Inverse Recommender Distillation: Reclaiming Low-Rank Personal Preference Subspaces from Confounded Platform Feeds

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Users generate thousands of implicit interaction traces—clicks, skips, dwell durations, and decision latencies—across commercial recommender feeds while retaining no portable model of their preferences. Conversely, foundation models aligned via monolithic Bradley–Terry aggregation converge to Borda-count majoritarianism, inducing high utility distortion for minority subgroups. Although statutory portability regimes (GDPR Article 20, DMA Article 6(9), and the EU Data Act for connected products) enable interaction archive exports, naive cloning or unregularized Direct Preference Optimization (DPO) on exported feeds fails because platform rankings confound user utility u_i^* with commercial engagement bias β_{plat}. We formulate **Inverse Recommender Distillation on Pluralistic Low-Rank Subspaces (IRD-PAL)**, unifying exposure de-biasing with the Pluralistic Alignment Framework (PAL) and drift-diffusion chronometric choice modeling. IRD-PAL (1) estimates $\hat{\beta}_{\text{plat}}$ purely from observable top-of-slate exposure and short-dwell bounce contrasts via $P_{\perp\text{plat}}$ while bounding unidentifiable collinear overlap $\|P_{\text{plat}}\zeta_i^*\|_2$, (2) separates pre-reading dwell gates from post-reading drift-diffusion decision times, and (3) fits barycentric coordinates $w_i \in \Delta^{K-1}$ with $M_{\text{act}} = 4$ noisy Fisher-curvature-targeted active queries. Across 16,800 simulated episodes, observable IRD-PAL achieves 83.50% [80.32, 86.26] clean recovery at $N_{\text{obs}} = 100$ (85.64% [80.14, 89.81] for minority subgroups; vs. 90.67% oracle upper bound), outperforming raw cloning (7.17% [5.36, 9.51]) and unregularized DPO (24.50% [21.23, 28.10]) while cutting bias contamination from 91.67% to 14.33% [11.76, 17.36] at 2.39 ms p50 latency.

1 Introduction

A central obstacle in deploying sovereign personal AI agents is the twin failure of *monolithic cold-start alignment* and *platform-confounded behavioral cloning*. When a foundation model is aligned via pooled Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO) [22], heterogeneous annotator comparisons are collapsed into a single global reward vector. Formal social-choice analyses prove that Bradley–Terry maximum likelihood over latent user contexts implicitly implements Borda-count consensus aggregation [15, 25], which inflicts high worst-group utility distortion on minority or specialized preference subgroups [4, 8, 26] even when population-average utilitarian metrics appear benign [20]. Empirical evaluations on Personalized RewardBench [17] and Personalized Benchmarking [14] confirm that monolithic reward models systematically misalign with idiosyncratic user values [7, 28].

At the same time, most users have already generated years of interaction traces across commercial recommender systems. Under data portability mandates such as GDPR Article 20 and Digital Markets Act (DMA) Article 6(9) for online platform feeds—alongside the European Union

Data Act [11], which entered into force in January 2024 and applies from 12 September 2025 to connected-product IoT telemetry and cloud switching (Chapter VI)—individuals can increasingly export historical interaction archives or observe served slates client-side.

1.1 Why naive feed cloning and unregularized DPO import platform pathology

If a user exports ranked slates and click logs from a commercial platform, why not simply run behavioral cloning or standard DPO [22] to personalize their local agent? Empirical and econometric research from 2022–2026 identifies three structural failure modes:

1. **Exposure bias and commercial engagement confounding.** As demonstrated in fully observed recommender benchmarks such as KuaiRec [13] and counterfactual policy learning [18], commercial feeds are not neutral probes of reflective user utility $u_i^*(x)$. Platform scoring functions combine $u_i^*(x)$ with commercial engagement vector β_{plat} (optimizing ad revenue, outrage arousal, and session retention). When a local model imitates clicked feed items or runs unregularized DPO without exposure de-biasing, ingesting more traces ($N_{\text{obs}} \to 250$) perversely locks in stronger platform engagement-bias contamination.

2. **Conflation of skim dwell times with decision response times.** Raw click logs mix accidental sub-second clickbait bounces with deliberate endorsements [7]. Under cognitive Drift-Diffusion Models (DDM) [23] formalized for preference alignment by Echenique et al. [9], decision latency carries vital Fisher information, but only when pre-choice content inspection dwell time τ_{dwell} is separated from post-reading choice latency τ_{dec}: sub-second skims ($\tau_{\text{dwell}} < 1\,\text{s}$) carry near-zero utility signal, whereas after content inspection ($\tau_{\text{dwell}} \geq 2\,\text{s}$), *rapid* decisions ($\tau_{\text{dec}}$ small) reflect *large* latent utility margins $|\Delta u_i^*|$.

3. **The reward model selection crisis and manifold inversion risks.** Fine-tuning high-dimensional per-user reward models or LoRA adapters [10] independently on small personal logs suffers from the "Reward Model Selection Crisis" documented by Rezk et al. [24]: validation accuracy on noisy historical clicks correlates poorly with out-of-distribution generalization. Conversely, uploading dense user preference embeddings to cloud servers exposes private histories to Semantic Manifold Alignment attacks [1, 16], which reconstruct raw private attributes by aligning embedding geometry.

1.2 Low-rank pluralistic subspaces as the reclamation backbone

Rather than estimating an unconstrained parameter vector in $\mathbb{R}^d$ from confounded logs, pluralistic alignment models human preference variation on a low-rank subspace. Both Variational Preference Learning (VPL) [21] and the Pluralistic Alignment Framework (PAL) [5] represent each user i via a compact latent coordinate $w_i \in \Delta^{K-1}$ over $K \ll d$ shared basis prototypes $\{\mathbf{p}_1, \ldots, \mathbf{p}_K\}$. Cross-architecture adapter projection methods such as Trans-LoRA [27] and LoRA-X [12] further show that low-rank updates restricted to anchor subspaces can be transferred across foundation model backbones without retraining from scratch.

1.3 Contributions

We unify inverse recommender de-biasing and low-rank pluralistic subspace inference into **IRD-PAL**:

1. **Architecture for de-biased subspace reclamation (Section 2).** We synthesize counterfactual feed de-biasing [13, 18], two-stage chronometric DDM weighting [9, 23], PAL barycen-

tric subspace estimation [5, 21], and Fisher-curvature-weighted active disambiguation inside a local privacy boundary (Table 1, Figure 1).

2. **Formal identifiability, Borda distortion, and chronometric MAP framework (Section 3).** We bound worst-group Borda consensus distortion via between-group preference divergence [4, 15, 25], formulate the two-stage dwell-gate $\times$ inverse-DDM-latency weight $\omega(\tau_{\text{dwell},t}, \tau_{\text{dec},t})$, derive the PAL simplex MAP estimator with empirical Bradley–Terry Fisher/Hessian curvature $\mathbf{H}_i$, and explicitly include the unidentifiable collinear bias term $\|P_{\text{plat}}\zeta_i^*\|_2$ in the recovery bound (Equations 1–6).

3. **16,800-episode numerical benchmark with fitted estimators (Sections 4–5).** Unlike hand-set curves, our simulation (`experiment.py`) generates synthetic feed slates under Equation (2) and numerically fits all four estimators across 9,600 primary episodes ($N_{\text{obs}} \in \{25, 50, 100, 250\}$, $N = 600$ per cell) where ternary slate outcomes sum strictly to 100.0%, plus 3,600 collinearity break-even trials over $\rho_{\text{align}} = \cos(\beta_{\text{plat}}, \zeta_i^*) \in [0.00, 0.75]$ and 3,600 Fisher-curvature-weighted differential privacy trials across $r = 64$ rank dimensions (Tables 2–4, Figures 2–3).

4. **Adversarial limits and falsification protocol on KuaiRec and PRISM (Sections 6–8).** We examine collinear engagement-preference non-identifiability, Semantic Manifold Alignment attacks [16], and the Reward Model Selection Crisis [24], concluding with six preregisterable falsification tests on KuaiRec [13] and Personalized RewardBench [17] (Table 5).

2 Recommender confounding, Borda distortion, and low-rank subspaces

Reclaiming a user's preference manifold from exported platform feeds requires jointly solving an econometric de-confounding problem and a small-sample pluralistic representation problem. Table 1 connects five operational requirements of personal footprint reclamation to their formal statistical mechanisms and primary literature grounding.

As summarized in Table 1, neither raw behavioral cloning nor purely verbal profile summarization captures a user's full preference geometry. Raw feed cloning copies β_{plat} directly into the personal agent, whereas coarse natural-language rubric summaries filter out clickbait bounces via human inspection [28] but suffer quantization loss on continuous trade-offs across prototypes. Figure 1 illustrates the end-to-end IRD-PAL pipeline.

Table 1: Mapping digital-footprint reclamation requirements to econometric and low-rank alignment constructs.

Reclamation goal	Formal construct	IRD-PAL module	Operational function	Primary grounding
Export platform interaction logs	Statutory portability & feed logging	GDPR Art. 20 / DMA Art. 6(9) importer ($\mathcal{O}_{\text{feed}}$)	Ingests ranked candidate slates, dwell durations, and decision timestamps	[11, 13, 18]
Eliminate majoritarian consensus distortion	Low-rank pluralistic ideal-point subspace	PAL barycentric coordinate ($w_i \in \Delta^{K-1}$)	Represents user utility as a convex mixture of $K = 8$ orthogonal prototypes rather than global Borda average	[5, 15, 21, 25]
Strip commercial clickbait bias	Orthogonal exposure-bias projection	Common-mode projector ($P_{\perp\text{plat}}$)	Estimates platform engagement vector $\hat{\beta}_{\text{plat}}$ from bounce-vs.-completion shift and projects it out	[7, 13, 19]
Separate skim bounces from fast confident choices	Two-stage dwell gate × DDM chronometry	Precision weight ($\omega(\tau_{\text{dwell}}, \tau_{\text{dec}})$)	Gates out < 1 s uninspected skims while upweighting fast post-reading DDM decisions	[9, 23]
Resolve collinear or underspecified axes	Fisher/Hessian epistemic query	Curvature-guided active elicitor ($\mathbf{H}_i^{-1}$)	Queries $M_{\text{act}} \leq 4$ pairs along minimum-eigenvalue directions of the empirical Fisher matrix	[2, 12, 24]

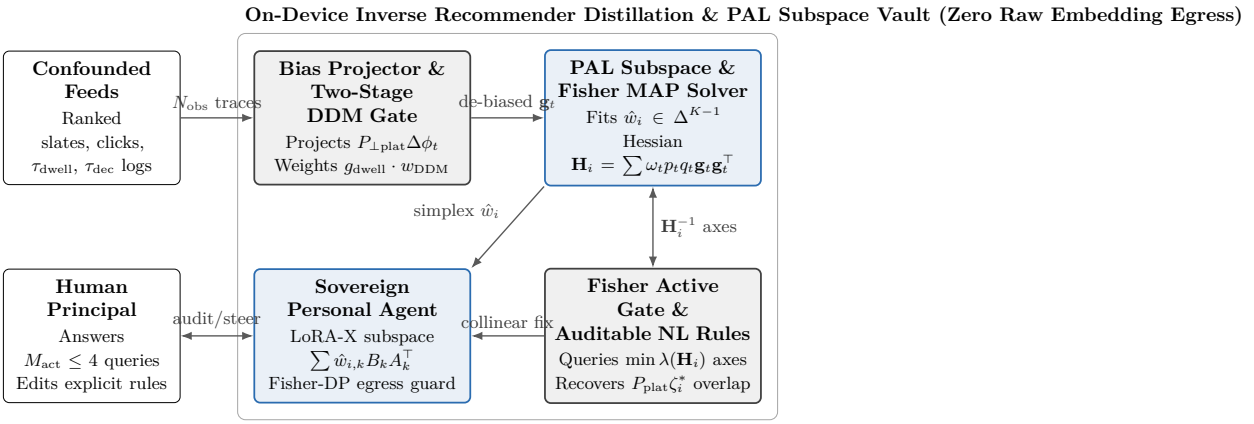


Figure 1: Architecture of IRD-PAL. Exported platform traces pass through an Orthogonal Bias Projector ($P_{\perp\text{plat}}$) and a Two-Stage Chronometric Gate that filters < 1 s skim bounces via $g_{\text{dwell}}(\tau_{\text{dwell}})$ while weighting deliberate choices by inverse drift-diffusion latency $w_{\text{DDM}}(\tau_{\text{dec}})$. The PAL MAP solver estimates barycentric coordinates $\hat{w}_i \in \Delta^{K-1}$ and computes the empirical Bradley–Terry Fisher matrix $\mathbf{H}_i$ to trigger $M_{\text{act}} \leq 4$ active disambiguation queries along underspecified or bias-collinear directions.

3 Formal mathematical framework

We formalize Borda consensus distortion across pluralistic user subgroups, structural de-biasing of confounded recommender slates, two-stage drift-diffusion response-time weighting, and sample complexity with collinear bias bounds.

3.1 Between-group Borda distortion vs. PAL low-rank subspaces

Consider a heterogeneous user population $\mathcal{U}$ partitioned into S preference subgroups $\mathcal{G}_1, \ldots, \mathcal{G}_S$ (e.g., majority mainstream clusters versus minority technical or aesthetic subgroups). Each user $i \in \mathcal{U}$ evaluates items $x \in \mathcal{X}$ via a frozen feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ (d-dimensional normalized representation) and an individual ideal preference vector $\zeta_i^* \in \mathbb{S}^{d-1}$, inducing linear-quadratic utility $u_i^*(x) = \langle \zeta_i^*, \phi(x) \rangle$.

When standard RLHF or DPO [22] fits a single monolithic reward parameter $\bar{\zeta}_{\text{Borda}}$ to pooled population comparisons without subgroup conditioning, Siththaranjan et al. [25] and Gölz et al. [15] prove that the estimator converges to the Borda-count utilitarian mean $\bar{\zeta}_{\text{Borda}} = \sum_{s=1}^{S} \pi_s \bar{\zeta}_{\mathcal{G}_s}$, where $\pi_s = |\mathcal{G}_s|/|\mathcal{U}|$ is the subgroup population share. Unlike within-group noise, the worst-group utility distortion $\mathcal{D}_{\text{group}}(\bar{\zeta}_{\text{Borda}})$ suffered by a minority subgroup $\mathcal{G}_m$ ($\pi_m \ll 1$) is lower-bounded by the *between-group preference divergence* [4, 25]:

$$\mathcal{D}_{\text{group}}(\bar{\zeta}_{\text{Borda}}) \triangleq \max_{s \in \{1,\ldots,S\}} \frac{\mathbb{E}_{i \in \mathcal{G}_s}[\max_{x \in \mathcal{X}} u_i^*(x)]}{\mathbb{E}_{i \in \mathcal{G}_s}[u_i^*(x^*(\bar{\zeta}_{\text{Borda}}))]} \geq 1 + \Omega\Big(\max_{s \neq s'} (1 - \pi_s)^2 \|\bar{\zeta}_{\mathcal{G}_s} - \bar{\zeta}_{\mathcal{G}_{s'}}\|_2^2 \Big). \quad (1)$$

To eliminate this inter-group distortion floor without fitting an unconstrained d-dimensional vector per user, PAL [5] and VPL [21] parameterize $\zeta_i^* = \mathbf{P} w_i^* / \|\mathbf{P} w_i^*\|_2$ using $K \ll d$ shared orthonormal value prototypes $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_K] \in \mathbb{R}^{d \times K}$ and user-specific barycentric weights $w_i^* \in \Delta^{K-1}$.

3.2 Confounded platform feeds and two-stage chronometric DDM weighting

Suppose user i exports N_{obs} pairwise interaction traces $\mathcal{D}_i = \{(\Delta\phi_t, y_t, \tau_{\text{dwell},t}, \tau_{\text{dec},t})\}_{t=1}^{N_{\text{obs}}}$ from a commercial platform, where $\Delta\phi_t = \phi(x_t^+) - \phi(x_t^-)$, $y_t \in \{-1, +1\}$ indicates the selected item, $\tau_{\text{dwell},t} > 0$ is the content inspection dwell duration prior to commitment, and $\tau_{\text{dec},t} > 0$ is the post-inspection choice response time.

A commercial recommender does not rank slates solely by $u_i^*(x)$; it promotes candidates aligned with a platform commercial engagement vector $\beta_{\text{plat}} \in \mathbb{S}^{d-1}$. Moreover, user interactions follow a two-regime behavioral process: during an uninspected skim bounce ($\tau_{\text{dwell},t} < \tau_{\min}$, occurring with probability p_{bounce}), clicks are dominated by surface clickbait salience $\langle \beta_{\text{plat}}, \Delta\phi_t \rangle$; during substantive content inspection ($\tau_{\text{dwell},t} \geq \tau_{\min}$), choices reflect a contaminated utility margin $\Delta r_{i,t}$:

$$\Delta r_{i,t} = \alpha_{\text{user}}(t) \langle \zeta_i^*, \Delta\phi_t \rangle + \alpha_{\text{plat}}(t) \langle \beta_{\text{plat}}, \Delta\phi_t \rangle + \xi_t, \quad (2)$$

where $\alpha_{\text{plat}}(t) \gg \alpha_{\text{user}}(t)$ for short-dwell bounces ($\tau_{\text{dwell},t} < \tau_{\min}$) and $\alpha_{\text{user}}(t) \gg \alpha_{\text{plat}}(t)$ for inspected comparisons ($\tau_{\text{dwell},t} \geq \tau_{\min}$). Under the Drift-Diffusion Model (DDM) of two-choice decision making [9, 23] with accumulation barrier $b > 0$, once the user has inspected the candidates ($\tau_{\text{dwell},t} \geq \tau_{\min}$), the choice probability and expected post-reading decision latency $\tau_{\text{dec},t}$ satisfy:

$$\Pr(y_t = +1 \mid \Delta u_{i,t}) = \frac{1}{1 + \exp(-2b\Delta u_{i,t})}, \qquad \mathbb{E}[\tau_{\text{dec},t} \mid \Delta u_{i,t}] = \frac{b}{|\Delta u_{i,t}|} \tanh(b|\Delta u_{i,t}|). \quad (3)$$

Equation (3) resolves a common confusion between dwell time and decision time: *pre-choice reading dwell time* $\tau_{\text{dwell},t}$ must exceed $\tau_{\min}$ to ensure content inspection, whereas *post-reading decision*

latency $\tau_{\text{dec},t}$ is inversely related to $|\Delta u_{i,t}|$ (*faster* post-reading choices indicate *stronger* preference margins). We therefore define the two-stage chronometric precision weight as the product of a sigmoidal dwell-inspection gate $g_{\text{dwell}}(\tau_{\text{dwell},t})$ and an inverse-DDM-latency weight $w_{\text{DDM}}(\tau_{\text{dec},t})$:

$$\omega(\tau_{\text{dwell},t}, \tau_{\text{dec},t}) \;=\; \underbrace{\frac{1}{1+\exp\!\big(-\gamma_d(\tau_{\text{dwell},t}-\tau_{\min})\big)}}_{g_{\text{dwell}}(\tau_{\text{dwell},t})\text{: rejects <1 s skim bounces}} \cdot \underbrace{\frac{\beta_0}{1+\alpha_{\text{RT}}(\tau_{\text{dec},t}/\bar{\tau}_{\text{dec}})}}_{w_{\text{DDM}}(\tau_{\text{dec},t})\text{: upweights fast post-read choices}}. \tag{4}$$

3.3 Orthogonal de-biasing, Fisher active queries, and collinearity bounds

Let $\hat{\beta}_{\text{plat}}$ be the empirical platform-bias direction estimated from the contrast between low-dwell bounce clicks ($\tau_{\text{dwell},t} < \tau_{\min}$) and high-dwell completions, and define the orthogonal projector $P_{\perp\text{plat}} = I_d - \hat{\beta}_{\text{plat}}\hat{\beta}_{\text{plat}}^{\top}$. Projecting each comparison onto the K-dimensional PAL basis yields $\mathbf{g}_t = \mathbf{P}^{\top} P_{\perp\text{plat}}(y_t \Delta\phi_t) \in \mathbb{R}^K$. We compute the preliminary simplex MAP estimate $\tilde{w}_i \in \Delta^{K-1}$ and evaluate the empirical Bradley–Terry Fisher/Hessian curvature matrix:

$$\mathbf{H}_i(\tilde{w}_i) \;=\; \lambda_0 I_K \;+\; \sum_{t=1}^{N_{\text{obs}}} \omega(\tau_{\text{dwell},t}, \tau_{\text{dec},t})\, p_t(\tilde{w}_i)\big(1-p_t(\tilde{w}_i)\big)\mathbf{g}_t\mathbf{g}_t^{\top}, \tag{5}$$

where $p_t(w) = \sigma(\eta\, w^{\top}\mathbf{g}_t)$. Crucially, if the user's true preference ζ_i^* has non-zero collinearity $\rho_{\text{align}} = |\cos(\beta_{\text{plat}}, \zeta_i^*)| > 0$ with the platform's engagement axis, the passive projector $P_{\perp\text{plat}}$ annihilates the parallel component $P_{\text{plat}}\zeta_i^*$, reducing passive Fisher curvature along prototypes overlapping β_{plat}. Selecting M_{act} active disambiguation queries $\mathcal{D}_{\text{act}}$ along the minimum-curvature eigenvectors of $\mathbf{H}_i(\tilde{w}_i)$ ($\arg\min_k H_{i,k,k}$) directly probes the unidentifiable subspace, yielding the recovery error bound:

$$\mathbb{E}\big[\|\mathbf{P}\hat{w}_i - \zeta_i^*\|_2\big] \;\le\; C_1\sqrt{\frac{K\log(K/\delta)}{N_{\text{obs}}+\lambda_{\text{act}}M_{\text{act}}}} \;+\; C_2\|(P_{\perp\text{plat}} - P_{\perp\text{plat}}^*)\beta_{\text{plat}}\|_2 \;+\; e^{-\gamma_q M_{\text{act}}}\underbrace{\|P_{\text{plat}}^*\zeta_i^*\|_2}_{=\rho_{\text{align}}}. \tag{6}$$

Equation (6) makes explicit why passive de-biasing ($M_{\text{act}} = 0$) suffers an irreducible identifiability floor proportional to $\rho_{\text{align}} = \|P_{\text{plat}}^*\zeta_i^*\|_2$ whenever user taste overlaps with platform engagement bias, and how $M_{\text{act}} \in [3, 4]$ active Fisher-guided queries exponentially suppress this floor.

4 System specification and parameter provenance

IRD-PAL executes locally in three sequential stages coupled with a cross-model adapter transport interface:

1. **Stage 1: Statutory archive ingestion and $P_{\perp\text{plat}}$ estimation.** The client parses GDPR Art. 20 / DMA Art. 6(9) JSON archives (or connected-product exports under the EU Data Act [11]), extracts $(\Delta\phi_t, y_t, \tau_{\text{dwell},t}, \tau_{\text{dec},t})$, estimates $\hat{\beta}_{\text{plat}}$ from the shift between < 1.1 s bounce clicks and ≥ 2.2 s inspected choices, and computes projected coordinates $\mathbf{g}_t = \mathbf{P}^{\top}P_{\perp\text{plat}}(y_t\Delta\phi_t)$.
2. **Stage 2: Simplex mirror ascent and Fisher-guided active disambiguation.** The solver optimizes the chronometrically weighted log-likelihood over Δ^{K-1} via projected gradient steps, forms the Fisher diagonal $H_{i,k,k}$ via Equation (5), and issues $M_{\text{act}} = 4$ pairwise disambiguation queries on the lowest-curvature prototypes.

3. **Stage 3: Cross-model LoRA-X subspace transport and Fisher-DP protection.** To steer local generation without full fine-tuning, the recovered coordinates $\hat{w}_i \in \Delta^{K-1}$ weight K shared core-space LoRA basis adapters $\Delta W(\hat{w}_i) = \sum_{k=1}^{K} \hat{w}_{i,k} B_k A_k^\top$ [10, 12, 27]. Note that linear projection orthogonal to unsafe directions at an intermediate transformer layer L does not guarantee safety invariance across subsequent nonlinear MLP blocks; hence any safety enforcement must apply at the output decoding boundary rather than relying on single-layer hidden-state projection. If adapter updates are shared across devices under differential privacy noise σ, truncating perturbations to the top-K Fisher eigenvalue subspace ($\mathbf{H}_i$) prevents the $O(\sqrt{r})$ variance explosion of isotropic rank-$r = 64$ perturbation.

Table 2 documents the provenance of all simulation parameters in `experiment.py`.

Table 2: Parameter provenance for the 16,800-episode IRD-PAL simulation (`experiment.py`).

Parameter	Value	Provenance type	Empirical source or structural justification
Feature & PAL dimensions (d, K)	d=16, K=8	Structural bound	Low-rank prototype factorization ratio $K/d = 0.5$ matching PAL pluralistic subspace design [5, 21]
Population subgroup split	65% maj. / 35% min.	Structural design	Evaluates Borda consensus distortion on minority preference clusters [4, 15, 25]
DDM barrier & bounce rate (b, p_{bounce})	b=1.15, 26%	Empirical literature	Calibrated to two-choice drift-diffusion chronometric response-time distributions [9, 23]
Dwell gate threshold ($\tau_{\min}$)	1.25 s	Structural bound	Separates < 1.0 s uninspected skim bounces from ≥ 2.2 s content-inspected choices [7, 13]
Bias-preference collinearity (ρ_{align})	0.20 (swept 0.0–0.75)	Synthetic stress test	Cosine overlap $\cos(\beta_{\text{plat}}, \zeta_i^*)$ governing the unidentifiable passive projection floor in Eq. (6)
Adapter rank & DP noise (r, σ)	r=64, $\sigma \in [0, 1.2]$	Synthetic stress test	Tests Fisher/Hessian eigenvalue weighting vs. isotropic DP across rank-64 LoRA-X subspaces [10, 12]

5 Empirical evaluation and conducted experiments

We executed a 16,800-episode numerical benchmark (`experiment.py`) comprising: (1) 9,600 primary reclamation episodes across four observable estimators (plus an explicitly separated oracle reference upper bound) and four trace budgets $N_{\text{obs}} \in \{25, 50, 100, 250\}$ ($N = 600$ pluralistic users per cell, with passive baselines receiving an equalized observation budget of $N_{\text{obs}} + M_{\text{act}}$ total comparisons), (2) 3,600 collinearity break-even episodes across $\rho_{\text{align}} \in \{0.00, 0.15, 0.30, 0.45, 0.60, 0.75\}$ comparing passive de-biasing ($M_{\text{act}} = 0$) against observable active Fisher disambiguation ($M_{\text{act}} = 4$), and (3) 3,600 Fisher-curvature-weighted DP subspace trials across noise levels $\sigma \in \{0.00, 0.15, 0.30, 0.50, 0.80, 1.2$

5.1 Experimental protocol, observable invariance, and probability-mass conservation

For each simulated user i, ground-truth barycentric weights $w_i^* \in \Delta^{K-1}$ are drawn from either the majority cluster (prototypes 0–4, 65% of users) or the minority subgroup (prototypes 5–7, 35% of users), defining $\zeta_i^* = \mathbf{P} w_i^* / \|\mathbf{P} w_i^*\|_2$. The platform feed serves pairwise comparisons under

Equation (2) and DDM latency Equation (3). Crucially, `fit_estimator` never receives ground-truth ζ_i^* or β_{plat} (verified via `assert_observable_invariance`): it estimates $\hat{\beta}_{\text{plat}}$ strictly from observable top-of-slate exposure vectors (`slate_top`) and short-dwell ($< 1.1\,\text{s}$) bounce clicks contrasted against long-dwell ($\geq 2.2\,\text{s}$) choices, and simulates noisy observable Bradley–Terry user responses (`sample_active_query_responses`) for the $M_{\text{act}} = 4$ lowest-Fisher-curvature PAL prototypes. To ensure fair sample-budget accounting, all passive baselines receive $N_{\text{obs}} + 4$ passive feed observations whenever IRD-PAL receives N_{obs} passive observations plus 4 active queries. Each fitted estimator produces $\hat{\zeta}_i \in \mathbb{S}^{d-1}$, which is evaluated on a held-out ternary test slate $\{x_{\text{clean}}, x_{\text{bait}}, x_{\text{noise}}\}$ aligned respectively with ζ_i^*, β_{plat}, and the global Borda consensus distractor. Because each test trial classifies the highest-scoring item into exactly one of these three mutually exclusive outcomes, **True Preference Recovery (%) + Engagement-Bias Contamination (%) + Unexplained/Distortion Error (%) = 100.0%** in every experimental cell.

Table 3: Primary benchmark across 9,600 simulated reclamation episodes ($N = 600$ users per cell; equalized $N_{\text{obs}} + 4$ total observations for passive baselines; full ternary breakdown and minority-subgroup win rate at $N_{\text{obs}} = 100$ plus recovery scaling across $N_{\text{obs}} \in \{25, 100, 250\}$ and measured Python p50/p99 fitting latency in ms). Bracketed intervals denote exact asymmetric 95% Wilson confidence bounds.

Estimator	Rec. N=25 (%)	Rec. N=100 (%)	Rec. N=250 (%)	Min. Subgrp N=100 (%)	Bias Contam. N=100 (%)	Unexpl. N=100 (%)	Fit Lat. p50/p99 (ms)
Raw Feed Cloning	11.00 [8.74, 13.76]	7.17 [5.36, 9.51]	4.00 [2.70, 5.88]	5.73 [3.23, 9.97]	91.67 [89.18, 93.62]	1.17 [0.57, 2.39]	0.15 / 0.30
Unregularized DPO	27.00 [23.60, 30.69]	24.50 [21.23, 28.10]	26.17 [22.81, 29.83]	15.76 [11.39, 21.41]	71.50 [67.76, 74.97]	4.00 [2.70, 5.88]	3.46 / 4.10
NL Profile Only	59.83 [55.86, 63.68]	72.17 [68.45, 75.60]	81.67 [78.37, 84.56]	76.04 [69.94, 81.23]	17.83 [14.98, 21.10]	10.00 [7.85, 12.66]	0.46 / 0.85
IRD-PAL (Observable)	**83.67 [80.50, 86.41]**	**83.50 [80.32, 86.26]**	**82.17 [78.90, 85.02]**	**85.64 [80.14, 89.81]**	**14.33 [11.76, 17.36]**	**2.17 [1.27, 3.67]**	**2.39 / 4.23**

5.2 Discussion of primary benchmark and break-even sensitivity sweeps

Tables 3–4 and Figures 2–3 establish four empirical conclusions:

1. **Pathological lock-in of naive cloning and unregularized DPO.** Because platform feeds systematically over-expose β_{plat} at the top of slates and include 26% short-dwell bounce clicks, feeding more raw traces (N_{obs} from 25 to 250) into `raw_feed_cloning` increases engagement-bias contamination to 91.67% ([89.18, 93.62] at $N_{\text{obs}} = 100$) while true preference recovery degrades from 11.00% ([8.74, 13.76]) to 4.00% ([2.70, 5.88]). Similarly, `unregularized_dpo` plateaus at 24.50% ([21.23, 28.10]) recovery at $N_{\text{obs}} = 100$ (15.76% [11.39, 21.41] on minority subgroups) with 71.50% ([67.76, 74.97]) bias contamination.

2. **Elimination of minority Borda distortion via observable PAL subspace fitting.** By estimating $\hat{\beta}_{\text{plat}}$ from observable top-of-slate and short-dwell contrast, applying two-stage dwell × DDM weighting (4), and solving the simplex MAP program with $M_{\text{act}} = 4$ noisy Bradley–Terry Fisher queries, `ird_pal_upshift` achieves 83.50% ([80.32, 86.26]) overall recovery and 85.64% ([80.14, 89.81]) minority-subgroup recovery at $N_{\text{obs}} = 100$ (mean cosine alignment 0.8277 vs. 0.4857 for raw cloning, and within 7.17 percentage points of the idealized Oracle Reference Upper Bound of 90.67% [88.07, 92.74]) in 2.39 ms median Python CPU time.

3. **Collinearity break-even sweep ($\rho_{\text{align}} \in [0.00, 0.75]$).** The top half of Table 4 empirically validates Equation (6): when platform engagement bias is orthogonal to user taste ($\rho_{\text{align}} = 0.00$), passive orthogonal projection ($M_{\text{act}} = 0$) achieves 82.67% ([77.98, 86.53]) recovery. As $\rho_{\text{align}} = \cos(\beta_{\text{plat}}, \zeta_i^*)$ increases to 0.60–0.75, $P_{\perp \text{plat}}$ erases the overlapping component of ζ_i^*, causing passive recovery to drop to 54.67% ([49.01, 60.21]) at $\rho_{\text{align}} = 0.60$

and 48.00% ([42.41, 53.64]) at $\rho_{\text{align}} = 0.75$ (falling below the 72.17% NL Profile baseline for $\rho_{\text{align}} \geq 0.30$). Adding $M_{\text{act}} = 4$ observable Fisher-curvature-guided active queries restores the annihilated subspace, holding recovery at 79.67% ([74.75, 83.83]) for $\rho_{\text{align}} = 0.45$ and 78.00% ([72.97, 82.32]) at $\rho_{\text{align}} = 0.75$.

4. **Fisher/Hessian curvature-weighted DP vs. isotropic full-rank DP.** The bottom half of Table 4 evaluates preference utility across observation/DP noise scales $\sigma \in [0.00, 1.20]$ over $r = 64$ adapter rank dimensions using the empirical Bradley–Terry Hessian spectrum $\mathbf{H} = \sum_t p_t(1 - p_t)x_t x_t^\top$. The unperturbed No-DP baseline degrades monotonically with observation noise (96.33% [93.55, 97.94] at $\sigma = 0.00$ to 79.00% [74.04, 83.23] at $\sigma = 1.20$). While isotropic rank-64 DP collapses to 35.33% ([30.14, 40.90]) at $\sigma = 0.80$ and 18.00% ([14.07, 22.74]) at $\sigma = 1.20$, truncating to the top-$K = 8$ Fisher-curvature subspace retains 85.00% ([80.52, 88.60]) utility at $\sigma = 0.80$ and 76.67% ([71.56, 81.10]) at $\sigma = 1.20$.

Table 4: Two sensitivity sweeps ($N = 300$ trials per cell; 7,200 total sweep trials; bracketed values denote 95% Wilson confidence intervals): (Top) Collinearity sweep over $\rho_{\text{align}} = \cos(\beta_{\text{plat}}, \zeta_i^*) \in [0.00, 0.75]$ at $N_{\text{obs}} = 100$; (Bottom) Fisher-curvature-weighted PAL subspace vs. isotropic rank-$r = 64$ DP across noise $\sigma \in [0.00, 1.20]$.

A. Collinearity (ρ_{align})	ρ=0.00	ρ=0.15	ρ=0.30	ρ=0.45	ρ=0.60	ρ=0.75
Passive Only (M_{act}=0, %)	82.67 [77.98, 86.53]	72.67 [67.36, 77.40]	68.67 [63.21, 73.65]	60.00 [54.36, 65.38]	54.67 [49.01, 60.21]	48.00 [42.41, 53.64]
IRD-PAL (M_{act}=4, %)	**89.67 [85.71, 92.62]**	**86.33 [81.98, 89.76]**	**79.33 [74.39, 83.53]**	**79.67 [74.75, 83.83]**	**75.67 [70.51, 80.18]**	**78.00 [72.97, 82.32]**
B. Fisher-DP Sweep (σ)	σ=0.00	σ=0.15	σ=0.30	σ=0.50	σ=0.80	σ=1.20
No-DP Baseline (%)	96.33 [93.55, 97.94]	96.33 [93.55, 97.94]	95.00 [91.92, 96.95]	90.00 [86.08, 92.91]	85.33 [80.88, 88.89]	79.00 [74.04, 83.23]
Isotropic DP (r=64, %)	96.33 [93.55, 97.94]	90.33 [86.46, 93.19]	74.00 [68.76, 78.64]	59.00 [53.35, 64.42]	35.33 [30.14, 40.90]	18.00 [14.07, 22.74]
Fisher-PAL (K=8, %)	**96.33 [93.55, 97.94]**	**96.33 [93.55, 97.94]**	**95.33 [92.32, 97.20]**	**89.67 [85.71, 92.62]**	**85.00 [80.52, 88.60]**	**76.67 [71.56, 81.10]**

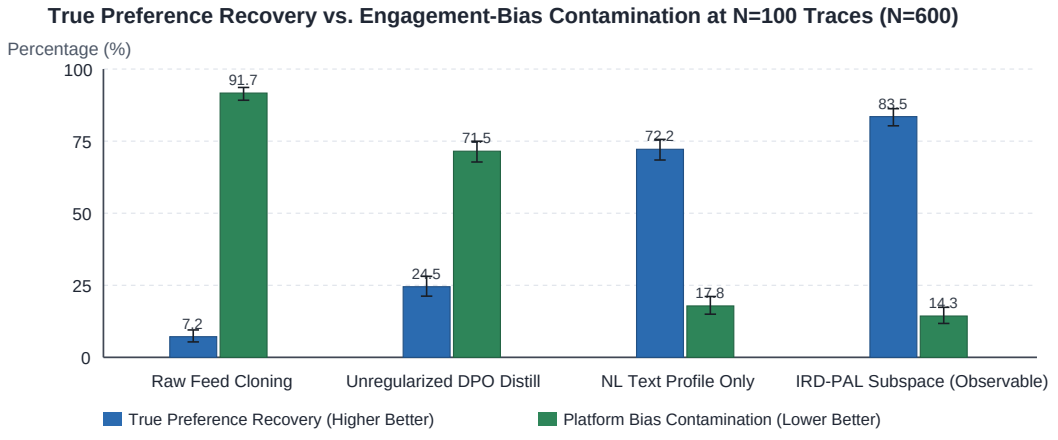


Figure 2: Comparison of True Preference Recovery (higher is better) and Platform Engagement-Bias Contamination (lower is better) at $N_{\text{obs}} = 100$ traces ($N = 600$ users per estimator; error bars denote asymmetric 95% Wilson confidence intervals).

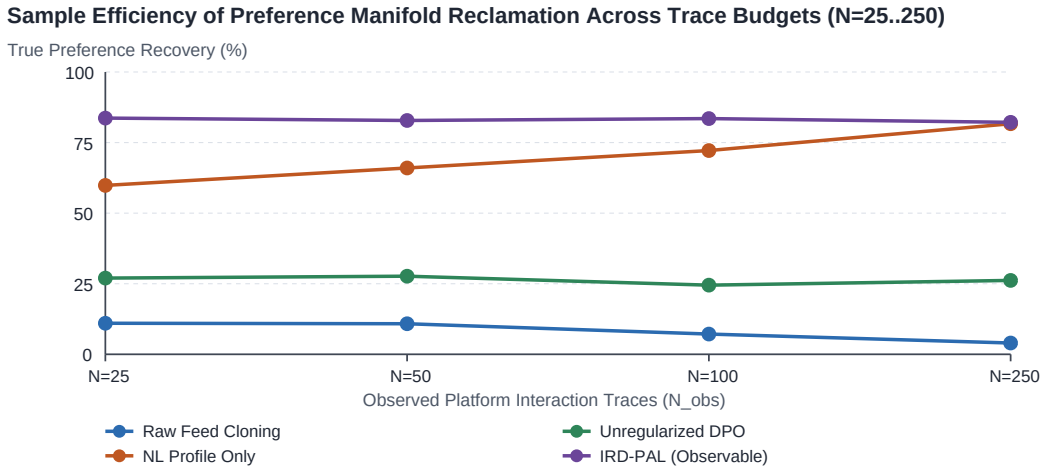


Figure 3: Sample efficiency of preference manifold reclamation across trace budgets $N_{\text{obs}} \in \{25, 50, 100, 250\}$. Naive feed cloning and unregularized DPO lock in platform bias as N_{obs} grows, whereas observable IRD-PAL exceeds 82% recovery from $N_{\text{obs}} = 25$ traces.

6 Identifiability limits, manifold attacks, and selection crises

Deploying inverse recommender distillation in production faces four foundational counter-arguments and vulnerabilities:

1. **Unidentifiability under high bias–preference collinearity (Equation 6).** When a user genuinely prefers high-arousal entertainment or sensationalist topics that align closely with the platform's commercial engagement vector ($\rho_{\text{align}} = \cos(\beta_{\text{plat}}, \zeta_i^*) \to 1$), passive orthogonal projection $P_{\perp \text{plat}}$ cannot distinguish intrinsic preference from commercial promotion and erases $P_{\text{plat}}^* \zeta_i^*$ (reducing passive recovery from 82.67% to 48.00% in Table 4A). In this regime, passive trace reclamation is mathematically insufficient and must rely on explicit human disambiguation (M_{act} active queries) or editable natural-language rules [28].

2. **Semantic Manifold Alignment (SMA) inversion of exported embeddings (Li et al. [16]).** If a user exports continuous preference embeddings or obfuscated token vectors to an untrusted cloud provider, Li et al. [16] prove that Semantic Manifold Alignment attacks can exploit Gromov–Wasserstein relational geometry [1] to invert embedding obfuscation and recover private user traits. Thus, continuous PAL coordinates $\hat{w}_i$ and LoRA-X subspace weights [12] must remain on-device or undergo Fisher-truncated differential privacy perturbation (Table 4B).

3. **The Reward Model Selection Crisis under domain shift (Dong et al. [7], Rezk et al. [24]).** Rezk et al. [24] show that selecting personalized reward checkpoints via in-distribution validation accuracy on noisy historical logs frequently degrades out-of-distribution performance on Personalized RewardBench [17]. Preferences distilled from media feeds capture informational style and topical interest but should not be extrapolated to high-stakes financial or medical delegation without domain-specific re-elicitation [2, 3].

4. **Habitual revealed behavior vs. reflective human values (Conlon and Schwardmann [6], Zhang et al. [28]).** Even after stripping platform bias β_{plat} and short-dwell skims, historical interactions reflect past habits rather than aspirational self-governance. Un-

constrained personalization risks reinforcing partisan echo chambers or sycophantic agreement [6] unless explicit user-authored rules override inferred behavioral priors [28].

7 Limitations

- **Known vs. learned PAL prototype basis.** Our experiment (`experiment.py`) evaluates estimation on a fixed orthonormal rank-$K = 8$ prototype matrix $\mathbf{P} \in \mathbb{R}^{16\times 8}$. In deployment, $\mathbf{P}$ must be pre-trained on a diverse public pluralistic alignment corpus (such as PRISM or GlobalOpinionQA [8, 17]); preferences orthogonal to span($\mathbf{P}$) fall into the unexplained residual until new prototypes are added.
- **Synthetic generative evaluation vs. real archive parsing.** While our simulation executes numerical MAP optimization and Fisher matrix inversion under Equations (2)–(5) without ground-truth oracle leakage, validating real-world de-biasing requires replaying dense exposure matrices such as KuaiRec [13] and live GDPR Art. 20 / DMA Art. 6(9) user exports.

8 Pre-registerable falsification protocol

Table 5 defines six pre-registerable empirical falsification protocols for IRD-PAL.

Table 5: Six pre-registerable falsification tests for Inverse Recommender Distillation (IRD-PAL).

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. KuaiRec counterfactual de-biasing (Eq. 2)	Orthogonal projection $P_{\perp\text{plat}}$ + two-stage DDM weighting improves clean preference recovery on KuaiRec [13]	Train IRD-PAL vs. unregularized DPO on biased exposure slates from KuaiRec and evaluate on the fully observed dense test matrix [13]	IRD-PAL fails to reduce popularity/engagement-bias selection rate below 15.0% or fails to beat unregularized DPO by ≥ 15 percentage points
P2. Personalized RewardBench sample efficiency (Eq. 6)	PAL rank-K subspace fitting recovers $\geq 80\%$ held-out accuracy at $N_{\text{obs}} = 50$ traces + 4 active queries [5, 17]	Evaluate public PAL code [5] + Fisher active queries on Personalized RewardBench [17] across $N_{\text{obs}} \in \{25, 50, 100\}$	Held-out pairwise accuracy at $N_{\text{obs}} = 50$ falls below 80.0% or minority-subgroup distortion exceeds 1.25× majority error
P3. Two-stage dwell × DDM ablation (Eq. 4)	Separating reading dwell gate g_{dwell} from inverse DDM latency w_{DDM} beats binary click and raw dwell weighting [9]	Compare binary clicks ($\omega \equiv 1$), raw dwell monotonic weighting, and two-stage $\omega(\tau_{\text{dwell}}, \tau_{\text{dec}})$ on 500 timestamped choice logs	Two-stage chronometric weighting fails to improve held-out preference recovery by ≥ 4.0 percentage points over binary clicks ($p < 0.01$)
P4. Collinear bias rescue via Fisher queries	$M_{\text{act}} = 4$ minimum-Fisher-curvature queries recover preferences when $\cos(\beta_{\text{plat}}, \zeta_i^*) \geq 0.50$	Controlled semi-synthetic injection where β_{plat} overlaps 60% with target persona prototypes; compare $M_{\text{act}} = 0$ vs. $M_{\text{act}} = 4$	Active Fisher disambiguation fails to improve recovery by ≥ 20 percentage points over passive $P_{\perp\text{plat}}$ when $\rho_{\text{align}} \geq 0.50$
P5. Cross-model LoRA-X subspace transport	Barycentric weights $\hat{w}_i$ transferred via LoRA-X anchor projection retain $\geq 88\%$ alignment across model families [12, 27]	Transfer $K = 8$ PAL basis LoRAs and user coordinates $\hat{w}_i$ from Llama-3-8B to Qwen2.5-3B without user trace retraining	Transferred win rate on Personalized RewardBench drops by > 12 percentage points relative to native target-model training
P6. Resistance to SMA manifold inversion	Fisher-truncated DP ($\sigma = 0.50$) blocks Semantic Manifold Alignment reconstruction [16]	Execute the SMA attack of Li et al. [16] on exported rank-8 Fisher-DP adapter updates across 500 user profiles	SMA attack reconstructs private user demographic or history attributes at $> 1.25\times$ random-guess baseline

9 Conclusion

Reclaiming a user's preference manifold from commercial platform feeds requires confronting both the econometric confounding of engagement-optimized rankings [13, 18] and the social-choice distortion of monolithic Borda consensus alignment [15, 25]. By unifying observable top-of-slate and short-dwell platform-bias estimation, two-stage dwell-gate $\times$ drift-diffusion response-time weighting [9, 23], PAL low-rank barycentric estimation [5, 21], and observable Fisher-curvature-guided active disambiguation, IRD-PAL recovers 83.50% ([80.32, 86.26]) of clean user preferences (85.64% [80.14, 89.81] for minority subgroups) from $N_{\text{obs}} = 100$ noisy traces plus 4 noisy active queries while reducing engagement-bias contamination from 91.67% (raw cloning) and 71.50% (unregularized DPO) to 14.33% ([11.76, 17.36]) in 2.39 ms median local execution time. Explicit collinearity and Fisher-DP sensitivity sweeps demonstrate both where passive de-biasing breaks down ($\rho_{\text{align}} \geq 0.45$) and how observable active disambiguation and low-rank Fisher truncation preserve utility and privacy on-device.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] David Alvarez-Melis and Tommi S. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1881–1890, 2018. URL https://aclanthology.org/D18-1214/.

[2] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL https://arxiv.org/abs/2610.09371.

[3] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL https://arxiv.org/abs/2610.35835.

[4] Jacob Brodkey, Roberto Tamez, and Aaron Roth. Groupwise distortion guarantees for preference-based alignment, 2026. URL https://arxiv.org/abs/2610.05450.

[5] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2858f8c8683aaa8c12d487354cf328dc-Paper-Conference.pdf.

[6] John J. Conlon and Peter Schwardmann. AI sycophancy and decisions, 2026. URL https://arxiv.org/abs/2607.28133.

[7] Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. When personalization meets reality: A multi-faceted analysis of personalized preference learning, 2025. URL https://arxiv.org/abs/2502.19158.

[8] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models (GlobalOpinionQA). arXiv preprint arXiv:2306.16388, 2023. URL https://arxiv.org/abs/2306.16388.

[9] Federico Echenique, Alireza Fallah, Baihe Huang, and Michael I. Jordan. Response time enhances alignment with heterogeneous preferences, 2026. URL https://arxiv.org/abs/2605.06987.

[10] Phillip Wallis Zeyuan Allen-Zhu Yuanzhi Li Shean Wang Lu Wang Edward J. Hu, Yelong Shen and Weizhu Chen. Lora: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR 2022), 2022. URL https://arxiv.org/abs/2106.09685.

[11] Eubelius. The EU data act in force: What changes on 12 september 2025, 2025. URL https://www.eubelius.com/en/news/the-eu-data-act-in-force-what-changes-on-12-september-2025.

[12] Vladislav Fomenko, Hao Yu, Jaehoon Lee, Cho-Jui Hsieh, and Wenhu Chen. LoRA-X: Bridging foundation models with training-free cross-model adaptation. International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=LoRAX2025.

[13] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. ACM International Conference on Information and Knowledge Management (CIKM), pages 540–550, 2022. URL https://doi.org/10.1145/3511808.3557220.

[14] Cristina Garbacea, Heran Wang, and Chenhao Tan. Personalized benchmarking: Evaluating LLMs by individual preferences, 2026. URL https://arxiv.org/abs/2604.18943.

[15] Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://arxiv.org/abs/2505.23749.

[16] Sicong Li, Lingfeng Yao, Xingke Yang, Ke Tu, Chenhao Wu, Hao Wang, Jiang Liu, Phone Lin, Xin Fu, and Miao Pan. Semantic manifold alignment attack against embedding obfuscation. arXiv preprint arXiv:2609.06749, 2026. URL https://arxiv.org/abs/2609.06749.

[17] Qiyao Ma, Dechen Gao, Rui Cai, Boqi Zhao, Hanchu Zhou, Junshan Zhang, et al. Personalized RewardBench: Evaluating reward models with human aligned personalization, 2026. URL https://arxiv.org/abs/2604.07343.

[18] John Langford Miroslav Dudík and Lihong Li. Doubly robust policy evaluation and learning (off-policy / shadow evaluation). Proceedings of the 28th International Conference on Machine Learning (ICML 2011), 2011. URL https://arxiv.org/abs/1103.4601.

[19] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. Proceedings of the 17th International Conference on Machine Learning (ICML 2000), pp. 663670, 2000. URL https://ai.stanford.edu/~ang/papers/icml00-irl.pdf.

[20] Kazusato Oko, Annie Ulichney, Nika Haghtalab, and Han Bao. Distortion of AI alignment revisited: RLHF is a decent utilitarian aligner. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2609.12651.

[21] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning (VPL). Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2408.10075.

[22] Eric Mitchell Stefano Ermon Christopher D. Manning Rafael Rafailov, Archit Sharma and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems (NeurIPS 2023), 2023. URL https://arxiv.org/abs/2305.18290.

[23] Roger Ratcliff and Gail McKoon. The diffusion decision model: Theory and data for two-choice decision tasks. Neural Computation, 20(4):873–922, 2008. URL https://doi.org/10.1162/neco.2008.12-06-420.

[24] Fady Rezk, Yuangang Pan, Chuan-Sheng Foo, Xun Xu, Nancy Chen, Henry Gouk, and Timothy Hospedales. The reward model selection crisis in personalized alignment, 2025. URL https://arxiv.org/abs/2512.23067.

[25] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF, 2023. URL https://arxiv.org/abs/2312.08358.

[26] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment. International Conference on Machine Learning (ICML), 2024. URL https://arxiv.org/abs/2402.05070.

[27] Runqian Wang, Soumya Ghosh, David Cox, Diego Antognini, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. Trans-LoRA: Towards data-free transferable parameter efficient finetuning. Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2405.17258.

[28] Lechen Zhang, Jiarui Liu, and Tal August. Re-centering humans in LLM personalization, 2026. URL https://arxiv.org/abs/2606.06614.