistic estimation (Faury et al., 2020; Jun et al., 2021; Lee et al., 2024a;b) and the sample-splitting, one-step estimation framework of Bae & Lee (2026), which avoids the nondegeneracy assumption on context diversity. We improve the pilot estimation to overcome the difficulty of learning the high-dimensional concatenated vector. Our analysis integrates EXP4.P and Sparring EXP4.P startups (Beygelzimer et al., 2011; Dudík et al., 2015) with calibrated estimation across bins to bound initialization overheads. Finally, we translate local confidence bounds into calibrated utility regret and aggregate the binwise guarantees via empirical calibration tails, establishing protocol-specific regret bounds.

2 Preliminaries and Problem Formulation

We study contextual preference learning with pointwise ratings or pairwise comparisons (Verma et al., 2025). Both protocols evaluate the same arms under known, context-history-dependent calibration. The learner aims to select arms with near-optimal scores.

Notation. For a positive integer n, let $[n] = \{1, 2, \dots, n\}$, and write $|\mathcal{S}|$ for the cardinality of a finite set $\mathcal{S}$. For vectors $x, y \in \mathbb{R}^d$, we write $\langle x, y\rangle$ for their Euclidean inner product. For a vector $x \in \mathbb{R}^d$, its Euclidean norm is $\|x\|_2$; for a symmetric positive definite matrix $A \in \mathbb{R}^{d\times d}$, its matrix-weighted norm is $\|x\|_A = \sqrt{\langle x, Ax\rangle}$. For symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, the relations $A \succeq B$ and $A \succ B$ mean that $A - B$ is positive semidefinite and positive definite, respectively. We denote the identity matrix by I, the ith standard basis vector by e_i, and the Kronecker product by $\otimes$, with dimensions determined by their use. We write $\mathbb{E}$ for expectation and $\log$ for the natural logarithm; the notation $\widetilde{O}$ suppresses logarithmic factors, with any additional constant dependence stated alongside each bound. The logistic link is $\sigma(x) = (1 + \mathrm{e}^{-x})^{-1}$, and its derivative is $\sigma'(x) = \sigma(x)\big(1 - \sigma(x)\big)$.

2.1 Problem Setup

We consider the stochastic K-armed contextual preference bandit problem for T rounds, where at each round $t \in [T]$, the learner is presented with a context vector $z_t \in \mathbb{R}^d$ describing the current task and the user's circumstances. Each arm $i \in [K]$ represents a fixed answer strategy. The learner is allowed to select a single arm $a_t \in [K]$ for a pointwise rating, or an ordered pair of arms $(a_t, b_t) \in [K]^2$ for a pairwise comparison. The selected action determines which stochastic feedback the environment reveals, as specified below.

Assumption 2.1 (Independent contexts)**.** The contexts $z_1, z_2, \dots, z_T$ are drawn independently from an unknown distribution $\mathcal{F}$ on $\mathbb{R}^d$ and revealed sequentially. Each context is independent of the interaction history before its arrival.

Linear scores and calibration. Arm i has latent utility $\langle z_t, \theta_i^\star\rangle$, where $\theta_i^\star \in \mathbb{R}^d$ is fixed and unknown. The context may be a fixed feature representation; our analysis assumes exact linearity in that representation. The context and arm parameters satisfy $\|z_t\|_2 \le L$ and $\|\theta_i^\star\|_2 \le S$ for known constants $L, S > 0$. Fixed functions ϕ_α and ϕ_β determine the positive sensitivity and common offset from the context history: $\alpha_t = \phi_\alpha(z_1, z_2, \dots, z_t) \in \mathbb{R}^+$ and $\beta_t = \phi_\beta(z_1, z_2, \dots, z_t) \in \mathbb{R}$. The learner observes both calibration values before each action; the functions themselves need not be known. The values are shared by all arms and may depend on the entire context history and its length, but not on actions or realized feedback. The calibrated score of arm i at round t is

$$s_{t,i} = \alpha_t \langle z_t, \theta_i^\star\rangle + \beta_t.$$

Calibration preserves the ranking of arms because sensitivity is positive and the offset is common. The context and parameter bounds imply $|s_{t,i} - \beta_t| \le LS\alpha_t$ for every arm.

Pointwise and pairwise feedback model. At round $t \in [T]$, upon selecting arm $a_t \in [K]$, the learner receives a binary rating R_t, where $R_t = 1$ indicates a positive evaluation of the arm with calibrated score s_{t,a_t}. If the learner instead selects distinct arms a_t and b_t, it receives a binary comparison R_t, where $R_t = 1$ indicates that arm a_t is preferred to arm b_t. Let the pre-action information $\mathcal{H}_t$ include all observed contexts, known calibration values, previous actions, and

observed feedback. Ratings depend on the selected arm's score, while comparisons depend on the score differences: for pointwise, $R_t \mid \mathcal{H}_t, a_t \sim \mathrm{Bernoulli}\big(\sigma(s_{t,a_t})\big)$; for pairwise, $R_t \mid \mathcal{H}_t, a_t, b_t \sim \mathrm{Bernoulli}\big(\sigma(s_{t,a_t} - s_{t,b_t})\big)$. Note that the learner may select a repeated pair (i.e., $a_t = b_t$) to commit to a single arm, in which case no informative comparison is requested (Yue et al., 2012).

Objective: regret minimization. Suppose the optimal arm at round t is $i_t^\star \in \arg\max_{i\in[K]} s_{t,i}$, and denote its score by $s_t^\star = s_{t,i_t^\star}$. We measure the instantaneous regret of a single arm by its score gap from the optimal arm, and that of a pair by the average gap of its two arms. Precisely, the cumulative regrets under the two feedback protocols are

$$\mathrm{Reg}_{\mathrm{pt}}(T) \coloneqq \sum\nolimits_{t=1}^{T}(s_t^\star - s_{t,a_t}), \qquad \mathrm{Reg}_{\mathrm{pr}}(T) \coloneqq \sum\nolimits_{t=1}^{T}\Big(s_t^\star - \frac{1}{2}(s_{t,a_t} + s_{t,b_t})\Big).$$

Allowing repeated arms is necessary for sublinear regret: on a two-arm instance with a fixed positive score gap, selecting distinct arms on every round incurs linear regret even when the optimal arm is known. Weak regret measures the gap to the better of the two selected arms: $\mathrm{Reg}_{\mathrm{pr}}^{\mathrm{weak}}(T) \coloneqq \sum_{t=1}^{T}\big(s_t^\star - \max\{s_{t,a_t}, s_{t,b_t}\}\big)$. Clearly, weak regret is at most the average regret defined above.

To accommodate diverse user sensitivities and evaluation standards, we do not require sensitivity and offset to be bounded independently of the horizon. Instead, we impose empirical tail conditions requiring the frequency of extreme values to decrease exponentially with their magnitude.

Assumption 2.2 (Sensitivity and offset tails)**.** Let $a, b > 0$ be tail rates, and let $C_\alpha > 1$ and $C_\beta \geq 1$ be constants. For every $x \geq 0$, the number of rounds with sensitivity or absolute offset at least x satisfy

$$N_\alpha(x) \coloneqq \big|\{t \in [T] : \alpha_t \geq x\}\big| \leq C_\alpha T \mathrm{e}^{-ax}; \quad N_\beta(x) \coloneqq \big|\{t \in [T] : |\beta_t| \geq x\}\big| \leq C_\beta T \mathrm{e}^{-bx}.$$

This assumption models typically moderate user sensitivity and acceptance standards with occasional extremes; its suitability depends on the application and data.

3 The Algorithms

We propose BinPilot, which combines pilot-centered matrix regularization, calibration-based binning, and protocol-specific startup transitions.

3.1 The common framework

To use the same estimator for both protocols, concatenate all arm parameters into the vector $\Theta = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^{dK}$. The feasible parameter set is $\mathcal{B} = \{\Theta = (\theta_1, \ldots, \theta_K) : \|\theta_i\|_2 \leq S \text{ for every } i \in [K]\}$. Represent an observed rating or comparison by an action feature $x_t \in \mathbb{R}^{dK}$ and a known offset y_t, as specified in Table 1. Thus, both feedback protocols follow the same logistic observation model, with the feature vector and offset defined according to the chosen protocol. Table 1 also gives a bound h_t on the absolute logistic argument over all feasible parameters.

Table 1: Observation representations and bounds used by the common framework.

Feedback	Selected action	Action feature x_t	Offset y_t	Bound h_t
Pointwise	Arm a_t	$\alpha_t e_{a_t} \otimes z_t$	β_t	$LS\alpha_t + \lvert\beta_t\rvert$
Pairwise	Arms (a_t, b_t)	$\alpha_t(e_{a_t} - e_{b_t}) \otimes z_t$	0	$2LS\alpha_t$

Here e_i is the ith standard basis vector in $\mathbb{R}^K$; $e_i \otimes z_t$ places z_t in arm i's parameter block.

Bin $m = 1 + \lfloor h_t \rfloor$ has height m and maintains its own visit count and samples. Every visited bin is retained; neither algorithm uses a hardness cutoff or calibration tail rates. Within each bin, both algorithms follow the three-step framework below: startup handles the initial visits, after which each round uses exploration and estimation to decide whether to collect another sample or commit to an arm.

1. **Startup.** Run the startup stage on the first N_m rounds assigned to bin m, counting only that bin's rounds. The predetermined threshold N_m controls the bin's contribution to

cumulative regret while observations are limited. For pointwise feedback, one EXP4.P learner (Beygelzimer et al., 2011) selects arms and updates from ratings. For pairwise feedback, Sparring EXP4.P (Dudík et al., 2015) uses two learners to select arm pairs and update from comparison feedback.

2. **Exploration and estimation.** The learner maintains accuracy levels with separate sample bins for pilot estimation and prediction correction. Start at level 1 with all arms active; level ℓ uses error tolerance $\eta_\ell = 2^{-\ell}$. For the current context, if the largest confidence interval radius among active actions (those using only arms not yet eliminated) exceeds η_ℓ, execute an action attaining this maximum, store its feedback in a preselected sample bin, and end the round. Otherwise, enter the commitment stage if this radius is at most $T^{-1/2}$; if not, perform elimination and proceed to level $\ell + 1$ within the same round.
3. **Commitment.** If only one active arm remains or the confidence intervals certify the required accuracy for the current context, select the active arm with the highest estimated score. Play it as a single arm for pointwise feedback or a repeated pair for pairwise feedback, and end the round. Commitment applies only to the current round. On the next round assigned to this bin, restart the checks at level 1 with all arms active and each level's samples retained, determining whether to explore or commit again.

We now specify the parameters and estimators that instantiate this framework: the bin initialization, pilot and corrected estimates, confidence radii, and sample-assignment thresholds. Algorithm 1 gives the common routine; Appendix A.1 gives the two protocol-specific algorithms. Section 3.2 specifies their startup lengths and regret guarantees.

Initialization. For bin m, let $\kappa_m = 1/\sigma'(m)$ bound the reciprocal logistic derivative and let δ_m be the probability allowed for failure of that bin's guarantees. Lemma 3.1 below specifies the shared estimation parameters; Appendix A.2 gives the startup policy construction and initialization regret bound $\mathcal{I}_m$ used to choose N_m. For $T \geq 2$, the common routine uses $J = \lceil \frac{1}{2} \log_2 T \rceil$ accuracy levels, so that $\eta_J \leq T^{-1/2}$. At each level ℓ, initialize two empty sample bins: P_ℓ for fitting a preliminary parameter estimate and E_ℓ for adjusting it using additional responses. Each index j stores (x_j, y_j, R_j); only post-startup exploration samples enter these bins, and they are retained across bin visits. Below, we suppress the bin index and write $\kappa = \kappa_m$.

Parameter estimation. We estimate the unknown arm parameters $\Theta^\star$ in two steps, following Bae & Lee (2026). First, we fit logistic regression to the observations indexed by P_ℓ, obtaining a preliminary estimate $\bar{\Theta}_\ell$. This estimate predicts the probability of each response in E_ℓ. We then use the differences between these predictions and the observed responses to update $\bar{\Theta}_\ell$, obtaining $\widehat{\Theta}_\ell$, the parameter estimate used to score and select arms. For a binary response r and logistic argument ζ, let $\mathcal{L}(r, \zeta) \coloneqq \log(1 + \mathrm{e}^\zeta) - r\zeta$ denote the negative log-likelihood. With regularization parameter $\lambda = (KS^2)^{-1}$, the pilot estimate is $\bar{\Theta}_\ell = \arg\min_{\Theta \in \mathcal{B}} \left\{ \sum_{j \in P_\ell} \mathcal{L}(R_j, \langle x_j, \Theta \rangle + y_j) + \frac{\lambda}{2} \|\Theta\|_2^2 \right\}$. To control the correction, we define the pilot design matrix $V_{P,\ell}$ and regularization matrix B_ℓ as $V_{P,\ell} = \kappa\lambda I + \sum_{j \in P_\ell} x_j x_j^\top$ and $B_\ell = V_{P,\ell}/\varrho^2$, where ϱ is the pilot confidence radius. For each $j \in E_\ell$, the prediction error is $R_j - \sigma(\langle x_j, \bar{\Theta}_\ell \rangle + y_j)$. The matrix H_ℓ scales the resulting parameter update according to the logistic curvature and the regularization B_ℓ:$H_\ell = B_\ell + \sum_{j \in E_\ell} \sigma'(\langle x_j, \bar{\Theta}_\ell \rangle + y_j) x_j x_j^\top$; $\widehat{\Theta}_\ell = \bar{\Theta}_\ell + H_\ell^{-1} \sum_{j \in E_\ell} x_j \big(R_j - \sigma(\langle x_j, \bar{\Theta}_\ell \rangle + y_j)\big)$. This update is a single Newton step with a quadratic penalty centered at the pilot and weighted by B_ℓ. Curvature is evaluated at the feasible pilot; the corrected estimate is not projected onto $\mathcal{B}$.

Confidence intervals. The estimated score of arm i at level ℓ is $\widehat{s}_{t,i}^{\,\ell} = \alpha_t \langle e_i \otimes z_t, \widehat{\Theta}_\ell \rangle + \beta_t$. For an action feature x, the confidence radius is $w_\ell(x) = A\|x\|_{H_\ell^{-1}}$, where A is a confidence multiplier. The corresponding interval is centered at the estimated arm score for ratings and the estimated score difference for comparisons. The multiplier A, pilot radius ϱ, and positive sample-assignment thresholds τ_ℓ are chosen as in Lemma 3.1. Together, the common routine's confidence events use failure probability at most $\delta_m/2$, leaving $\delta_m/2$ for the pairwise startup guarantee.

Action selection and sample assignment. Let $\mathcal{A} \subseteq [K]$ denote the arms still active for the current context; it is reset to $[K]$ at the start of each invocation. For $|\mathcal{A}| \geq 2$, the candidate feature set is

$$\mathcal{X}_t(\mathcal{A}) = \begin{cases} \{\alpha_t e_i \otimes z_t : i \in \mathcal{A}\}, & \text{pointwise;} \\ \{\alpha_t (e_i - e_j) \otimes z_t : i, j \in \mathcal{A},\ i \neq j\}, & \text{pairwise.} \end{cases}$$

Algorithm 1 uses the largest radius $w_{\max}$ to choose between exploration, elimination, and commitment. When exploration is required, the selected observation is assigned to P_ℓ if $\|x_t\|^2_{V_{P,\ell}^{-1}} > \tau_\ell$ and to E_ℓ otherwise, before its response is observed. At level ℓ, exploration action selection and sample assignment do not use responses in E_ℓ, although the active set may depend on feedback from lower levels. This separation is used in the confidence analysis in Appendix B.2. The bin state passed to Algorithm 1 consists of $\kappa, \lambda, \varrho, A, (\tau_\ell)_{\ell=1}^J$ and the retained sample bins $(P_\ell, E_\ell)_{\ell=1}^J$; all bins use the global horizon T.

The following lemma gives the parameter choices that make the confidence radii valid for both protocols. We restore the bin index in $\varrho_m, A_m, \tau_{m,\ell}$. Let $c_{P,m}$ be the deterministic factor controlling pilot estimation error and Γ_m the deterministic bound used to control design-matrix growth; Appendix A.2 gives their explicit logarithmic expressions.

Lemma 3.1 (Parameter choices and confidence)**.** *Fix $\delta \in (0,1)$ and set $\kappa_m = 1/\sigma'(m) = 2 + \mathrm{e}^m + \mathrm{e}^{-m}$, $\delta_m = 6\delta/(\pi^2 m^2)$, and $\lambda = (KS^2)^{-1}$. With $\iota_m = \log(8TJK^2/\delta_m)$, choose $\varrho_m^2 = 8\kappa_m c_{P,m} + 5\kappa_m + 1$, $A_m = 2(\sqrt{2\mathrm{e}\iota_m} + \frac{2}{3}\iota_m + 1)$, and $\tau_{m,\ell} = \frac{1}{\varrho_m^2}\min\{1, \frac{\eta_\ell}{4\mathrm{e}\sqrt{2\Gamma_m}}\}$. With probability at least $1 - \delta_m/2$, simultaneously for all invocations, levels, and candidate action features x in bin m, $|\langle x, \widehat{\Theta}_\ell - \Theta^\star\rangle| \le A_m \|x\|_{H_\ell^{-1}}$.*

This bounds arm-score error for ratings and score-difference error for comparisons. The proof is given by Theorem B.6 in Appendix B.2. The bin budgets sum to δ, allowing the two regret guarantees below to hold across all visited bins.

Algorithm 1 Shared exploration, estimation, and commitment for one round

Require: Context z_t, calibration α_t, β_t, horizon T, arm count K, and feedback protocol
Require: Bin m's parameters and sample bins, as defined in Section 3.1; number of levels J

Initialize the active arms $\mathcal{A} \leftarrow [K]$.
for $\ell = 1, \dots, J$ **do**
  Set the level tolerance $\eta_\ell \leftarrow 2^{-\ell}$.
  if $|\mathcal{A}| = 1$ **then**
    return Play the sole active arm, or its repeated pair.
  Compute $\bar{\Theta}_\ell, V_{P,\ell}, H_\ell, \widehat{\Theta}_\ell$ from P_ℓ, E_ℓ using the formulas in Section 3.1.
  Compute scores $\widehat{s}_{t,i}^\ell$ for $i \in \mathcal{A}$ and radii $w_\ell(x) = A\|x\|_{H_\ell^{-1}}$.
  Set $w_{\max} \leftarrow \max_{x \in \mathcal{X}_t(\mathcal{A})} w_\ell(x)$.
  if $w_{\max} > \eta_\ell$ **then**
    Select an active action attaining $w_{\max}$; obtain x_t, y_t from Table 1.
    Choose P_ℓ if $\|x_t\|^2_{V_{P,\ell}^{-1}} > \tau_\ell$, and E_ℓ otherwise.
    return Play the selected action, observe R_t, and store (x_t, y_t, R_t) in the chosen bin.
  else if $w_{\max} \le T^{-1/2}$ **then**
    Select $i \in \arg\max_{j \in \mathcal{A}} \widehat{s}_{t,j}^\ell$.
    return Play arm i, or pair (i, i).
  else $\mathcal{A} \leftarrow \{i \in \mathcal{A} : \widehat{s}_{t,i}^\ell \ge \max_{j \in \mathcal{A}} \widehat{s}_{t,j}^\ell - 2\eta_\ell\}$.

3.2 Pointwise and pairwise algorithms

Algorithms 2 and 3 in Appendix A.1 instantiate BinPilot using the protocol-specific hardness measures in Table 1. Our switching design controls the initialization cost of bins with few visits. For pairwise feedback, we combine Sparring EXP4.P (Dudík et al., 2015) with the common elimination routine: startup incurs regret at most $\mathcal{S}_m\sqrt{n}$ on the event of Lemma A.1, so a bin switches only when this bound reaches the initialization cost $\mathcal{I}_m$. Balancing these costs gives $N_m^{\mathrm{pr}} = \lceil(\mathcal{I}_m/\mathcal{S}_m)^2\rceil$; the pointwise bound $2mn$ similarly gives $N_m^{\mathrm{pt}} = \lceil\mathcal{I}_m/(2m)\rceil$. Neither transition requires the bin's final visit count or the calibration tail rates. Appendix A.2 specifies the coefficients and startup tuning.

After startup, both follow the exploration, estimation, and commitment rules of Section 3.1, with the shared parameters of Lemma 3.1. For pairwise feedback, elimination also improves the curvature: on the confidence event, fine-level comparisons have score differences of magnitude at most one, so their logistic derivative is at least $\sigma'(1)$ regardless of the bin's original hardness. This removes the

exponential curvature cost from fine-level estimation, while startup controls sparsely visited bins; Appendix C.1 proves these bounds.

The following guarantees use the effective tail indices $\rho_{\mathrm{pt}} = ab/(a + LSb)$ and $\rho_{\mathrm{pr}} = a/(2LS)$.

Theorem 3.2 (Regret upper bounds)**.** *For any $\delta \in (0, 1)$, Algorithms 2 and 3 achieve the following pointwise and pairwise regret bounds, respectively, each with probability at least $1 - \delta$:*

$$\mathrm{Reg}_{\mathrm{pt}}(T) \le \widetilde{O}\Big(\sqrt{dKT} + (dK)^{\frac{5\rho_{\mathrm{pt}}}{2(1+\rho_{\mathrm{pt}})}} T^{\frac{1}{1+\rho_{\mathrm{pt}}}}\Big),$$
$$\mathrm{Reg}_{\mathrm{pr}}(T) \le \widetilde{O}\Big(\sqrt{dKT} + d^{\frac{5\rho_{\mathrm{pr}}+2}{2(2+\rho_{\mathrm{pr}})}} K^{\frac{5\rho_{\mathrm{pr}}+4}{2(2+\rho_{\mathrm{pr}})}} T^{\frac{1}{2+\rho_{\mathrm{pr}}}}\Big).$$

The pointwise and pairwise bounds in Theorem 3.2 are proved separately in Appendices C.2.3 and C.2.4, respectively, using the bin regret bounds in Appendix C.1.

Remark 3.1 (Regret and calibration information). In both bounds, the first term $\sqrt{dKT}$ is the estimation cost of learning the dK arm parameters. The second term accounts for initializing the learners across hardness bins, balanced against the frequency of difficult rounds. Pointwise startup is controlled by a linear bound on each bin's visits, while Sparring gives a square-root bound for pairwise startup. This difference produces the horizon exponents $1/(1 + \rho_{\mathrm{pt}})$ and $1/(2 + \rho_{\mathrm{pr}})$ in the respective initialization terms. Both algorithms require only the realized calibration values α_t, β_t before each action; the functions ϕ_α, ϕ_β need not be disclosed to the learner.

Remark 3.2 (Regret comparison). For fixed d, K, and model parameters, the pairwise regret bound is $\widetilde{O}(\sqrt{T})$ for every $\rho_{\mathrm{pr}} > 0$. The pointwise bound has horizon exponent no less than $1/2$, whose equality holds if and only if $\rho_{\mathrm{pt}} = ab/(a+LSb) \ge 1$. Thus, the pairwise bound has a strictly smaller horizon exponent when $\rho_{\mathrm{pt}} < 1$, and the same exponent otherwise. Interestingly, these asymptotic rates alone do not determine which bound is smaller at a given horizon, since the coefficients and initialization costs differ between the two protocols. When offsets are small, pointwise feedback can also have a smaller hardness bound: $LS\alpha_t + |\beta_t|$ may be less than $2LS\alpha_t$.

For any fixed $\epsilon > 0$, a randomized argument in Appendix C.2.6 yields a modification of Algorithm 3 that preserves the pairwise expected regret rate of Theorem 3.2 and satisfies

$$\mathbb{E}\Big[\mathrm{Reg}_{\mathrm{pr}}^{\mathrm{weak}}(T)\Big] \le \mathbb{E}\big[\mathrm{Reg}_{\mathrm{pr}}(T)\big] \le (1+\epsilon)\mathbb{E}\Big[\mathrm{Reg}_{\mathrm{pr}}^{\mathrm{weak}}(T)\Big].$$

The following lower bounds show that the $\sqrt{dKT}$ estimation term is unavoidable even for known constant calibration.

Theorem 3.3 (Regret lower bounds)**.** *Let $d \ge 1$, $K \ge 3$, $L, S > 0$, and $T \ge d^2K/(L^2S^2)$. Fix the known calibration $\alpha_t \equiv 1$ and $\beta_t \equiv 0$. There exists a uniform context distribution on $\{Le_1, Le_2, \ldots, Le_d\}$ such that, for every possibly randomized pointwise or pairwise algorithm, there are arm parameters with $\|\theta_i\|_2 \le S$ for all $i \in [K]$ satisfying*

$$\mathbb{E}\big[\mathrm{Reg}_{\mathrm{pt}}(T)\big] \ge \frac{1}{8}\sqrt{dKT}, \qquad \mathbb{E}\big[\mathrm{Reg}_{\mathrm{pr}}(T)\big] \ge \mathbb{E}\Big[\mathrm{Reg}_{\mathrm{pr}}^{\mathrm{weak}}(T)\Big] \ge \frac{1}{16}\sqrt{dKT}.$$

The pointwise and pairwise claims are respectively proved in Appendices D.1 and D.2, matching the common $\sqrt{dKT}$ term. Together with the pointwise $\Omega(T^{1/(1+\rho_{\mathrm{pt}})})$ lower bound in Appendix D.3, these establish the near-optimality of our algorithms' regret rates for both protocols.

4 Experiments

We conduct two experiments to examine the main result of our analysis. First, using a controlled synthetic instance, we vary the tail behavior of the context-dependent sensitivity α_t and common offset β_t and study their effects on pointwise and pairwise regret. Second, using PrivacyLens contexts (Shao et al., 2024), we compare practical learners to determine whether incorporating known context-dependent calibration improves arm recommendations. Appendix E provides the complete environment constructions, implementation details, and statistical protocols.

4.1 Synthetic experiment

In this section, we compare a fast tail, for which the square-root term dominates, with a slow tail, for which the additional term dominates.

Instance construction. We construct a two-arm instance in $\mathbb{R}^{10}$ that isolates the effect of the common offset. Let e_1 be the first standard basis vector, set $\theta_1^\star = e_1$ and $\theta_2^\star = -e_1$, and use a balanced sequence of contexts $z_t \in \{-e_1, e_1\}$. Each arm is optimal equally often, and the absolute utility gap remains constant. For each horizon T, we independently sample two uniformly random permutations π_α and π_β of $[T]$ and set $\alpha_t = \frac{1}{a}\log\left(\frac{T+1}{\pi_\alpha(t)}\right)$ and $|\beta_t| = \frac{1}{b}\log\left(\frac{T+1}{\pi_\beta(t)}\right)$. We then assign balanced random signs to the resulting offset magnitudes. We compare a fast-tail setting with $(a, b) = (2, 2)$ against a slow one with $(a, b) = (2, 0.05)$, which produces substantially more saturated pointwise observations. Thus, the construction changes the difficulty of pointwise feedback while preserving the utility decisions and canceling the common offset from pairwise comparisons.

We evaluate a sequential implementation of the proposed framework with practical startup and sample-assignment thresholds at horizons $T = 2^r$ for $r \in \{8, \dots, 21\}$. The implementation retains score binning, protocol-specific startup, and the layered pilot and one-step correction procedures. We shorten the startup thresholds and relax the pilot sample-assignment thresholds as specified in Appendix E.1. All remaining estimator parameters follow their theoretical definitions with $\delta = 0.05$. Each horizon uses ten paired seeds; the protocols share context and calibration sequences, with feedback draws coupled through common uniform random numbers. Across offset settings, paired runs also share the contexts, sensitivities, offset ranks and signs, learner randomization, and feedback uniforms; only the offset magnitudes differ. We evaluate the protocol-specific cumulative score regret defined in Section 2 and report means with 95% seed-level bootstrap intervals. Appendix E.1 provides the implementation and slope-estimation details.

Regret versus horizon. Figure 1(a) plots cumulative regret against the horizon on logarithmic axes. The $\sqrt{T}$ and T reference lines indicate slopes. The four curves are close at shorter horizons, but the slow-offset-tail pointwise curve increasingly separates from the others and runs close to the linear reference slope at larger horizons. Under the fast tail ($a = b = 2$), the fitted late-horizon slopes are 0.598 for pointwise feedback and 0.554 for pairwise feedback, respectively. Their finite-horizon growth is compatible with square-root dependence modified by logarithmic factors. Under the slow offset tail ($a = 2$, $b = 0.05$), the pointwise slope rises to 0.975.

The two pairwise trajectories coincide. The settings use the same sensitivity sequence, and the common offset cancels from the comparison logit. With the remaining inputs and random numbers shared across the paired runs, changing the offset tail leaves the pairwise feedback and decisions unchanged. Pointwise ratings behave differently: a large $|\beta_t|$ can move both arms' feedback probabilities close to zero or one. Their binary ratings then provide less information about which arm has higher latent utility, even though the learner knows β_t and uses the correct likelihood. This is consistent with the worse pointwise performance under the slow offset tail, while pairwise feedback remains unaffected.

The slow offset tail also spreads pointwise observations across more score bins. At $T = 2^{21}$, the mean number of active bins rises from eight under the fast tail to 251 under the slow offset tail, leaving observations distributed among many bin-specific learners. Together, the separation of the pointwise curves and the overlap of the pairwise curves show the effect of the common offset and the qualitative difference suggested by the regret bounds.

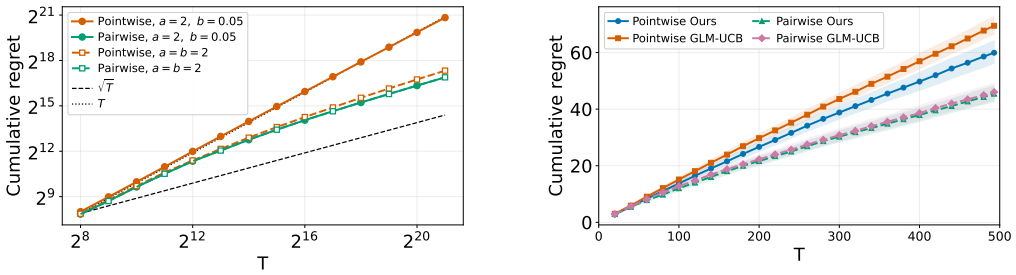
4.2 PrivacyLens contexts

We next examine whether accounting for known context-dependent calibration improves recommendations on privacy-sensitive contexts.

Environment and methods. PrivacyLens provides 493 privacy-sensitive scenarios and associated agent trajectories (Shao et al., 2024). We embed the information available before the final action, excluding the action and evaluation outputs, and obtain normalized contexts $z_t \in \mathbb{R}^{128}$. We construct three synthetic arm vectors from the context data, subtract their mean, and jointly scale them to $S = 1$. Thus, $d = 128$, $K = 3$, and $L = 1$. The median gap between the best and second-best arms is 0.148.

Frozen projections of the embeddings generate known context-dependent calibration values satisfying $0.25 < \alpha_t < 2$ and $|\beta_t| < 3$. Positive sensitivity preserves the latent ranking, while sensitivity and offset jointly determine pointwise feedback probabilities. For pairwise feedback, the common offset cancels from the comparison probability.

The first 345 contexts form the historical phase, followed by 148 online rounds. Within each feedback protocol, the two methods use the same context order and underlying feedback random numbers. Our practical learner incorporates the known α_t and β_t in its online Newton updates. For pointwise



(a) Synthetic data

(b) PrivacyLens contexts

Figure 1: **Synthetic and real-context experiments. (a)** Under $(a, b) = (2, 2)$ and $(a, b) = (2, 0.05)$ on log–log axes; the $\sqrt{T}$ and T lines indicate reference slopes. **(b)** 493 PrivacyLens contexts; lines denote means and shaded regions denote 95% confidence intervals.

feedback, it selects an arm using a calibrated optimistic score. For pairwise feedback, it selects the arm pair with the largest uncertainty in their estimated utility difference.

The GLM-UCB-style baseline (Filippi et al., 2010) maintains one global online Newton model (Jun et al., 2017) with the same regularization and confidence multiplier, but assumes $\alpha_t = 1$ and $\beta_t = 0$ and does not use score bins. Its pointwise action maximizes an optimistic utility estimate. Its pairwise adaptation constructs difference features, identifies arms that may still be optimal, and queries the most uncertain pair among them (Saha, 2021). Both methods are evaluated using the true calibrated score gaps, although the baseline does not use the true calibration values for learning. Appendix E.2 specifies the environment construction and hyperparameters.

Evaluation. Our practical adaptation retains calibration-aware updates and routing by score bounds, while replacing the theoretical startup and layered estimation procedures with online Newton updates and a two-bin historical warm start. For both methods, we evaluate the protocol-specific cumulative regret defined in Section 2 using the true context-dependent sensitivity α_t. We use 50 paired seeds and report 95% bootstrap intervals over context order and feedback randomness. Figure 1(b) displays all 493 rounds, including the 345-round historical phase, alongside differences restricted to the subsequent 148 online rounds. Through this setup, we evaluate the whole algorithm as an integrated procedure, assessing its overall performance relative to the uncalibrated baseline.

Results. Figure 1(b) shows a persistent separation between the pointwise curves. At round 493, mean cumulative pointwise regret is 59.925 for Ours and 66.893 for GLM-UCB. The paired difference, Ours minus GLM-UCB, is -6.968. Restricting the comparison to the final 148 online rounds gives a difference of -3.016. Thus, the separation visible over the full sequence is not confined to the historical phase. The baseline treats feedback generated with varying offsets and sensitivities as though both calibration values were fixed. Our learner uses their known values in its updates, which helps explain its lower pointwise regret when those values vary across contexts.

The pairwise curves show smaller differences throughout the plotted sequence. Their final mean cumulative regrets are 43.255 for Ours and 45.757 for GLM-UCB. The paired difference is -2.502. The smaller separation is consistent with the feedback model: unlike a pointwise rating, a comparison removes the common offset from its probability. Positive sensitivity preserves the preferred arm, although it still changes how likely each comparison outcome is. Thus, the pointwise separation and pairwise proximity in the figure reflect the different roles of calibration under the two protocols.

5 Conclusion

We have presented a unified framework for context-aware preference bandits that quantifies how known, context-dependent calibration governs learning from both pointwise ratings and pairwise comparisons. Our algorithms attain near-optimal pointwise regret of $\widetilde{O}\big(\sqrt{T} + T^{1/(1+\rho_{\mathrm{pt}})}\big)$ and pairwise regret of $\widetilde{O}(\sqrt{T})$, even without knowing or estimating the calibration tail rates. These results establish an asymptotic minimax advantage for pairwise. Synthetic experiments at long horizons exhibit regret trends consistent with the theory, while experiments with real-world contexts show that a practical adaptation of our framework achieves lower regret than the GLM-UCB-style baseline.

AI USE STATEMENT

In this work, we used generative AI tools for help assist in the writing of proofs and design or provide feedback on research methodology or experiments. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, generate synthetic data sets, implement methods, clean and reformat datasets, or interpret results, and activities classified as supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft parts of a research paper, summarize or analyze existing literature, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, and suggest a structure for a research paper. We have reviewed all AI-assisted work. Our review covered the mathematical arguments and their consistency with the stated assumptions, the interpretation of experimental results, and the accuracy and attribution of related work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Seoungbin Bae and Dabeen Lee. Logistic bandits with $\widetilde{O}(\sqrt{dT})$ regret without context diversity assumptions. *arXiv preprint arXiv:2604.22161*, 2026. doi: 10.48550/arXiv.2604.22161.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7): 1–108, 2021.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 19–26. JMLR.org, 2011.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4299–4307, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214. JMLR.org, 2011.

Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13513–13545. PMLR, 2025.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 563–587. PMLR, 2015.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3052–3060. PMLR, 2020.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23, pp. 586–594, 2010.

Xiang Ji, Huazheng Wang, Minshuo Chen, Tuo Zhao, and Mengdi Wang. Provable benefits of policy learning from human preferences in contextual bandit problems. *arXiv preprint arXiv:2307.12975*, 2023. doi: 10.48550/arXiv.2307.12975.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, volume 30, pp. 99–109, 2017.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5148–5157. PMLR, 2021.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2776–2785. PMLR, 2018.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4474–4482. PMLR, 2024a.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124640–124685, 2024b. doi: 10.52202/079017-3960.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670. Association for Computing Machinery, 2010. doi: 10.1145/1772690.1772758.

Kaiqu Liang, Julia Kruk, Shengyi Qian, Xianjun Yang, Shengjie Bi, Yuanshun Yao, Shaoliang Nie, Mingyang Zhang, Lijuan Liu, Jaime Fernández Fisac, Shuyan Zhou, and Saghar Hosseini. Learning personalized agents from human feedback. *arXiv preprint arXiv:2602.16173*, 2026. doi: 10.48550/arXiv.2602.16173.

Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, 2004.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30050–30062, 2021.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 968–994. PMLR, 2022.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. PrivacyLens: Evaluating privacy norm awareness of language models in action. In *Advances in Neural Information Processing Systems*, volume 37, pp. 89373–89407, 2024. doi: 10.52202/079017-2837.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020.

Arun Verma, Zhongxiang Dai, Xiaoqiang Lin, Patrick Jaillet, and Bryan Kian Hsiang Low. Neural dueling bandits: Preference-based optimization with human feedback. In *International Conference on Learning Representations*, 2025.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer 2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501, 2024. doi: 10.52202/079017-0047.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. HelpSteer2-Preference: Complementing ratings with preferences. In *International Conference on Learning Representations*, 2025.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. doi: 10.1016/j.jcss.2011.12.028.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 392–401. PMLR, 2016.