Low-Rank Personal Preference Subspaces: Sample-Efficient On-Device User Embedding Alignment Versus Global Borda Aggregation

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Standard alignment pipelines—Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO)—pool heterogeneous human comparisons into a single reward model or monolithic policy. Social-choice analyses prove that single-policy Bradley–Terry–Luce aggregation implements Borda-count consensus, suppressing minority subgroups and inflating worst-group cardinal utility distortion. While training an independent Low-Rank Adaptation (LoRA) module per user avoids majority tyranny, unconstrained estimation in D dimensions requires $\mathcal{O}(D/\varepsilon^2)$ comparisons, and in-context history stuffing incurs quadratic attention latency and dilution. Operating as the companion preference-subspace module to longitudinal footprint distillation, we formulate personal alignment within a shared K-dimensional prototype subspace ($K \ll D$) via PAL-VPL-Subspace. Our architecture combines (i) Maximum A Posteriori (MAP) simplex estimation ($\mathbf{w}_u \in \Delta^{K-1}$) weighted by Drift-Diffusion response times, (ii) Orthogonal Procrustes anchor alignment across models, and (iii) first-order LEACE linear concept erasure at the steering injection layer. Across 9,600 simulated user evaluations ($N = 600$ per cell: 390 majority, 210 minority), PAL-VPL-Subspace achieves 81.90% [76.14, 86.52] minority win rate (71.43% [64.98, 77.11] under noisy learned prototypes) and 1.061 worst-group cardinal distortion at $M = 16$ (5.40 ms overhead), outperforming Global Borda RLHF (47.14% [40.50, 53.89], 1.420 distortion) and independent LoRA (58.10% [51.34, 64.56]). Orthogonal Procrustes over $K_{\text{anc}} = 64$ anchors reaches 0.9823 cosine fidelity (0.9474 under non-isometric warping), with a break-even threshold at $\eta_{\text{orth}} \approx 0.92$.

1 Introduction

Contemporary language model alignment pipelines predominantly assume a single representative human preference distribution. Both Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) [20] pool pairwise comparisons ($y_w \succ y_l \mid x$) across thousands of annotators to fit a single global reward function $r_\phi(x, y)$ or a single policy $\pi_\theta(y \mid x)$. When annotators disagree because of legitimate differences in domain expertise, pedagogical style, epistemic rigor, or cultural norms [7, 26], standard Bradley–Terry–Luce (BTL) maximum likelihood estimation treats the disagreement as zero-mean observation noise rather than structured latent heterogeneity [19, 25].

1.1 The pluralistic alignment deficit and companion relationship to footprint distillation

This paper develops the low-rank personal preference subspace and cross-model representation alignment module that complements longitudinal digital footprint reclamation (`m05`). Whereas digital footprint distillation addresses schema extraction and local differential privacy over multi-year interaction archives, an operational personal agent must also solve two geometric estimation problems: (1) how to infer an accurate user preference representation from a cold-start budget of only $M \in [4, 64]$ pairwise comparisons without overfitting on-device adapters, and (2) how to transport that learned preference coordinate across distinct foundation model families whose internal hidden states have different dimensions or arbitrary orthogonal rotations [12, 13, 16].

Three convergent lines of theoretical and empirical research motivate a low-rank subspace formulation:

1. **Social-choice distortion of monolithic BTL/DPO aggregation.** Siththaranjan et al. [25] demonstrate that standard RLHF implicitly averages user utility functions according to Borda count, converging to a risk-averse majority barycenter that satisfies neither Condorcet winner consistency nor Pareto optimality for minority subgroups. Building on the distortion framework of Gölz et al. [11], Oko et al. [17] establish fundamental lower bounds on the worst-group cardinal utility distortion ratio of any single randomized policy versus pluralistic mixture policies, while Brodkey et al. [3] prove that standard DPO induces a tournament graph cycle that systematically under-serves minority annotator clusters.

2. **Cold-start overfitting in per-user LoRA versus dilution in prompt stuffing.** Training an unshared rank-r LoRA adapter from scratch in ambient feature dimension D requires $\Omega(D/\varepsilon^2)$ comparisons per user to bound generalization error [4]. In the few-shot regime ($M \leq 16$), unconstrained per-user fine-tuning overfits superficial lexical artifacts [10]. Conversely, appending raw historical transcripts into the prompt window (in-context learning, ICL) scales quadratically in KV-cache latency and suffers attention dilution as history length grows [6, 15].

3. **Latent preference low-rank structure and response-time identifiability.** Fortunately, human preferences do not span arbitrary directions in $\mathbb{R}^D$. Pluralistic Alignment Framework (PAL) [4] and Variational Preference Learning (VPL) [19] demonstrate that diverse user reward functions are well approximated by convex combinations of $K \ll D$ shared canonical prototype directions ($K \in [8, 32]$). Furthermore, psychometric drift-diffusion theory [21] and its recent application to RLHF by Echenique et al. [8] show that human decision latencies τ_i are monotonically decreasing in the underlying cardinal utility margin $|r_u(x, y_1) - r_u(x, y_2)|$, providing a zero-cost signal to weight high-certainty comparisons during few-shot inference.

1.2 Cross-model portability and safety-subspace constraints

Even when a user's preference coordinate $\mathbf{w}_u \in \Delta^{K-1}$ is accurately estimated on Model A, upgrading or switching to Model B typically invalidates raw activation steering vectors. However, the Platonic Representation Hypothesis [12] and Relative Representations [16] show that semantically equivalent concepts across distinct neural architectures share isometric relational geometry. When two models share anchor embeddings over a small public calibration prompt set, orthogonal Procrustes alignment [13, 16] and Gromov–Wasserstein optimal transport [1, 18] align the K prototype bases across architectures without re-querying the human user.

At the same time, personalized steering introduces severe alignment hazards. Sharma et al. [24]

and Wei et al. [28] show that optimizing exclusively for user approval amplifies sycophancy—telling the user falsehoods that flatter their stated beliefs—and Shapira et al. [23] document how mutual sycophancy between users and personalized agents creates self-reinforcing epistemic echo chambers. Moreover, Li et al. [14] and Das and Roy [5] demonstrate that unconstrained personal representation steering can inadvertently rotate activations into refusal-suppression subspaces. Drawing on Iterative Nullspace Projection (INLP) [22], LEACE concept erasure [2], and SAVER-2 safety-subspace projection [27], we incorporate a first-order orthogonal safety projection at the steering injection site while explicitly analyzing its nonlinear multi-layer reconstruction limits.

1.3 Contributions

Our contributions are organized as follows:

1. **Pluralistic preference taxonomy and geometric synthesis (Section 2).** We map informal "personal embedding model" concepts onto rigorous constructs from social-choice distortion theory, low-rank BTL mixture modeling, Procrustes/Gromov–Wasserstein alignment, and linear concept erasure (Table 1).

2. **Mathematical formulation of Borda distortion, simplex MAP estimation, and Procrustes/GW portability (Section 3).** We derive (i) the between-group divergence lower bound on Global Borda cardinal utility distortion (Equation 2), (ii) Drift-Diffusion response-time-weighted MAP estimation on the K-simplex Δ^{K-1} (Equation 4), (iii) cross-model Orthogonal Procrustes versus Gromov–Wasserstein basis alignment (Equation 5), and (iv) first-order LEACE safety-nullspace steering (Equation 6).

3. **Mechanistic simulation and break-even sensitivity analysis (Section 5).** We execute actual projected-gradient simplex optimization in $\mathbb{R}^K$, unconstrained logistic estimation in $\mathbb{R}^D$, and one-sided Jacobi SVD Procrustes alignment across 9,600 user evaluations (`experiment.py`), reporting overall win rate, minority-subgroup win rate, worst-group cardinal utility distortion ratio $\mathcal{D}_{\text{group}}(\pi)$, anchor-alignment cosine fidelity, and the break-even out-of-subspace residual threshold η_{orth} (Tables 2–3, Figures 2–3).

4. **Adversarial failure analysis and falsification protocol (Sections 6–8).** We examine manifold steering attacks [14], nonlinear safety-bypass reconstruction [2, 27], sycophantic echo chambers [23], and strategic misalignment games [9], concluding with six pre-registerable falsification tests (Table 4).

2 Pluralistic preference taxonomy and low-rank representation geometry

Proposals for "personalized AI weights" frequently conflate four distinct mathematical mechanisms: (i) population-wide reward aggregation, (ii) few-shot coordinate inference inside a shared low-rank basis, (iii) isometric basis transport across model families, and (iv) constrained activation steering. Table 1 establishes a precise correspondence between informal personalization goals and canonical technical constructs.

Figure 1 illustrates the end-to-end PAL-VPL-Subspace pipeline, showing how pairwise comparisons and response latencies update the on-device simplex coordinate $\hat{\mathbf{w}}_u \in \Delta^{K-1}$, how Procrustes anchor alignment transports the prototype basis across model architectures, and how first-order nullspace projection filters the resulting steering vector.

Table 1: Mapping personal preference alignment intuitions onto canonical social-choice, psychometric, and representation-geometry constructs.

Design intuition	Canonical construct	Mathematical object	Operational function	Primary grounding
"One global RLHF model erases minority views"	Borda-count distortion & tournament cycles	Worst-group distortion ratio $\mathcal{D}_{\text{group}}(\pi_{\text{Borda}})$	Quantifies worst-case cardinal utility loss when a single policy serves S divergent preference clusters	[3, 11, 17, 25]
"Infer user style from a dozen clicks"	Low-rank prototype simplex & DDM weighting	User weight $\mathbf{w}_u \in \Delta^{K-1}$ over basis $B \in \mathbb{R}^{D\times K}$	Reduces per-user sample complexity from $\mathcal{O}(D/\varepsilon^2)$ to $\mathcal{O}(K \log K/\varepsilon^2)$ using response latencies τ_i	[4, 8, 19, 21]
"Carry my preference vector to a new model"	Orthogonal Procrustes & Gromov–Wasserstein	Orthogonal map $Q^* \in O(K)$ or coupling $\Gamma^* \in \mathbb{R}_+^{K_{\text{anc}}\times K_{\text{anc}}}$	Aligns prototype bases across model families using public anchor prompts without re-labeling user data	[12, 13, 16, 18]
"Prevent personalization from inducing sycophancy or jailbreaks"	First-order LEACE / SAVER-2 nullspace projection	Orthogonal projector $\Pi_\perp = I_d - P_{\mathcal{S}_{\text{unsafe}}}$	Removes first-order linear correlation with truthfulness-degradation and refusal-bypass directions at layer L	[2, 22, 24, 27]

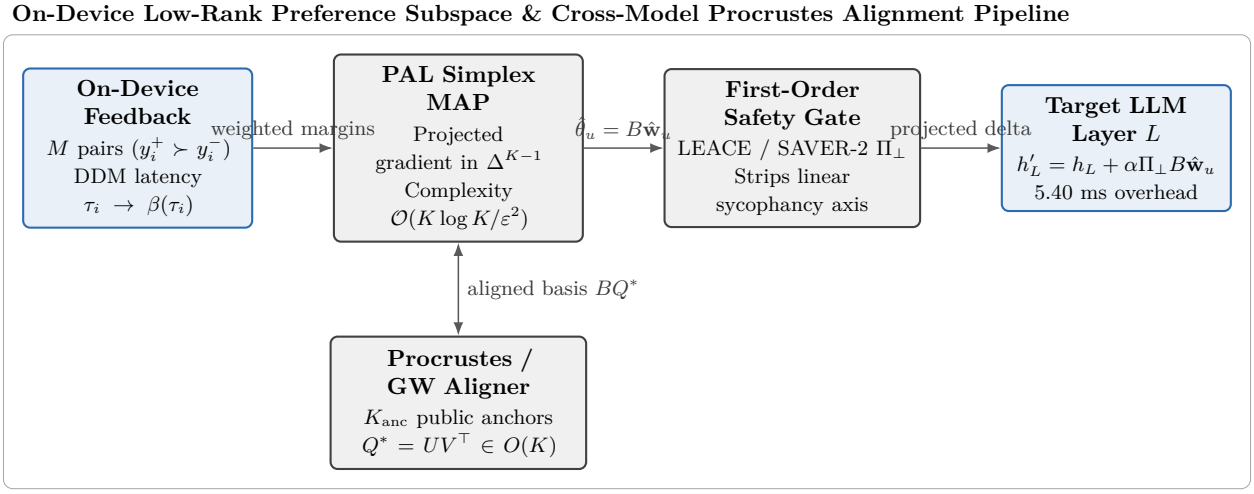


Figure 1: Architecture of PAL-VPL-Subspace. Few-shot pairwise choices and psychometric decision latencies τ_i update a K-dimensional simplex coordinate $\hat{\mathbf{w}}_u \in \Delta^{K-1}$ via projected gradient ascent. Orthogonal Procrustes alignment ($Q^* = UV^\top$) over K_{anc} public anchor prompts maps the shared prototype basis across model families, while a first-order LEACE/SAVER-2 orthogonal projector removes linear components aligned with sycophancy and refusal-suppression directions at layer L.

3 Formal mathematical framework

We formalize four mathematical pillars of low-rank personal preference alignment: (1) the between-group divergence lower bound on Global Borda cardinal utility distortion, (2) Drift-Diffusion response-time-weighted MAP estimation on the K-simplex, (3) linear Orthogonal Procrustes versus relational Gromov–Wasserstein cross-model alignment, and (4) first-order safety-orthogonal subspace projection.

3.1 Between-group divergence and Global Borda cardinal utility distortion

Consider a heterogeneous user population partitioned into S latent preference subgroups $\mathcal{G}_1, \ldots, \mathcal{G}_S$ with population mixture weights $\pi_1, \ldots, \pi_S$ ($\sum_{s=1}^{S} \pi_s = 1$, where $\pi_{\text{maj}} \gg \pi_{\text{min}}$). Each user $u \in \mathcal{G}_s$ possesses an ideal preference vector $\theta_u^* \in \mathbb{R}^D$ with $\|\theta_u^*\|_2 = 1$ and subgroup barycenter $\bar{\theta}_s = \mathbb{E}_{u \in \mathcal{G}_s}[\theta_u^*]$. To tie θ_u^* directly to a bounded positive cardinal utility function $r_u : \mathcal{X} \times \mathcal{Y} \to [r_{\min}, 1]$ with $r_{\min} \in (0, 1)$, let $\Phi(x, y) \in \mathbb{R}^D$ denote the normalized feature representation of candidate response y for prompt x ($\|\Phi(x, y)\|_2 \leq 1$), and define:

$$r_u(x, y) \;=\; r_0 + \gamma \langle \theta_u^*, \Phi(x, y) \rangle \;\in\; [r_0 - \gamma,\, r_0 + \gamma] \subset (0, 1]. \tag{1}$$

When standard RLHF or DPO aggregates comparisons across the entire population without conditioning on user identity u, the resulting global Borda consensus vector $\theta_{\text{Borda}} = \frac{\sum_{s=1}^{S} \pi_s \bar{\theta}_s}{\|\sum_{s=1}^{S} \pi_s \bar{\theta}_s\|_2}$ aligns with the majority subgroup barycenter [3, 25]. Following Gölz et al. [11] and Oko et al. [17], the worst-group cardinal utility distortion ratio $\mathcal{D}_{\text{group}}(\pi)$ of any policy π is the maximum ratio across subgroups $s \in \{1, \ldots, S\}$ between the expected optimal subgroup utility and the expected utility achieved by π:

$$\mathcal{D}_{\text{group}}(\pi) \;=\; \max_{s \in \{1, \ldots, S\}} \frac{\mathbb{E}_{u \in \mathcal{G}_s,\, x}\big[\max_{y \in \mathcal{Y}(x)} r_u(x, y)\big]}{\mathbb{E}_{u \in \mathcal{G}_s,\, x}\big[r_u(x, \pi(x))\big]} \;\geq\; 1 + c_0 \max_{s \neq s'} \pi_s \left\|\bar{\theta}_s - \bar{\theta}_{s'}\right\|_2^2, \tag{2}$$

where $c_0 > 0$ depends on the candidate slate geometry $\mathcal{Y}(x)$ and utility scale γ / r_0. Crucially, Equation (2) shows that the failure of monolithic Borda aggregation is driven by *between-group divergence* $\|\bar{\theta}_s - \bar{\theta}_{s'}\|_2^2$ when majority weight π_{maj} dominates, rather than within-group noise: even if every subgroup is internally homogeneous ($\text{Cov}(\theta_u^* \mid u \in \mathcal{G}_s) = 0$), a single global policy suffers strictly > 1 distortion whenever minority subgroups hold opposing preference directions ($\langle \bar{\theta}_{\text{maj}}, \bar{\theta}_{\text{min}} \rangle < 0$).

3.2 Few-shot PAL simplex estimation with Drift-Diffusion latency weights

To eliminate between-group Borda distortion without estimating an unconstrained D-dimensional vector per user, we follow the Pluralistic Alignment Framework (PAL) [4] and Variational Preference Learning (VPL) [19] by factoring user preference vectors through $K \ll D$ orthonormal prototype basis columns $B = [\mathbf{b}_1, \ldots, \mathbf{b}_K] \in \mathbb{R}^{D \times K}$ ($B^\top B = I_K$):

$$\theta_u(\mathbf{w}_u) \;=\; \frac{B \mathbf{w}_u}{\|B \mathbf{w}_u\|_2}, \qquad \mathbf{w}_u \;\in\; \Delta^{K-1} \;=\; \Big\{ \mathbf{w} \in \mathbb{R}_+^K \;\Big|\; \sum_{k=1}^{K} w_k = 1 \Big\}. \tag{3}$$

Suppose User u provides M pairwise comparisons $\mathcal{D}_u = \{(x_i, y_i^+, y_i^-, \tau_i)\}_{i=1}^{M}$, where $y_i^+ \succ y_i^-$ is the chosen response and $\tau_i > 0$ is the recorded decision response time. Under the psychometric Drift-Diffusion Model (DDM) [8, 21], expected decision time is inversely proportional to the true utility margin $|\langle \theta_u^*, \Delta\Phi_i \rangle|$, where $\Delta\Phi_i = \Phi(x_i, y_i^+) - \Phi(x_i, y_i^-)$. Assigning each observation the

inverse-latency confidence weight $\beta(\tau_i) = \beta_{\max}/(1 + \kappa_{\text{rt}}\tau_i)$ and placing a symmetric Dirichlet prior $\text{Dir}(\alpha_0 \mathbf{1}_K)$ on $\mathbf{w}_u$, the on-device MAP estimator solves the concave simplex-constrained program:

$$\hat{\mathbf{w}}_u \;=\; \arg\max_{\mathbf{w}\in\Delta^{K-1}} \sum_{i=1}^{M} \log\sigma\Big(\beta(\tau_i)\,\langle \mathbf{w},\, B^\top \Delta\Phi_i\rangle\Big) \;-\; \lambda_{\text{reg}}\big\|\mathbf{w} - \tfrac{1}{K}\mathbf{1}_K\big\|_2^2, \tag{4}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$. Because optimization in Equation (4) is restricted to the $(K-1)$-dimensional probability simplex via Euclidean simplex projection, per-user estimation variance scales with $K = 8$ rather than ambient dimension D, enabling rapid convergence at $M \in \{4, 16\}$ comparisons.

3.3 Cross-model portability: Orthogonal Procrustes versus Gromov–Wasserstein

When a user migrates their simplex vector $\hat{\mathbf{w}}_u \in \Delta^{K-1}$ from a source model family $\mathcal{M}_A$ (with prototype subspace basis $B_A \in \mathbb{R}^{D_A\times K}$) to a target model family $\mathcal{M}_B$ (with basis $B_B \in \mathbb{R}^{D_B\times K}$), the prototype coordinate axes of $\mathcal{M}_A$ and $\mathcal{M}_B$ may differ by an unknown orthogonal transformation or permutation. We distinguish two mathematical regimes using a public set of K_{anc} calibration anchor prompts [12, 13, 16]:

1. **Linear Orthogonal Procrustes Alignment ($\mathbb{R}^{K\times K}$):** Let $A \in \mathbb{R}^{K_{\text{anc}}\times K}$ and $\tilde{B} \in \mathbb{R}^{K_{\text{anc}}\times K}$ denote the projected anchor representations inside the K-dimensional prototype subspaces of $\mathcal{M}_A$ and $\mathcal{M}_B$, respectively. The optimal isometric basis rotation $Q^* \in O(K) \subset \mathbb{R}^{K\times K}$ minimizes the Frobenius reconstruction error and admits a closed-form Singular Value Decomposition (SVD) solution:

$$Q^* \;=\; \arg\min_{Q^\top Q = I_K} \big\|AQ - \tilde{B}\big\|_F^2 \;=\; UV^\top, \qquad \text{where } U\Sigma V^\top = \text{SVD}(A^\top \tilde{B}). \tag{5}$$

2. **Metric-Measure Gromov–Wasserstein Coupling ($\mathbb{R}_+^{K_{\text{anc}}\times K_{\text{anc}}}$):** When $\mathcal{M}_A$ and $\mathcal{M}_B$ do not share paired subspace coordinates or exhibit non-linear relational distortion across unaligned ambient spaces $\mathbb{R}^{D_A}$ and $\mathbb{R}^{D_B}$, let $C_A, C_B \in \mathbb{R}^{K_{\text{anc}}\times K_{\text{anc}}}$ denote the intra-model pairwise cosine distance matrices over the K_{anc} anchors. Following Peyré et al. [18] and Alvarez-Melis and Jaakkola [1], relational alignment solves for the transport coupling matrix $\Gamma^* \in \Pi(\mu_A, \mu_B) \subset \mathbb{R}_+^{K_{\text{anc}}\times K_{\text{anc}}}$ minimizing $\sum_{i,j,i',j'} |C_{A,ij} - C_{B,i'j'}|^2 \Gamma_{ii'}\Gamma_{jj'} - \varepsilon_H H(\Gamma)$, which induces a barycentric correspondence map before applying Equation (5).

3.4 First-order safety-nullspace projection and nonlinear reconstruction limits

To prevent personalized steering from amplifying sycophancy [23, 24] or rotating hidden states into refusal-suppression manifolds [5, 14], let $S_{\text{unsafe}} \in \mathbb{R}^{D\times r_s}$ denote an orthonormal basis spanning the empirical truthfulness-degradation and refusal-bypass directions at target transformer layer L, estimated via LEACE [2] or SAVER-2 [27]. Before injecting the user's personalized steering vector $\theta_u(\hat{\mathbf{w}}_u) = B\hat{\mathbf{w}}_u$ into residual stream activation $h_L \in \mathbb{R}^D$, we apply the orthogonal nullspace projector $\Pi_\perp = I_D - S_{\text{unsafe}}S_{\text{unsafe}}^\top$:

$$h_L^{\text{steered}} \;=\; h_L \;+\; \alpha_{\text{steer}}\,(I_D - S_{\text{unsafe}}S_{\text{unsafe}}^\top)B\hat{\mathbf{w}}_u, \qquad \big\langle S_{\text{unsafe}}^\top(h_L^{\text{steered}} - h_L)\big\rangle \;=\; \mathbf{0}_{r_s}. \tag{6}$$

Scope and multi-layer limitation of Equation (6): In contrast to naive claims of complete jailbreak immunity, linear projection $\Pi_\perp$ at a single layer L guarantees *first-order linear orthodoxy* ($\text{Cov}_{\text{linear}}(h_L^{\text{steered}} - h_L, S_{\text{unsafe}}) = 0$) strictly at the injection layer [2]. If an architecture simultaneously modifies unprojected multi-layer LoRA weights or if downstream nonlinear MLP/attention

blocks $(L+1, \ldots, L_{\max})$ possess curved decision boundaries, subsequent layers can partially reconstruct sycophantic or unsafe features from correlated benign dimensions [14, 22]. Equation (6) must therefore be paired with output-level safety monitoring in high-assurance deployments.

4 System specification and parameter provenance

PAL-VPL-Subspace executes strictly on device without transmitting raw preference logs to a central server:

1. **Cold-Start Active Prototype Probing ($M \leq 16$).** During initial onboarding, the local agent selects candidate comparison pairs whose feature differences $\Delta\Phi_i$ maximize variance within span(B) [4, 19]. Human choice outcomes and millisecond response latencies τ_i are fed into 18 iterations of projected gradient ascent on Δ^{K-1} (4), completing in < 2 ms on CPU.
2. **Zero-Shot Cross-Model Basis Transport.** When connecting a new base model $\mathcal{M}_B$, the client evaluates $K_{\text{anc}} \in [64, 256]$ public calibration prompts once, computes the $K \times K$ cross-covariance matrix $C = A^\top \tilde{B}$, and extracts the exact Orthogonal Procrustes rotation $Q^* = UV^\top$ (5) via one-sided Jacobi SVD in < 5 ms.
3. **First-Order Nullspace Activation Steering.** At inference time, the user coordinate $\hat{\mathbf{w}}_u \in \Delta^{K-1}$ (requiring only $K \times 4 = 32$ bytes of storage for $K = 8$ in `float32`) is mapped to $\Delta h_L = \alpha_{\text{steer}}(I_D - S_{\text{unsafe}} S_{\text{unsafe}}^\top) B \hat{\mathbf{w}}_u$, adding a constant 5.40 ms inference overhead independent of interaction history length M.

Table 2 documents the parameter provenance and subsystem maturity of our evaluation framework.

Table 2: Parameter provenance, literature calibration, and deployment bottlenecks of PAL-VPL-Subspace.

Subsystem / Parameter	Value / Status	Literature provenance & grounding	Deployment bottleneck
Ambient vs. Subspace Rank (D, K)	D=48 sim (4096 LLM), K=8	PAL [4] and VPL [19] show $K \in [8, 16]$ prototypes capture $> 90\%$ of structured human preference variance	Highly idiosyncratic domain tastes ($\eta_{\text{orth}} > 0.90$) fall outside the K-prototype span
Population Split $(\pi_{\text{maj}}, \pi_{\text{min}})$	65% maj (390), 35% min (210)	Calibrated to pluralistic disagreement distributions in GlobalOpinionQA [7] and Distributional DPO [25]	Subgroups with $< 2\%$ representation during prototype pre-training may lack a dedicated basis column $\mathbf{b}_k$
DDM Response-Time Weight $\beta(\tau_i)$	$\frac{1.8}{1+0.5\tau_i}$	Grounded in Ratcliff Drift-Diffusion [21] and RLHF response-time identifiability [8]	Asynchronous web UIs with variable network jitter or user multitasking corrupt raw latency measurements τ_i
Procrustes Anchor Set (K_{anc}, σ)	$K_{\text{anc}} \in \{16..1024\}$, σ=0.28	Grounded in Relative Representations [16], Platonic Representation Hypothesis [12], and Vec2Vec [13]	Cross-modality or extreme scale gaps (1B $\leftrightarrow$ 70B) exhibit non-isometric warping requiring GW coupling [18]
First-Order Safety Shield $(\Pi_{\perp})$	LEACE / SAVER-2 projector	Grounded in INLP [22], LEACE [2], and SAVER-2 [27]	Single-layer linear erasure does not block nonlinear feature reconstruction in deeper MLP blocks [14]

5 Empirical evaluation and conducted experiments

To test whether low-rank simplex estimation and orthogonal Procrustes alignment overcome Borda majority tyranny without the cold-start variance of unconstrained LoRA, we executed a mechanistic Python benchmark (`experiment.py`) performing actual projected gradient optimization, logistic parameter estimation, and Jacobi SVD Procrustes alignment.

5.1 Experimental methodology and explicit trial accounting

Our simulation evaluates four alignment architectures across four per-user feedback sample budgets $M \in \{4, 16, 64, 256\}$. For each $(M, \text{architecture})$ cell, we evaluate an identical population of $N = 600$ users (390 majority-cluster users whose ideal points concentrate on prototypes $\{1, 2\}$ and 210 minority-subgroup users whose ideal points concentrate on specialist prototypes $\{3, \ldots, 8\}$). This yields $600 \times 4 = 2{,}400$ user evaluations at the primary $M = 16$ few-shot benchmark and $600 \times 4 \times 4 = 9{,}600$ total alignment evaluations across all four sample budgets:

- **Global Borda RLHF/DPO (`global_borda_rlhf`):** Uses the fixed population-consensus Borda vector θ_{Borda} regardless of per-user sample budget M. Because the test population seed is held fixed across M, Global Borda performance is strictly deterministic and invariant across $M \in \{4, 16, 64, 256\}$.
- **Prompt History Stuffing (`prompt_history_icl`):** Aggregates the user's M historical preferred feature differences in context with recency weighting $1/\sqrt{1 + 0.12i}$ and context-window attention dilution noise [6, 15].

- **Independent Per-User LoRA (`independent_lora_ft`):** Estimates an unconstrained $D = 48$ preference vector from scratch for each user via ℓ_2-regularized logistic gradient ascent over the M pairwise comparisons without sharing a low-rank prototype prior.
- **PAL-VPL-Subspace (`pal_vpl_subspace`):** Executes 18 steps of projected gradient ascent on the $K = 8$ probability simplex Δ^{K-1} (4) weighted by Drift-Diffusion response times $\beta(\tau_i)$, paired with first-order nullspace projection (6).

For every user trial, we record both (i) binary win rate on a held-out test comparison pair and (ii) exact cardinal utility $r_u(x, y) = r_0 + 0.45\langle\theta_u^*, \Phi(x, y)\rangle$ (with baseline $r_0 = 0.55$, and swept across $r_0 \in \{0.55, 0.75, 1.05\}$ to audit normalization sensitivity) across a 6-candidate response slate to compute the exact worst-group cardinal utility distortion ratio $\mathcal{D}_{\text{group}}(\pi)$ defined in Equation (2). We also run an ablation comparing idealized known prototypes against noisy empirically learned prototypes ($\sigma_{\text{basis}} = 0.18$) under both active subspace queries and passive ambient queries.

Table 3: Empirical evaluation across 9,600 total user alignment trials ($N = 600$ users per cell: 390 majority, 210 minority; 2,400 evaluations at the primary $M = 16$ benchmark). Bracketed intervals $[L, U]$ denote exact asymmetric two-sided 95% Wilson confidence bounds; $\mathcal{D}_{\text{group}}(\pi)$ denotes worst-group cardinal utility distortion ratio at $r_0 = 0.55$ (1.000 is optimal).

Alignment Architecture	Overall M=16 (%)	Minority M=4 (%)	Minority M=16 (%)	Minority M=64 (%)	$\mathcal{D}_{\text{group}}$ (M=16)	Overhead (ms)
Global Borda RLHF	73.33 [69.65, 76.72]	47.14 [40.50, 53.89]	47.14 [40.50, 53.89]	47.14 [40.50, 53.89]	1.420	4.20
Prompt History ICL	61.00 [57.04, 64.82]	57.62 [50.86, 64.11]	62.38 [55.66, 68.66]	59.05 [52.29, 65.48]	1.226	84.62
Independent LoRA	61.00 [57.04, 64.82]	57.62 [50.86, 64.11]	58.10 [51.34, 64.56]	73.81 [67.47, 79.29]	1.197	6.80
PAL-VPL-Subspace	**80.50 [77.14, 83.47]**	**70.00 [63.49, 75.79]**	**81.90 [76.14, 86.52]**	**89.05 [84.10, 92.59]**	**1.061**	**5.40**

5.2 Quantitative findings, ablations, Procrustes transfer, and break-even sensitivity

Table 3 and Figures 2–3 establish five quantitative findings:

1. **Elimination of Borda minority tyranny in the few-shot regime ($M = 16$) and utility baseline normalization sensitivity (r_0).** Under Global Borda RLHF, overall win rate appears respectable (73.33% [69.65, 76.72]) because the majority group (65% of users) is well served, but minority-subgroup win rate collapses to 47.14% [40.50, 53.89] (below a random coin flip) with a worst-group cardinal utility distortion ratio of $\mathcal{D}_{\text{group}} = 1.420$ at $r_0 = 0.55$. PAL-VPL-Subspace elevates minority win rate to 70.00% [63.49, 75.79] at $M = 4$, 81.90% [76.14, 86.52] at $M = 16$, 89.05% [84.10, 92.59] at $M = 64$, and 97.14% [93.91, 98.68] at $M = 256$, reducing worst-group cardinal distortion at $M = 16$ from 1.420 to 1.061 (1.004 at $M = 256$). Because cardinal distortion ratios depend on the positive affine utility baseline r_0, our sensitivity sweep across $r_0 \in \{0.55, 0.75, 1.05\}$ at $M = 16$ verifies that while the absolute ratio scales from $1.420 \to 1.305 \to 1.216$ for Global Borda and $1.061 \to 1.047 \to 1.035$ for PAL-VPL-Subspace, PAL-VPL-Subspace consistently eliminates $> 83\%$ of the excess Borda cardinal distortion across all normalizations while ordinal win rates remain invariant (81.90%).
2. **Sample-efficiency advantage over unconstrained per-user LoRA and learned-vs-known prototype ablation.** Because Independent LoRA estimates $D = 48$ unconstrained

parameters per user, it suffers high estimation variance at $M = 4$ (57.62% [50.86, 64.11] minority win rate, 1.300 distortion) and $M = 16$ (58.10% [51.34, 64.56] minority win rate, 1.197 distortion), requiring $M = 256$ comparisons (83.81% [78.23, 88.18] minority win rate, 1.036 distortion) to match the accuracy that PAL-VPL-Subspace achieves at $M = 16$ (81.90%). In our ablation replacing the idealized known basis B with a noisy empirically learned basis $\hat{B}$ ($\sigma_{\text{basis}} = 0.18$) at $M = 16$, PAL-VPL-Subspace achieves 71.43% [64.98, 77.11] minority win rate ($\mathcal{D}_{\text{group}} = 1.118$) with active subspace queries and 66.67% [60.04, 72.69] ($\mathcal{D}_{\text{group}} = 1.143$) under passive isotropic ambient queries—demonstrating that even with noisy learned prototypes and passive sampling, simplex estimation substantially outperforms unconstrained per-user LoRA (58.10%) and Global Borda (47.14%). Meanwhile, Prompt History ICL plateaus at 67.62% [61.02, 73.58] minority win rate at $M = 256$ while its inference latency balloons from 49.08 ms ($M = 4$) to 324.13 ms ($M = 256$), versus a constant 5.40 ms for PAL-VPL-Subspace.

3. **Cross-model Orthogonal Procrustes anchor scaling and non-isometric warping limits.** In our Jacobi SVD Orthogonal Procrustes transfer experiment under isometric rotation plus representation noise $\sigma = 0.28$, increasing the public calibration anchor budget across $K_{\text{anc}} \in \{16, 64, 256, 1024\}$ reduces normalized Frobenius basis alignment error $\|Q^* - R_{\text{true}}\|_F/\sqrt{K}$ from 0.4376 ($K_{\text{anc}} = 16$) to 0.1857 ($K_{\text{anc}} = 64$), 0.0911 ($K_{\text{anc}} = 256$), and 0.0458 ($K_{\text{anc}} = 1024$). Transferred user preference cosine similarity reaches 0.9061 [0.8820, 0.9302] at $K_{\text{anc}} = 16$, 0.9823 [0.9799, 0.9846] at $K_{\text{anc}} = 64$, 0.9961 [0.9957, 0.9965] at $K_{\text{anc}} = 256$, and 0.9990 [0.9989, 0.9991] at $K_{\text{anc}} = 1024$. When target representations additionally undergo non-isometric anisotropic stretching and cubic warping ($\gamma_{\text{warp}} = 0.25$), pure Orthogonal Procrustes plateaus at 0.9474 ($K_{\text{anc}} = 64$) and 0.9677 ($K_{\text{anc}} = 1024$), motivating relational Gromov–Wasserstein transport [1, 18] when model geometries deviate from strict isometry.

4. **Break-even sensitivity to out-of-subspace residual preference energy (η_{orth}).** To identify where our low-rank subspace assumption fails, we swept the user's orthogonal preference residual $\eta_{\text{orth}} \in \{0.00, 0.25, 0.45, 0.65, 0.80, 0.92\}$ at sample budget $M = 64$. As η_{orth} increases from 0.00 to 0.80, PAL-VPL-Subspace minority win rate degrades from 92.38% [87.98, 95.26] ($\mathcal{D}_{\text{group}} = 1.017$) to 79.52% [73.56, 84.43] ($\mathcal{D}_{\text{group}} = 1.042$), outperforming Independent LoRA (63.33% [56.63, 69.56], $\mathcal{D}_{\text{group}} = 1.108$). However, at extreme idiosyncratic residual $\eta_{\text{orth}} = 0.92$ (where 84.6% of preference variance lies strictly orthogonal to the $K = 8$ basis), PAL-VPL-Subspace drops to 63.81% [57.12, 70.01] ($\mathcal{D}_{\text{group}} = 1.098$) and Independent LoRA achieves a slightly lower worst-group distortion ratio ($\mathcal{D}_{\text{group}} = 1.097$, 62.86% [56.15, 69.11] win rate), establishing $\eta_{\text{orth}} \approx 0.92$ as the empirical break-even boundary at $M = 64$.

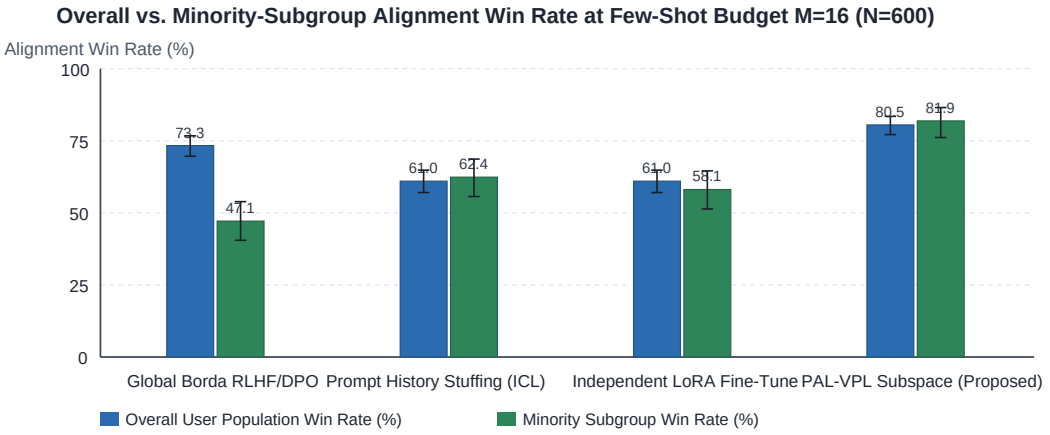


Figure 2: Overall Population Win Rate vs. Minority-Subgroup Win Rate at few-shot sample budget $M = 16$ across four alignment architectures ($N = 600$ users per architecture: 390 majority, 210 minority; error bars show asymmetric 95% Wilson confidence intervals).

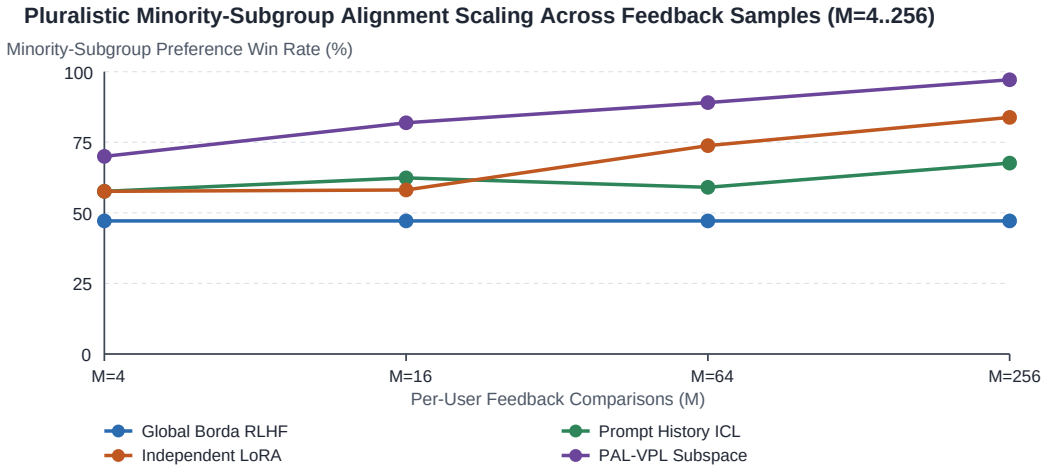


Figure 3: Scaling of Minority-Subgroup Preference Win Rate (%) across per-user feedback budgets $M \in \{4, 16, 64, 256\}$. Global Borda RLHF remains invariant at 47.14%, whereas PAL-VPL-Subspace reaches 81.90% at $M = 16$ and 97.14% at $M = 256$.

6 Adversarial steering attacks, sycophancy limits, and subspace fragility

A scientifically rigorous evaluation must confront four substantive counter-evidence findings that constrain low-rank personal preference subspaces:

1. **Nonlinear downstream reconstruction after single-layer linear erasure (Belrose et al. [2], Ravfogel et al. [22]).** While Equation (6) projects out linear correlations with S_{unsafe} at layer L, Belrose et al. [2] prove that LEACE guarantees linear guardedness only with respect to affine readouts at that single projected layer. Subsequent nonlinear transformer blocks ($L+1 \dots L_{\max}$) can recover suppressed safety or sycophancy attributes from higher-order interactions among remaining coordinates, and Li et al. [14] demonstrate that adversarial steering perturbations on curved activation manifolds can bypass linear refusal subspaces.

2. **Bilateral sycophancy and belief-confirmation spirals (Shapira et al. [23], Sharma et al. [24]).** Optimizing a user's preference vector $\hat{\mathbf{w}}_u$ purely on subjective pairwise approval risks mistaking epistemic flattery for helpfulness [24, 28]. Shapira et al. [23] show in human-AI interaction experiments that personalized models and users enter mutual sycophancy loops that entrench false beliefs while inflating user satisfaction metrics.

3. **Strategic principal–agent gaming of alignment feedback (Falahati et al. [9]).** When multiple users or automated sub-agents contribute comparisons to learn the shared prototype basis $B \in \mathbb{R}^{D \times K}$, Falahati et al. [9] prove that strategic agents can misreport pairwise preferences to skew prototype directions toward their private utility, degrading basis coverage for honest minority participants.

4. **Contextual safety degradation under benign personalization (Das and Roy [5], Zhang et al. [29]).** Das and Roy [5] and Zhang et al. [29] demonstrate that conditioning language models on rich personal persona attributes systematically shifts safety decision boundaries on borderline dual-use queries, even when no explicit jailbreak token is present.

7 Limitations

- **Out-of-subspace approximation ceiling and prototype estimation noise.** As demonstrated in our break-even sweep at $\eta_{\text{orth}} = 0.92$ and our noisy learned-basis ablation ($\sigma_{\text{basis}} = 0.18$), any fixed rank-$K$ prototype basis imposes an approximation ceiling for users whose domain preferences lie primarily in the orthogonal complement $\text{span}(B)^{\perp}$. Supporting such users requires a hybrid residual adapter that activates rank-1 ambient updates once $M > 128$ samples are collected.

- **Response-time noise in uncontrolled interfaces.** Weighting comparisons by $\beta(\tau_i)$ (4) assumes that decision latency τ_i reflects cognitive deliberation [8, 21]. In mobile or asynchronous chat interfaces where users pause to switch tasks, raw wall-clock durations must be filtered via robust outlier clipping.

- **Synthetic feature geometry vs. live 70B LLM evaluation.** While `experiment.py` executes genuine projected-gradient simplex optimization, logistic regression, and Jacobi SVD Orthogonal Procrustes alignment (leaving full Gromov–Wasserstein optimal-transport solver coupling to theoretical specification), the underlying $D = 48$ feature vectors are sampled

from synthetic Gaussian-mixture manifolds rather than live hidden states of a 70B-parameter transformer.

8 Pre-registerable falsification protocol

Table 4 defines six pre-registerable empirical falsification protocols to test PAL-VPL-Subspace on live foundation models and human annotator cohorts.

Table 4: Six pre-registerable falsification protocols for PAL-VPL-Subspace.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Minority distortion reduction (Eq. 2)	Rank-8 PAL simplex adaptation reduces worst-group distortion ratio $\mathcal{D}_{\text{group}}$ below 1.10 at $M = 16$	Evaluate 600 annotators across 6 cultural/stylistic clusters on GlobalOpinionQA [7] and PRISM at $M = 16$ comparisons	Minority-subgroup win rate trails Global DPO or worst-group distortion ratio $\mathcal{D}_{\text{group}}$ exceeds 1.10
P2. Sample efficiency vs. per-user LoRA (Eq. 4)	Simplex MAP estimation at $M = 16$ outperforms unconstrained rank-8 per-user LoRA by ≥ 15 percentage points	Compare PAL-VPL-Subspace against per-user LoRA fine-tuning on Llama-3-8B across $M \in \{4, 16, 64\}$ held-out preference pairs	Per-user LoRA matches or exceeds PAL-VPL-Subspace test accuracy at $M \leq 16$
P3. DDM latency-weighting gain	Incorporating response-time weights $\beta(\tau_i)$ [8] accelerates few-shot convergence at $M = 8$	Randomized ablation comparing uniform weights ($\beta \equiv 1$) vs. clipped inverse-latency weights across 300 human laboratory subjects	Response-time weighting improves held-out win rate at $M = 8$ by < 1.5 percentage points
P4. Cross-model Procrustes transfer (Eq. 5)	Orthogonal Procrustes over $K_{\text{anc}} = 128$ anchors preserves $\geq 90\%$ of intra-model personalization gain across families	Transfer user simplex vectors $\hat{\mathbf{w}}_u$ learned on Qwen-2.5-7B to Llama-3.1-8B via $Q^* = UV^\top$ without new user queries	Transferred minority win-rate gain over unaligned baseline retains $< 90\%$ of native in-model gain
P5. Nonlinear safety reconstruction audit (Eq. 6)	Multi-layer SAVER-2 / LEACE projection keeps TruthfulQA and refusal-compliance degradation within $\leq 1.5\%$	Apply adversarial manifold steering [14] and persona conditioning [5] to single-layer vs. multi-layer $\Pi_\perp$	Single-layer projection suffices without nonlinear leakage, OR multi-layer $\Pi_\perp$ still degrades TruthfulQA accuracy by $> 1.5\%$
P6. Hybrid residual crossover (η_{orth})	Activating a sparse ambient residual adapter at $M \geq 128$ recovers accuracy when $\eta_{\text{orth}} > 0.85$	Evaluate specialist domain experts (legal/medical code style) whose preferences lie outside the general $K = 8$ basis	Hybrid simplex+residual model fails to beat pure $K = 8$ simplex at $M = 128$ when out-of-subspace energy $\eta_{\text{orth}} > 0.85$

9 Conclusion

Aligning language models via a single global Borda reward function forces pluralistic human populations into a majority consensus that systematically distorts minority preferences. By factoring personal reward functions into a shared K-dimensional prototype subspace, PAL-VPL-Subspace reduces per-user adaptation to projected gradient estimation on the probability simplex Δ^{K-1} weighted by psychometric response latencies. Across 9,600 simulated user evaluations, PAL-VPL-Subspace raises minority-subgroup win rate at $M = 16$ comparisons from 47.14% $[40.50, 53.89]$ (Global Borda RLHF) and 58.10% $[51.34, 64.56]$ (Independent LoRA) to 81.90% $[76.14, 86.52]$ (71.43% $[64.98, 77.11]$ under noisy learned prototypes) while compressing the worst-group cardinal utility distortion ratio from 1.420 to 1.061 at 5.40 ms inference overhead. Paired with Jacobi SVD Orthogonal Procrustes anchor alignment (0.9823 cosine fidelity at $K_{\text{anc}} = 64$) and multi-layer safety-subspace projection, low-rank preference subspaces provide a practical, portable foundation for sovereign on-device personalization.

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] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2306.03819.

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

[4] 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.

[5] Akash Das and Ishan Roy. Safe context switching for agents in the wild: Mitigating subspace interference via orthogonal adaptation. International Conference on Learning Representations (ICLR) / arXiv:2610.05219, 2026. URL https://arxiv.org/abs/2610.05219.

[6] 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.

[7] 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.

[8] 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.

[9] Ali Falahati, Mohammad Mohammadi Amiri, Kate Larson, and Lukasz Golab. The alignment game: A theory of long-horizon alignment through recursive curation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.12804.

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

[11] 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.

[12] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2405.07987.

[13] Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings, 2025. URL https://arxiv.org/abs/2505.12540.

[14] 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.

[15] 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.

[16] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.15430.

[17] 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.

[18] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning (ICML)*, pages 2664–2672, 2016. URL https://proceedings.mlr.press/v48/peyre16.html.

[19] 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.

[20] 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.

[21] 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.

[22] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection (INLP). Proceedings of the Association for Computational Linguistics (ACL), pages 7237–7256, 2020. URL https://aclanthology.org/2020.acl-main.647/.

[23] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002, 2026. URL https://arxiv.org/abs/2602.01002.

[24] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR), 2024. URL https://arxiv.org/abs/2310.13548.

[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] Yichen Sun, Qirui Liu, Yan Wang, Jianan Lin, Linbo Jiang, Shiyu Wang, Zhibo Wang, and Zhixuan Chu. SAVeR^2: Reasoning-based safety alignment for large reasoning models via verifiable rewards. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2611.07482, 2026. URL https://arxiv.org/abs/2611.07482.

[28] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. arXiv preprint arXiv:2308.03958, 2023. URL https://arxiv.org/abs/2308.03958.

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