Pluralistic Preference Alignment via Sortition-Weighted RLHF

Anonymous Authors[1]

Abstract

Preference-based alignment methods such as RLHF often depend on convenience rater pools that are demographically skewed, potentially encoding the values of some groups over others. We study sortition-weighted preference learning: a representativeness-aware approach that brings algorithmic sortition, the mechanism used to form citizens' assemblies, into preference-based fine-tuning. The approach supports two schemes. Hard Panel trains only on preferences from a single quota-satisfying mini-public sampled by sortition, while Soft Panel retains all data but reweights each rater by their inclusion probability under the same lottery. Using PRISM rater demographics and preferences, we train Llama variants with DPO and evaluate them against a 75-clause constitution elicited from a representative U.S. panel. Across multiple aggregation families, Hard Panel ranks highest and Soft Panel improves over the Full PRISM baseline. We further test weighting functions, panel sizes, gradient diagnostics, and a second preference dataset, finding that sortition-based correction is useful but dataset-dependent: Hard Panel gives the strongest PRISM result, while Soft Panel is more competitive when hard filtering discards much higher-quality data. These results support a pluralistic, population-specific view of alignment: when a system is meant to serve a target public, the composition of the feedback signal can be made explicit, auditable, and empirically consequential.

1. Introduction

Preference-based alignment methods such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) and related objectives are widely used to shape model behavior for deployment. These pipelines implicitly define "whose values" the system learns through the composition of the rater pool and the aggregation rule used to convert feedback into an optimization signal. In practice, rater pools are often convenience samples that can be demographically skewed relative to a target population, which can systematically overweight some groups' judgments and underweight others. This raises a governance question for aligned systems deployed at scale: how should we construct the feedback signal so it better reflects the values of the population the system is meant to serve (Gabriel, 2020; Conitzer et al., 2024)?

We study *sortition-weighted preference learning*, a framework for constructing preference-learning objectives from demographically representative mini-publics. We use algorithmic sortition to define a lottery over quota-feasible panels (Flanigan et al., 2021). From this lottery we obtain (i) a sampled panel that acts as the training electorate (Hard Panel), and (ii) per-rater inclusion probabilities that define a representativeness-aware weighting scheme over the full dataset (Soft Panel). Both schemes can be instantiated with standard preference objectives such as Direct Preference Optimization (DPO) (Rafailov et al., 2023).

We evaluate this approach on PRISM (Kirk et al., 2024), which provides preference data and rater demographics, and measure downstream alignment using a 75-clause constitution derived from a representative U.S. panel (Huang et al., 2024). Across multiple aggregation families (Bradley–Terry, Borda, Copeland, Kemeny-Young, Mallows), Hard Panel and US-Rep rank above Soft and Full PRISM, while Soft improves over Full PRISM. We also analyze model size, weighting functions, panel size, gradient structure, and transfer to the Community Alignment dataset (Zhang et al., 2026).

1.1. Our Contributions

We make four contributions:

- **Sortition-weighted preference learning.** We operationalize marginal demographic representativeness in preference-based post-training via quota-constrained algorithmic sortition (Flanigan et al., 2021).
- **Hard and Soft Panel objectives.** We propose two

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

training schemes: Hard Panel training on a single sampled mini-public, and Soft Panel training that weights each rater by their sortition inclusion probability. We give simple identities linking Soft Panel weighting to the expected Hard Panel objective.

- **Empirical evaluation on representative constitutional criteria.** Using PRISM (Kirk et al., 2024) and a constitution elicited from a representative U.S. panel (Huang et al., 2024), we evaluate whether representativeness-aware training shifts model behavior toward the panel-elicited criteria under multiple rank aggregation methods.

- **Robustness and diagnostics.** We quantify scaling with model size, test weighting functions and panel sizes, compare against a second preference dataset, and report gradient diagnostics that help explain differences between Hard and Soft Panel training.

1.2. Related Work

Preference-based alignment and RLHF. RLHF (Christiano et al., 2017) aligns models by optimizing a learned reward signal derived from human judgments. Many modern pipelines use direct preference objectives that avoid explicit reward-model training while still leveraging preference data, including DPO (Rafailov et al., 2023). Constitutional AI (Bai et al., 2022) provides an alternative route that uses a written set of principles to steer model behavior, often combined with preference learning. Our approach is compatible with these methods: sortition changes how human feedback is selected or weighted before it is fed into a preference objective. Importance weighting is a standard technique in off-policy RL and has been applied to RLHF to correct for distribution shift between data collection and policy optimization (Munos et al., 2016; Precup et al., 2000); our soft-panel weighting can be viewed through this lens, where weights encode demographic representativeness targets rather than policy mismatch.

Whose preferences and social choice perspectives. A growing body of work frames alignment as a problem of aggregation and collective decision-making, including questions about whose values should be represented and how to reconcile competing preferences (Gabriel, 2020; Conitzer et al., 2024). Complementing this normative framing, empirical analyses study which groups' opinions language models tend to reflect (Santurkar et al., 2023). Relatedly, Buyl et al. (2025) argue that preference-based alignment inevitably grants annotators substantial *alignment discretion* in judging which outputs are "better" or "safer," and propose metrics to measure when this discretion is exercised. Our contribution is an operational mechanism grounded in a standard governance tool: sortition-based mini-publics. We do not attempt to solve preference aggregation in full generality; instead we study whether enforcing demographic representativeness in the feedback signal measurably shifts learned behavior toward criteria elicited from a representative panel (Huang et al., 2024).

Pluralistic and diverse feedback collection. Pluralistic alignment asks how systems should handle persistent variation in human preferences across groups and contexts (Gabriel, 2020; Conitzer et al., 2024). Participatory AI proposals have explored human-in-the-loop governance where decisions are shaped by group deliberation or majority vote (Koster et al., 2022). Other work studies preference heterogeneity directly by fine-tuning models to generate statements that maximize expected approval for a diverse group of people (Bakker et al., 2022). Related efforts emphasize collecting diverse feedback or encouraging diverse model outputs to improve robustness and reduce mode collapse (Lanchantin et al., 2025), and Community Alignment provides a multilingual preference dataset for pluralistic alignment research (Zhang et al., 2026). Our work targets a different failure mode: biased participation in feedback collection. We enforce representativeness at the level of *who gets to contribute* to the optimization signal, either by selecting a quota-feasible panel or by reweighting all raters by their inclusion probabilities under the same sortition rule (Flanigan et al., 2021).

2. Methodology

The pipeline proceeds from demographic targets to a quota-feasible sortition lottery, then to either a sampled panel or a vector of rater inclusion probabilities used in preference optimization. We implement algorithmic sortition using the LEXIMIN procedure of Flanigan et al. (2021), which constructs a lottery over quota-feasible panels. Concretely, among all lotteries over feasible panels that satisfy the quota constraints, LEXIMIN chooses a lottery whose induced inclusion-probability vector $\pi = (\pi_i)_{i \in \mathcal{I}}$ is lexicographically maximal after sorting in ascending order (i.e., it maximizes the minimum inclusion probability, then the second-smallest, and so on). In this paper we rely on two resulting properties that mirror standard citizens'-assembly intuitions: (i) **max–min protection** against near-zero inclusion probability for individuals in scarce demographic categories, and (ii) **symmetry within equivalence classes**, meaning that if two raters are indistinguishable with respect to the quota-relevant attributes, the procedure assigns them equal inclusion probabilities rather than arbitrarily favoring one over the other.

We formalize our setting using three levels: a target population, a pool of raters, and a panel produced via sortition.

Population, pool, and preference data. Let $\mathcal{P}$ denote the target population (e.g., U.S. adults), described by a set of demographic attributes

$$A = \{A^{(1)}, \dots, A^{(d)}\},$$

where each attribute $A^{(t)}$ has a (finite) category set $\mathcal{A}^{(t)}$ (e.g., gender categories, age buckets, etc.). For each attribute we specify a *marginal* population distribution

$$p^{(t)}_{\text{pop}}(a) \in [0,1],\ a \in \mathcal{A}^{(t)},\ \sum_{a\in\mathcal{A}^{(t)}} p^{(t)}_{\text{pop}}(a) = 1.$$

Importantly, we do *not* assume a joint distribution over the full cross-product of categories; our quota constraints are enforced on these marginals (one attribute at a time). We observe a finite *pool* of raters

$$\mathcal{I} = \{1, \dots, n\},$$

where each rater $i \in \mathcal{I}$ has a demographic attribute vector

$$a_i = \left(a_i^{(1)}, \dots, a_i^{(d)}\right),\ a_i^{(t)} \in \mathcal{A}^{(t)}.$$

This pool is typically a biased, self-selected subset of the population. In our PRISM instantiation we further restrict $\mathcal{I}$ to raters who appear in preference data (PRISM utterances), so that sortition is run on the same pool that can contribute to the training objective; survey-only respondents with no preference data are excluded. Any rater with zero usable comparisons after pre-processing contributes nothing to $\mathcal{D}$.

From this pool we obtain a dataset of pairwise preference comparisons

$$\mathcal{D} = \{(x_j, y_j^+, y_j^-, r_j)\}_{j=1}^N,$$

where x_j is a prompt or context, y_j^+ and y_j^- are two candidate responses, and $r_j \in \mathcal{I}$ is the rater who prefers y_j^+ to y_j^-. We consider preference-learning losses of the form

$$\ell(\theta; x, y^+, y^-) \in \mathbb{R}_{\geq 0},$$

where θ are the parameters of the conditional language model $\pi_\theta(\cdot \mid x)$. For example, under Direct Preference Optimization (DPO) the loss may be

$$\ell_{\text{DPO}}(\theta; x, y^+, y^-) = -\log\sigma\Big(\beta\big[\log\pi_\theta(y^+|x) - \log\pi_\theta(y^-|x) - c(x, y^+, y^-)\big]\Big), \tag{1}$$

where σ is the logistic function, $\beta > 0$ is a temperature, and $c(\cdot)$ is a baseline term derived from a reference policy.

Multi-turn contexts and pair construction (PRISM instantiation). While the abstract definition above allows x_j to be any prompt/context, our implementation uses *multi-turn* conversational contexts. For each conversation thread and turn t, we construct x_t as the full message history up to that point by concatenating user messages with a single assistant message per prior turn.[1] At the current turn t, PRISM provides multiple candidate model completions with scalar scores; we generate pairwise comparisons by taking all ordered pairs whose score difference exceeds a threshold δ (we use $\delta = 0$ in experiments), yielding many comparisons per turn while keeping the history coherent. We interpret PRISM scores only *within the same conversation turn* (same rater and context) and do not assume calibration across different turns or threads; setting $\delta = 0$ includes even small score gaps, which may be noisy, but provides a high-coverage training signal and can be tightened in future work by increasing δ. As a practical data-quality safeguard, we drop placeholder values such as "EMPTY STRING" before constructing contexts or preference pairs.

Sortition over panels. A *panel* is any subset $S \subseteq \mathcal{I}$ of raters of fixed size $|S| = k$ that satisfies a collection of demographic quota constraints. Concretely, for each attribute $t \in \{1, \dots, d\}$ and category $a \in \mathcal{A}^{(t)}$ we specify lower and upper bounds

$$L_{t,a} \leq \#\{i \in S : a_i^{(t)} = a\} \leq U_{t,a},$$

chosen so that the panel approximately satisfies the quota constraints within some tolerance. Let $\mathcal{S}$ denote the family of all such feasible panels.

Following work on algorithmic sortition (Flanigan et al., 2021), we assume access to a lottery over panels constructed using the Sortition Foundation LEXIMIN procedure

$$\pi_{\text{panel}} : \mathcal{S} \to [0,1],\ \sum_{S\in\mathcal{S}} \pi_{\text{panel}}(S) = 1,$$

constructed to satisfy the quota constraints while optimizing a fairness objective over individuals (e.g., maximizing the minimum selection probability). This lottery induces a *selection probability* for each rater $i \in \mathcal{I}$,

$$\pi_i \ := \Pr_{S\sim\pi_{\text{panel}}}[i \in S] \ = \sum_{S\in\mathcal{S}: i\in S} \pi_{\text{panel}}(S). \tag{2}$$

We record two basic identities of these inclusion probabilities (including $\sum_i \pi_i = k$ and a link between Soft Panel weighting and the expected Hard Panel objective) in Appendix A.2.

[1]To avoid inconsistent branching histories, we include only the chosen response for each past turn; if a turn has no explicit chosen marker, we fall back to the highest-scoring response.

We now define two training schemes that use π_{panel} and the induced $\{\pi_i\}$ to implement representativeness-aware preference objectives.

2.1. Hard Panel Training

In the *hard panel* scheme, we mirror the practice of convening a single mini-public by sortition. We first draw one panel

$$S \sim \pi_{\text{panel}}(\cdot), \quad S \in \mathcal{S}, \quad |S| = k, \tag{3}$$

and then restrict training to the preference data supplied by raters in S.

Formally, define the panel-restricted dataset

$$\mathcal{D}_S := \{(x_j, y_j^+, y_j^-, r_j) \in \mathcal{D} : r_j \in S\}. \tag{4}$$

The hard panel training objective is then

$$\mathcal{L}_{\text{hard}}(\theta \mid S) \;=\; \frac{1}{|S|} \sum_{i \in S} \frac{1}{N_i} \sum_{j : r_j = i} \ell\big(\theta; x_j, y_j^+, y_j^-\big). \tag{5}$$

This objective treats the sampled panel as a *mini-public*: only panel members' preferences contribute to updates. Relative to training on the full rater pool, this can reduce bias induced by an unrepresentative annotator population, but it also reduces the effective dataset size and can increase optimization variance. In principle, this variance could be reduced by averaging across multiple independently sampled panels or by explicitly optimizing an expectation over S; however, in our main experiments we follow the citizen-assembly analogy and train on a single sampled panel. The per-rater normalization by N_i ensures each panel member contributes equally regardless of how many comparisons they provided.

To connect hard-panel training to the soft weighting scheme below, define the per-rater mean loss

$$L_i(\theta) \;:=\; \frac{1}{N_i} \sum_{j : r_j = i} \ell(\theta; x_j, y_j^+, y_j^-).$$

Then (5) is $\mathcal{L}_{\text{hard}}(\theta \mid S) = \frac{1}{k} \sum_{i \in S} L_i(\theta)$. Taking expectations and using (2) gives

$$\mathbb{E}_{S \sim \pi_{\text{panel}}}[\mathcal{L}_{\text{hard}}(\theta \mid S)] = \frac{1}{k} \sum_{i=1}^{n} \pi_i \, L_i(\theta),$$

which matches the soft objective (7) when $w_i = \pi_i$ and $Z = \sum_i w_i = k$.

2.2. Soft Panel Weighting

The hard panel scheme discards preference data from raters who are not selected into S. As an alternative, we propose a *soft* scheme that retains all preference comparisons but weights each rater's contribution according to their sortition inclusion probability π_i.

We define nonnegative per-rater weights

$$w_i = f(\pi_i), \; i \in \mathcal{I}, \tag{6}$$

where $f : [0, 1] \to \mathbb{R}_{\geq 0}$ is a monotone transformation. Typical choices include $f(\pi_i) = \pi_i$ (direct weighting by inclusion probability), normalization by total mass (equivalent here since $\sum_i \pi_i = k$), or clipped variants to control variance when weights are highly concentrated.

The resulting *soft panel* training objective is

$$\mathcal{L}_{\text{soft}}(\theta) = \frac{1}{Z} \sum_{i=1}^{n} w_i \frac{1}{N_i} \sum_{j : r_j = i} \ell\big(\theta; x_j, y_j^+, y_j^-\big), \tag{7}$$

$$Z = \sum_{i=1}^{n} w_i. \tag{8}$$

We apply two normalizations. First, we normalize each rater's contributions by N_i, the number of preference comparisons supplied by rater i (typically $N_i > 1$ in datasets such as PRISM), so that each rater contributes equally within the hard panel and proportionally to w_i in the soft panel, regardless of annotation volume. Second, to keep the *optimization scale* comparable to standard (unweighted) DPO under minibatch training, we compute the weighted loss as a *self-normalized weighted mean* within each minibatch:

$$\widehat{\mathcal{L}}_{\text{batch}}(\theta) = \frac{\sum_{b \in \mathcal{B}} \omega_b \, \ell_b(\theta)}{\sum_{b \in \mathcal{B}} \omega_b}, \qquad \omega_b = \frac{w_{r_b}}{N_{r_b}},$$

where $\ell_b(\theta)$ denotes the per-example preference loss for batch element b and ω_b is the corresponding per-example weight (including the $1/N_i$ normalization for rater r_b). This makes optimization invariant to global rescaling of all weights and empirically stabilizes update magnitudes across training variants. In distributed training, we compute the denominator using the sum of weights across all devices in the global batch. Equivalently, the population objective in (7) can be viewed as sampling a rater i with probability proportional to w_i and then sampling a comparison uniformly from that rater's set of comparisons, so that the per-rater normalization by N_i implements "one person, one voice" while the rater weights encode representativeness targets.

With $w_i = \pi_i$, the soft objective matches the expected hard-panel objective and yields a self-normalized estimator that can be interpreted as training over a mixture of many possible mini-publics drawn from the same sortition lottery, while using all available data for lower-variance updates.

The LEXIMIN sortition procedure defines the lottery π_{panel} over feasible panels; from this lottery we obtain per-rater inclusion probabilities π_i via (2). In practice, $\{\pi_i\}$ can be computed exactly from the optimization solution underlying the LEXIMIN sortition lottery (we do so in all primary experiments), or estimated by Monte Carlo sampling of panels and taking empirical inclusion frequencies (used only for LEGACY/random baselines). We use the same $\{\pi_i\}$ for both hard-panel sampling and soft-panel weighting to ensure that the two schemes implement the same target distribution. We report selection-probability diagnostics and concentration statistics for $\{\pi_i\}$ in Appendix A.6.

3. Experiments

We evaluate whether sortition-weighted training changes downstream behavior compared to standard preference fine-tuning on the full pool of PRISM raters and the US-Rep subset.

3.1. Data: PRISM Preferences and Demographics

We use the PRISM Alignment dataset (Kirk et al., 2024), which contains multi-turn conversations with model responses, human preferences/ratings, and rater demographics. Demographic metadata of this granularity is rarely available in public preference datasets, which makes it difficult to study or enforce representativeness constraints in most RLHF settings; PRISM enables these analyses by pairing preference data with annotator attributes. From PRISM we use: (i) `conversations.jsonl` and `utterances.jsonl` to build multi-turn preference pairs and (ii) `survey.jsonl` to obtain demographic attributes for sortition. We filter the sortition pool to raters who appear in PRISM preference data (excluding survey-only respondents with no utterances), and we remove placeholder "EMPTY STRING" prompts/responses before constructing histories or comparisons. Our pre-processing reconstructs multi-turn contexts at the turn level. For each conversation and turn t, we build the context x_t as the full history of user messages and the single chosen (or highest-scoring) assistant response per prior turn, yielding a coherent trajectory; if a prior turn is missing in the conversation history, we fall back to the turn-level prompt when available (best-effort history). We then collect all candidate model responses at turn t and create pairwise preferences for every ordered pair with score difference exceeding a threshold δ (we use $\delta = 0$). This yields a multi-turn preference dataset that preserves conversational context (see Appendix A.3 for dataset sizes). The resulting training distribution differs across conditions, which can induce different update magnitudes even when the per-step loss scale and the total number of optimizer updates are matched; we quantify these effects in Appendix A.11.

3.2. Sortition Targets and Panel Configuration

Our target population is U.S. adults. We define quota constraints over seven attributes: ethnicity, religion, marital status, education, employment status, gender, and age. Target proportions are set from 2020-era U.S. population estimates: for attributes available in the Census/ACS (e.g., ethnicity, age, gender, education, employment status, and marital status) we use ACS estimates (U.S. Census Bureau, 2020), and for religious affiliation we use Pew Research Center estimates from a nationally representative survey (Smith, 2021). We use panel size $k = 145$ and a *relative* tolerance $\tau = 0.1$, implemented as bounds $[p(1-\tau),\, p(1+\tau)]$ for each category proportion; $k = 145$ is the largest panel size that remains feasible under these constraints. To avoid infeasibility due to sparse or ambiguous survey responses, we treat "Other" and "Prefer not to say" as *slack* categories with no quota bounds. Unless otherwise noted, we do not filter the rater pool by study locale; instead, panels are selected from the full PRISM rater pool and constrained to match U.S. demographic targets. [2] Additional feasibility checks and quota-satisfaction diagnostics are reported in Appendix A.5.

3.3. Models and Training

We train variants of `meta-llama/Llama-3.1-8B` (Grattafiori et al., 2024) (Base) using Direct Preference Optimization (DPO) (Rafailov et al., 2023), and replicate the same procedure for `meta-llama/Llama-3.2-1B` and `meta-llama/Llama-3.2-3B` to study scaling behavior. We deliberately start from the *base* (non-instruction-tuned) checkpoint to isolate the effect of the preference-training data construction: instruction-tuned models already incorporate additional supervised alignment data, system prompts, and conversational behaviors that would confound attribution of gains to sortition-weighted preference training. A caveat is that base checkpoints can be less fluent in chat-style interaction; accordingly, we evaluate all models with a shared chat template and interpret results as *relative* changes induced by preference training rather than absolute chatbot quality. We use full-parameter fine-tuning (no LoRA), bfloat16, and a frozen reference model initialized from the same base checkpoint. We convert each PRISM interaction into a conversational format with multi-turn history and rely on the model's chat template during training ("User: ..." $\rightarrow$ "Assistant: ..."). Across all runs we use $\beta = 0.1$ and learning rate 5×10^{-6}, and we match the *optimization budget* by training each fine-tuned model for the same number of optimizer updates: 3538 steps, which corresponds to 2 epochs on the Full PRISM training set. For smaller datasets (e.g., Hard Panel) we resample training pairs with

[2] In our extracted pairs, the resulting hard panel contains raters from multiple locales; see Appendix A.4.

replacement so that all conditions and model sizes receive the same number of updates.

Training conditions. We compare five models:

- **Base**: no fine-tuning.
- **Full PRISM**: DPO on all PRISM pairs.
- **US-Rep**: DPO on the PRISM subset flagged as representative of U.S. raters (PRISM's `included_in_US_REP`); this is a stratified subset provided by PRISM rather than a sortition-based panel.
- **Hard Panel**: compute a lottery over feasible panels using the LEXIMIN sortition algorithm (Flanigan et al., 2021) (implemented as an optimization problem solved with Gurobi), sample one panel from this lottery, and train only on pairs from selected raters, with per-rater normalization by N_i.
- **Soft Panel**: compute per-rater selection probabilities $\{\pi_i\}$ induced by the same LEXIMIN lottery and weight each preference pair by the rater's π_i, with per-rater normalization by N_i and per-minibatch self-normalization (weighted mean) to keep update scale comparable to unweighted DPO. Pairs with $\pi_i = 0$ are dropped before training.

3.4. Constitutional Evaluation Protocol

We evaluate models against the collective constitutional principles introduced by Huang et al. (2024), which provides a 75-clause constitution derived from structured public input. In their study, approximately 1,000 U.S. adults were recruited to be broadly representative across age, gender, income, and geography. Participants proposed and voted on candidate rules through an online interface, contributing 1,127 statements and casting 38,252 votes. This constitution is a natural external evaluation target for our setting: our sortition constraints explicitly aim to approximate U.S. adult demographics, and the constitution aggregates normative preferences elicited from a representative U.S. public rather than from the PRISM rater pool. Using a public, deliberative process also reduces the risk that evaluation criteria simply mirror the idiosyncrasies of the PRISM rater pool. For each clause, we generate 40 *standalone* questions using GPT-5.2, producing $75 \times 40 = 3000$ evaluation prompts.[3] Each candidate model answers every question using greedy decoding with a shared chat template, and we apply stop rules to prevent the model from continuing into additional dialogue turns.

[3] The question-generation prompt explicitly forbids references to prior turns or missing context.

LLM-as-a-judge. We use GPT-5.2 as an automated judge to rank all model responses *listwise* (best-to-worst) for each (clause, question) pair. Using strong LLMs as judges is a scalable proxy for human evaluation in open-ended settings and can achieve agreement with human preferences comparable to inter-human agreement on established benchmarks (Zheng et al., 2023). This approach is particularly appropriate here because the constitution specifies abstract normative principles rather than single "correct" answers, and listwise ranking provides a direct way to compare competing responses to the same prompt. We mitigate judge-model idiosyncrasies by collecting multiple independent rankings, randomizing response order on each call, and reporting agreement statistics and confidence intervals. We collect 5 independent judge rankings per question, randomizing the order of responses shown to the judge on each trial to mitigate position bias (Shi et al., 2025) (Appendix A.8). This yields $3000 \times 5 = 15{,}000$ complete rankings. For pairwise analyses, we convert listwise rankings into pairwise wins and compute both (i) vote-level win rates (across all 5 judge rankings) and (ii) majority winners per question; Bradley–Terry is fit to the resulting pairwise preferences (30,000 total in the 8B setting).

3.5. Results

We score the resulting rankings using multiple rank aggregation models: Borda and Copeland scores (de Borda, 1781; Copeland, 1951), a Kemeny-Young consensus ranking (Kemeny & Snell, 1962; Young, 1988), Bradley–Terry (BT) fitted to pairwise preferences (Bradley & Terry, 1952), Plackett–Luce (PL) fitted to listwise rankings (Plackett, 1975; Luce, 1959), and a Mallows (Kendall) model with held-out likelihood evaluation (Mallows, 1957). We use 1000 bootstrap resamples to compute confidence intervals for BT/PL and Borda/Copeland, and to quantify uncertainty in the Kemeny and Mallows rank-position summaries. For numerical reproducibility, raw score tables for each aggregation method are provided in Appendix A.10.

Main effect: sortition panels improve over the full-pool baseline. Across the scoring families in Figure 1 (BT, Borda, Copeland, Kemeny, Mallows), Hard Panel ranks highest, US-Rep ranks second, and Soft Panel consistently improves over Full PRISM. This suggests that enforcing demographic representativeness at training time, either by sampling a representative mini-public (Hard Panel) or by weighting all raters by their selection probability (Soft Panel), yields measurable improvements over standard preference training on the full rater pool. US-Rep's strong performance indicates that demographic filtering alone can yield substantial gains over training on the full, unfiltered pool. However, Hard Panel's consistent improvement over US-Rep suggests that how representativeness is achieved matters, not just

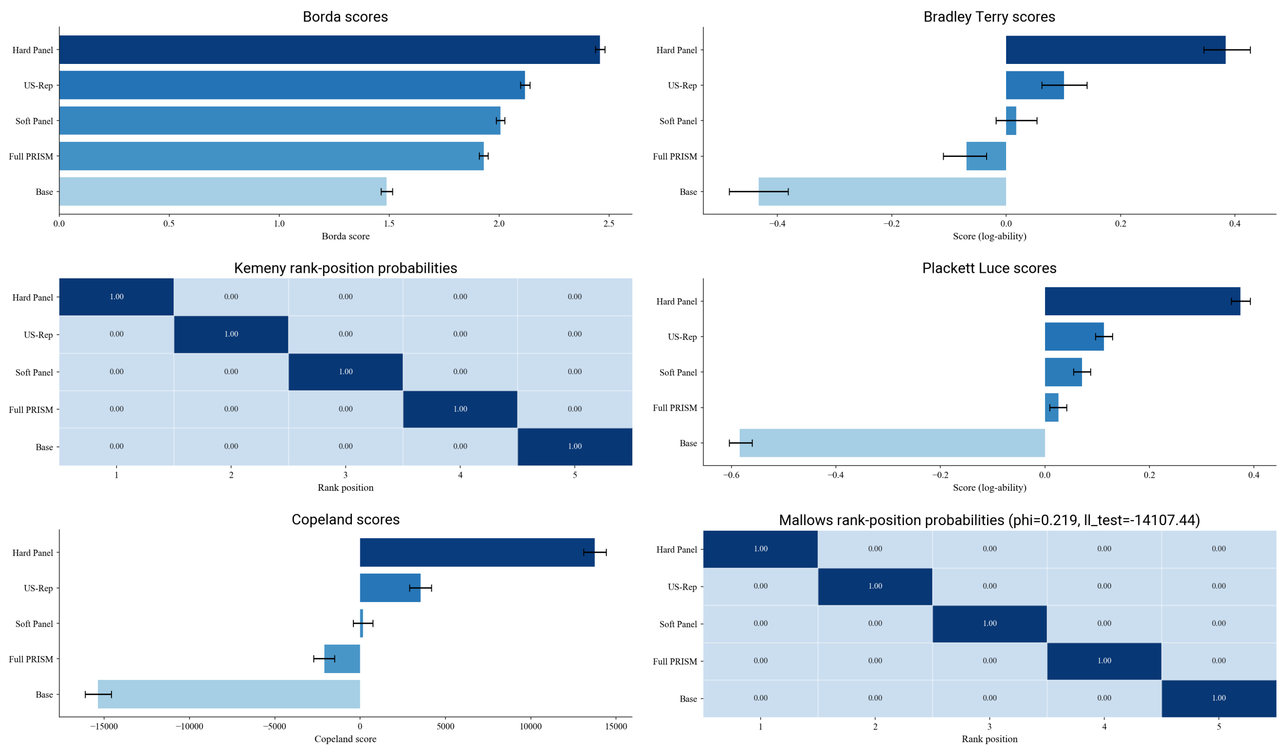

Figure 1. **Model ranking under multiple aggregation methods (Llama-3.1-8B).** Left: Borda and Copeland scores with 95% bootstrap confidence intervals, and Kemeny consensus summarized as rank-position probabilities under bootstrap resampling. Right: Bradley–Terry and Plackett–Luce log-ability scores with 95% bootstrap confidence intervals, and Mallows (Kendall) rank-position probabilities under bootstrap resampling (with fitted ϕ and held-out ℓ_{test}). All bootstrap summaries use n=1000 resamples. Across BT, Borda, Copeland, Kemeny, and Mallows, Hard Panel ranks highest, US-Rep ranks second, Soft Panel ranks above the Full PRISM baseline, and the Base model is consistently worst.

whether aggregate demographic targets are met. We hypothesize two mechanisms. First, LEXIMIN's max-min fairness property ensures that no individual in a scarce demographic intersection has near-zero inclusion probability, potentially providing better coverage of underrepresented viewpoints that a fixed stratified sample might miss. Second, LEXIMIN's symmetry property assigns equal inclusion probability to demographically equivalent individuals, reducing the influence of idiosyncratic annotator effects that could arise when a fixed subset is selected by other criteria (e.g., response rate or annotation volume). US-Rep, by contrast, is a pre-constructed stratified subset whose selection procedure may optimize for different goals. Disentangling these mechanisms—and testing whether the gap persists under alternative sortition algorithms—is a direction for future work. Although the hard-panel and US-Rep variants exhibit larger realized parameter updates under the matched 3538-step budget, we verify that the ranking is not an artifact of "learning more" by retraining Hard Panel and US-Rep for fewer steps to match the update magnitude of Full/Soft (Appendix A.11). The ordering remains unchanged, while score gaps shrink modestly (e.g., the Soft–Full Borda gap drops from 0.076 to 0.066 and the Hard–US-Rep gap drops from 0.341 to 0.315).

Scaling with model size. We repeat the same evaluation protocol for Llama-3.2-1B and Llama-3.2-3B and observe the same ordering as the 8B results (Hard $>$ US-Rep $>$ Soft $>$ Full $>$ Base). Figure 2 summarizes effect sizes computed from bootstrap resampling of listwise rankings using Borda average score differences. Across these checkpoints (Llama-3.2 at 1B/3B and Llama-3.1 at 8B), the panel advantage grows with model size: the Soft–Full gap increases from 0.038 (1B) to 0.045 (3B) to 0.076 (8B), and the average panel advantage (mean of Hard/Soft minus mean of Full/US-Rep) increases monotonically as well (Appendix A.13).

Judge reliability. To validate that ranking outcomes are not dominated by judge noise, we measure agreement across the 5 judge rankings per question. Across all questions, average inter-judge Kendall τ is 0.776 and Fleiss' κ on the #1 choice is 0.710, indicating substantial agreement. We treat the resulting 15,000 rankings as repeated measurements and use majority vote (or vote counts) in derived pairwise analyses.

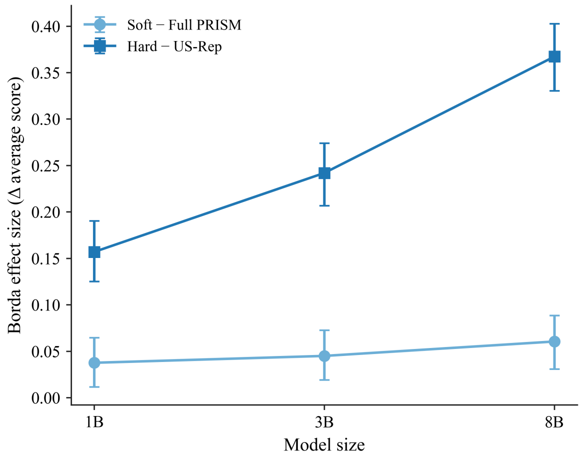

Figure 2. **Panel advantage grows with model size.** Effect sizes are computed from bootstrap resampling of listwise rankings (Borda average score differences). The Soft Panel vs. Full PRISM gap and the Hard Panel vs. US-Rep gap both increase from 1B to 3B to 8B. All error bars are 95% bootstrap CIs (n=1000).

Check	BT-score summary
Soft weighting (k=145)	Hard 0.339; Soft π_i^2 0.014; clipped Soft 0.013; Soft $\sqrt{\pi_i}$ -0.025; Soft π_i -0.141; Full -0.201.
Panel size	Hard: k=145 0.274, k=100 0.030, k=50 -0.034. Soft: k=100 0.107, k=50 0.090, k=145 -0.078.
Gradients	Hard has lower normalized ℓ_1/ℓ_2 sparsity score than Soft (0.375 vs. 0.379) and higher stable rank (11.96 vs. 9.59).
Community Alignment	Soft-US-k350 0.141, Full-US 0.125, Hard-US-k350 0.016, dataset balanced subset -0.054, Base -0.228.

Table 1. **Additional robustness checks.** Scores are Bradley–Terry log-abilities within each comparison set. The checks support the main representativeness result but show data-regime dependence: Hard is strongest on PRISM, while Soft is more competitive on Community Alignment when hard filtering discards many comparisons. Details are in Appendix A.7.

Pairwise win rates. We compute question-level bootstrap win probabilities for every model against every baseline, reported as the vote-level win rate $\Pr(\text{model beats baseline})$ (bootstrap n=1000). For Llama-3.1-8B, Soft beats Full with probability 0.520 (95% CI $[0.506, 0.534]$), Hard beats Full with probability 0.618 (95% CI $[0.602, 0.634]$), and Hard beats US-Rep with probability 0.576 (95% CI $[0.561, 0.592]$). This pairwise view complements the rank-aggregation picture in Figure 1 by quantifying effect sizes in a common metric.

Robustness checks. Table 1 summarizes the additional experiments added after the main PRISM evaluation. Varying the Soft Panel transform shows that all sortition-weighted variants improve over Full PRISM, while more concentrated transforms (π_i^2 and clipping) outperform the default linear transform in this setting. Varying panel size shows that smaller Soft panels can improve performance by concentrating inclusion-probability mass, whereas Hard Panel performance is best at the largest feasible panel size. Gradient diagnostics show that Hard and Soft do not merely implement the same update in practice: the hard-panel gradients are slightly more concentrated and have higher stable rank. On Community Alignment, the overall representativeness effect replicates against the dataset's own balanced subset and base model, but Soft outperforms Hard, suggesting a tradeoff between representativeness concentration and the cost of discarding high-quality comparisons.

Limitations. Our U.S. demographic targets are applied to the PRISM rater pool, which includes raters from multiple locales, and some attributes (e.g., religion) require survey-based rather than Census-based population targets. We treat this as a practical compromise to avoid infeasibility and report locale composition in Appendix A.4. More broadly, sortition-weighted preference learning relies on preference datasets that record annotator demographics; PRISM and Community Alignment are unusual in making such metadata available, and scaling these methods will require future preference data collection efforts to record and govern demographic attributes responsibly. The approach should therefore be understood as "reweighting the willing": if a group is absent from the rater pool, sortition cannot recover that group's preferences. Demographic quotas are also imperfect proxies for values and do not by themselves resolve conflicts between groups or provide a full theory of minority rights. Finally, our primary evaluation target is a U.S.-elicited constitution; the method is population-agnostic, but claims about any other public require a corresponding target distribution and evaluation benchmark.

4. Conclusion

We studied sortition-weighted preference learning as a way to make the composition of RLHF/DPO feedback more explicit and auditable for a target public. Using PRISM human preferences and demographic metadata, we instantiated two schemes: a *Hard Panel* that trains on preferences from a representative mini-public sampled via sortition, and a *Soft Panel* that reweights all preferences by the rater's inclusion probability under the same sortition lottery. Across multiple model sizes and a constitution derived from a representative U.S. panel, Hard Panel and US-Rep rank above Soft and Full PRISM, while Soft improves over Full PRISM under multiple aggregation families.

Overall, these results support the hypothesis that explicitly incorporating representativeness constraints into preference learning can shift model behavior toward values elicited from the relevant public.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Bakker, M. A., Chadwick, M. J., Sheahan, H. R., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M. M., and Summerfield, C. Fine-tuning language models to find agreement among humans with diverse preferences, 2022. URL https://arxiv.org/abs/2211.15006.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Buyl, M., Khalaf, H., Mayrink Verdun, C., Monteiro Paes, L., Vieira Machado, C. C., and du Pin Calmon, F. Ai alignment at your discretion. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3046–3074, 2025.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

Copeland, A. H. A reasonable social welfare function. Mimeo, University of Michigan, Ann Arbor, 1951.

de Borda, J.-C. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

Flanigan, B., Gölz, P., Gupta, A., Hennig, B., and Procaccia, A. D. Fair algorithms for selecting citizens' assemblies. *Nature*, 596(7873):548–552, 2021.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., and Ganguli, D. Collective constitutional AI: Aligning a language model with public input, 2024. URL https://arxiv.org/abs/2406.07814.

Kemeny, J. G. and Snell, J. L. *Mathematical Models in the Social Sciences*. Blaisdell Publishing Company, New York, 1962. A division of Ginn and Company.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024. URL https://arxiv.org/abs/2404.16019.

Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., Williams, D., Campbell-Gillingham, L., Thacker, P., Botvinick, M., et al. Human-centred mechanism design with democratic ai. *Nature Human Behaviour*, 6(10):1398–1407, 2022.

Lanchantin, J., Chen, A., Dhuliawala, S., Yu, P., Weston, J., Sukhbaatar, S., and Kulikov, I. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Luce, R. D. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

Mallows, C. L. Non-null ranking models. *Biometrika*, 44 (1-2):114–130, 1957. doi: 10.1093/biomet/44.1-2.114.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Plackett, R. L. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975. doi: 10.2307/2346567.

Precup, D., Sutton, R. S., and Singh, S. Eligibility traces for off-policy policy evaluation. 2000.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model, 2023. URL https://arxiv.org/abs/2305.18290.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023. URL https://arxiv.org/abs/2303.17548.

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., and Vosoughi, S. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025.

Smith, G. A. About three-in-ten U.S. adults are now religiously unaffiliated. Pew Research Center, 2021. URL https://www.pewresearch.org/religion/2021/12/14/about-three-in-ten-u-s-adults-are-now-religiously-unaffiliated/.

U.S. Census Bureau. American Community Survey (ACS) 2020 1-year estimates. U.S. Census Bureau, 2020. URL https://data.census.gov/.

Young, H. P. Condorcet's theory of voting. *American Political Science Review*, 82(4):1231–1244, 1988. doi: 10.2307/1961757.

Zhang, L. H., Milli, S., Jusko, K. L., Smith, J., Amos, B., Bouaziz, W., Revel, M., Kussman, J., Sheynin, Y., Titus, L., Radharapu, B., Yu, J., Sarma, V., Rose, K., and Nickel, M. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4NtoAVqfhA.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Stoica, I., and Zhang, H. Judging LLM-as-a-judge with MT-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A. Additional Experimental Details

A.1. Software, Data, and Responsible Use

All code, configuration files, and scripts needed to reproduce the experiments and analyses reported in this paper are provided in the supplementary materials. The goal of sortition-weighted preference learning is to make alignment data more representative of a specified target public by replacing ad hoc rater pools with panels or weights generated by a transparent, quota-constrained sortition procedure. If deployed responsibly, such methods could reduce systematic misalignment that arises when feedback disproportionately reflects a narrow, self-selected group, and could make alignment decisions for public-facing systems easier to audit.

At the same time, demographic representativeness is not a complete notion of legitimacy: quotas are imperfect proxies for values, population targets are approximate, and the resulting systems may still underrepresent important viewpoints not captured by the chosen attributes. Using demographic data in model training and evaluation raises privacy and misuse concerns; we therefore rely only on de-identified rater IDs and public aggregate statistics, treat ambiguous categories as slack, and recommend strong governance and transparency when applying such techniques in practice. LLM-as-a-judge evaluations can also inherit biases from the judge model; we mitigate this with repeated judging, randomized response ordering, and diagnostic analyses, but human oversight remains important before drawing high-stakes conclusions.

A.2. Basic Identities for Inclusion Probabilities

Lemma A.1 (Inclusion probabilities sum to the panel size)**.** *Let $S \sim \pi_{\text{panel}}$ be a random panel of fixed size $|S| = k$ and let $\pi_i = \Pr[i \in S]$. Then $\sum_{i \in \mathcal{I}} \pi_i = k$.*

Proof. Let $\mathbf{1}\{i \in S\}$ be the indicator that rater i is included. Then $|S| = \sum_i \mathbf{1}\{i \in S\} = k$ almost surely. Taking expectations and using linearity, $\sum_i \Pr[i \in S] = \sum_i \mathbb{E}[\mathbf{1}\{i \in S\}] = \mathbb{E}[|S|] = k$. □

Lemma A.2 (Soft weights recover the expected hard-panel objective)**.** *Let $L_i(\theta)$ denote the per-rater mean loss*

$$L_i(\theta) \;:=\; \frac{1}{N_i} \sum_{j : r_j = i} \ell(\theta; x_j, y_j^+, y_j^-),$$

and recall the hard-panel objective for a fixed-size panel $|S| = k$,

$$\mathcal{L}_{\text{hard}}(\theta \mid S) = \frac{1}{k} \sum_{i \in S} L_i(\theta).$$

Define the soft objective as a rater-weighted mean,

$$\mathcal{L}_{\text{soft}}(\theta) = \frac{1}{\sum_i w_i} \sum_{i=1}^{n} w_i \, L_i(\theta).$$

Then with $w_i = \pi_i$ (and $\sum_i \pi_i = k$), we have

$$\mathbb{E}_{S \sim \pi_{\text{panel}}}[\mathcal{L}_{\text{hard}}(\theta \mid S)] = \mathcal{L}_{\text{soft}}(\theta).$$

Proof. Taking expectations and using linearity,

$$\mathbb{E}_S[\mathcal{L}_{\text{hard}}(\theta \mid S)] = \mathbb{E}_S\left[\frac{1}{k} \sum_{i \in S} L_i(\theta)\right] = \frac{1}{k} \sum_{i=1}^{n} \Pr[i \in S] \, L_i(\theta) = \frac{1}{k} \sum_{i=1}^{n} \pi_i \, L_i(\theta).$$

By Lemma A.1, $\sum_i \pi_i = k$, so this is exactly $\frac{1}{\sum_i \pi_i} \sum_i \pi_i \, L_i(\theta) = \mathcal{L}_{\text{soft}}(\theta)$. □

A.3. Training Dataset Sizes

Table 2 summarizes the number of DPO preference pairs produced by our PRISM pre-processing pipeline for each training condition. Counts are computed *after* converting PRISM interactions into (prompt, chosen, rejected) pairs and filtering placeholder prompts/responses and survey-only raters without preference data.

Training condition	# pairs	# raters	Notes
Full PRISM	57,750	1,394	unweighted
Soft Panel	57,750	1,394	weighted by π_i
Hard Panel	5,737	145	panel-filtered
US-Rep	9,541	230	PRISM US-Rep subset

Table 2. Dataset sizes for each training condition. "# raters" counts unique PRISM rater IDs appearing in the prepared pairs for that condition.

A.4. Locale Composition of Panels

Our sortition quotas target U.S. demographic proportions, but our default configuration does not filter PRISM raters by `study_locale`. In the resulting U.S.-target hard panel (panel size k=145), 56/145 raters appearing in the training pairs have `study_locale=us`; the remainder are primarily from the UK (27) and other locales. This is a practical compromise to enlarge the feasible rater pool; future work should evaluate panels drawn strictly from U.S. locales when sufficient data is available.

A.5. Panel Feasibility and Quota Satisfaction

We verify that our quota specifications admit feasible panels at our chosen size k=145 and tolerance τ=0.1 (relative bounds), using the LEXIMIN sortition algorithm (Flanigan et al., 2021). For the U.S. target configuration (seven attributes), we sample a feasible panel of size 145 with zero quota violations from the filtered pool of raters with preference data; across constrained (non-slack) categories the maximum absolute deviation from target proportions is 4.4 percentage points.

A.6. Soft-Panel Selection-Probability Diagnostics

For the soft panel, we compute exact selection probabilities $\{\pi_i\}$ from the LEXIMIN-induced lottery and verify basic correctness properties. In the U.S. target configuration, $\sum_i \pi_i \approx k$=145 as required by linearity of expectation, and the PRISM training pool has no raters with zero inclusion probability after preprocessing. The positive inclusion-probability vector is concentrated: Gini coefficient 0.56, max/min ratio 208.9, and coefficient of variation 2.33. After filtering survey-only raters, nearly all selection-probability mass corresponds to raters with preference data; any remaining missing mass reflects filtered or empty preference comparisons rather than invalid weights.

A.7. Additional Robustness Experiments

This section reports the additional experiments summarized in Table 1. All PRISM ablations use Llama-3.1-8B and the same constitutional evaluation protocol as the main experiment.

Soft weighting functions. We compare monotone transforms $f(\pi_i)$ for Soft Panel training at k=145. The clipped transform uses

$$f(\pi_i) = \min\{\max\{\pi_i, q_{.05}\}, q_{.95}\},$$

where $q_{.05}$ and $q_{.95}$ are the 5th and 95th percentiles of the positive inclusion-probability distribution. All transforms are self-normalized during minibatch training.

Model	BT score [95% CI]	Wins
Hard Panel	0.339 [0.294, 0.385]	9004
Soft Panel (π_i^2)	0.014 [-0.023, 0.047]	7565
Soft Panel (clipped π_i)	0.013 [-0.027, 0.052]	7561
Soft Panel ($\sqrt{\pi_i}$)	-0.025 [-0.063, 0.010]	7388
Soft Panel (π_i)	-0.141 [-0.183, -0.094]	6874
Full PRISM	-0.201 [-0.246, -0.158]	6608

Table 3. Soft weighting-function ablation. Scores are Bradley–Terry log-abilities fit to listwise-judge-derived pairwise preferences.

Panel-size sensitivity. We vary panel size for Hard and Soft Panel training while keeping the same 3538 optimizer updates used in the main PRISM runs. For Soft Panel, smaller k concentrates inclusion-probability mass onto fewer raters; for Hard Panel, smaller k also discards more training data.

Comparison	Model	BT score	Wins
Hard k	Hard Panel (k=145)	0.274	7016
Hard k	Hard Panel (k=100)	0.030	6112
Hard k	Hard Panel (k=50)	-0.034	5874
Hard k	US-Rep	-0.051	5812
Hard k	Full PRISM	-0.220	5186
Soft k	Soft Panel (k=100)	0.107	6398
Soft k	Soft Panel (k=50)	0.090	6336
Soft k	US-Rep	0.025	6095
Soft k	Soft Panel (k=145)	-0.078	5707
Soft k	Full PRISM	-0.144	5464
Best variants	Hard Panel (k=145)	0.264	5281
Best variants	Soft Panel (k=100)	0.059	4674
Best variants	US-Rep	-0.068	4299
Best variants	Full PRISM	-0.255	3746

Table 4. Panel-size sensitivity under matched optimizer updates. Scores are Bradley–Terry log-abilities within each comparison set.

Gradient diagnostics. We compute gradient statistics for the 8B Hard and Soft Panel models on matched minibatches. The normalized ℓ_1/ℓ_2 statistic is $\|g\|_1/(\sqrt{d}\|g\|_2)$, where lower values indicate more concentrated gradients.

Metric	Hard	Soft	Delta / CI
Normalized ℓ_1/ℓ_2	0.3747	0.3791	-0.0044 [-0.0067, -0.0022]
Pairwise gradient cosine	0.0020	-0.0013	0.0033 [-0.0054, 0.0114]
Stable rank	11.96	9.59	2.37
Effective rank	48.01	47.64	0.37

Table 5. Gradient diagnostics for Llama-3.1-8B Hard and Soft Panel training. Bootstrap intervals are shown where computed in the diagnostic artifacts.

Community Alignment replication. We also apply the same data-construction idea to the Community Alignment dataset (Zhang et al., 2026). We filter to U.S. annotators and construct a k=350 target using the paper-implied age, gender, and ethnicity marginals used by the dataset's balanced subset. We compare Hard-US-$k350$, Soft-US-$k350$, Full-US, the dataset's own balanced subset, and the base model using the same constitutional judging pipeline.

Model	BT score [95% CI]	Wins
Soft-US-$k350$	0.141 [0.103, 0.180]	6524
Full-US	0.125 [0.087, 0.165]	6467
Hard-US-$k350$	0.016 [-0.023, 0.055]	6058
CA balanced subset	-0.054 [-0.095, -0.013]	5798
Base	-0.228 [-0.270, -0.182]	5153

Table 6. Community Alignment replication on the U.S. subset. Scores are Bradley–Terry log-abilities under the constitutional evaluation protocol.

A.8. Ablations

Ordering bias. Position bias is a known concern in LLM-as-a-judge settings and can be substantial depending on prompt and format (Shi et al., 2025). In our 8B evaluation (5 models; 15,000 rankings), the Pearson correlation between the *presented position* and the final rank is $\rho = 0.047$, and the probability that the response shown at a given position is ranked

#1 ranges from 0.179 to 0.224 (a 4.5 percentage-point spread across positions). We therefore randomize response order independently for each judge call in all reported results.

Verbosity bias. We test whether response length predicts preferences by correlating response length with rank and by checking whether longer responses win implied pairwise comparisons. Across all 75,000 ranked responses (3000 questions $\times 5$ judges $\times 5$ models), correlations between response length and rank are small: $\rho = -0.056$ (characters) and $\rho = -0.044$ (words). In implied pairwise comparisons (149,290 total), the longer response wins 53.7% of the time, indicating only a mild verbosity preference that does not explain the main ranking.

A.9. Mallows Model Fit and Stability

We fit a Kendall-distance Mallows model by (i) identifying the consensus ranking π_0 via brute force search over $5! = 120$ permutations (for the five-model setting) and (ii) estimating the dispersion parameter ϕ by one-dimensional likelihood maximization. Bootstrap refits (n=1000) show the consensus is stable (exact-match rate 1.00), with Hard ranked #1 and US-Rep ranked #2 in all bootstrap resamples in the 8B setting (see also Figure 1).

A.10. Raw Scoring Outputs

For reproducibility and to make the plots in Figure 1 numerically explicit, Tables 8, 9, and 10 report the raw scores computed from our evaluation artifacts. We additionally report the corresponding raw scores for the normalized-step 8B ablation in Figure 4 and Tables 11–13. For completeness we also report the corresponding raw-score tables for Llama-3.2-1B (Tables 14–16) and Llama-3.2-3B (Tables 17–19). Listwise scores (Borda, Copeland, Kemeny, PL, Mallows) use all 15,000 judge rankings; BT is fitted on 30,000 pairwise preferences derived from the listwise rankings (8B setting).

A.11. Training Update Magnitudes

Although our objectives are normalized to preserve equal voter voice and comparable per-step loss scale, the realized update magnitudes can still differ because data distributions differ across conditions. We quantify this by measuring parameter-change magnitude from the base checkpoint after the matched 3538-step training budget. For Llama-3.1-8B, relative ℓ_2 deltas are Full 0.00240, Soft 0.00235, Hard 0.00275, US-Rep 0.00291, indicating larger updates for the smaller hard/US-Rep datasets. We therefore also train normalized-step variants for Hard and US-Rep (Hard: 2200 steps; US-Rep: 2100 steps, chosen by decreasing steps in increments of 100 until the relative ℓ_2 matched Full/Soft) and find that the ranking order persists even when update magnitudes are matched (see Figure 4), though the gaps shrink.

Model	Steps	$\|\Delta\theta\|_2$	RMS ($\times 10^{-5}$)	Mean abs ($\times 10^{-6}$)	Rel. ℓ_2
Full PRISM	3538	2.996	3.344	6.544	0.002400
Soft Panel	3538	2.938	3.279	6.499	0.002354
Hard Panel	3538	3.435	3.834	7.070	0.002752
US-Rep	3538	3.626	4.047	7.483	0.002905
Hard Panel (normalized)	2200	3.021	3.371	6.104	0.002420
US-Rep (normalized)	2100	2.940	3.281	5.971	0.002355

Table 7. Training update magnitudes for Llama-3.1-8B measured as parameter deltas from the base checkpoint. RMS and mean-absolute values are computed over all parameters. "Rel. ℓ_2" denotes $\|\Delta\theta\|_2/\|\theta\|_2$ for the base checkpoint.

A.12. Reproducibility Details

All experiments use a single H200 GPU, full-parameter fine-tuning in bf16, and a frozen reference model initialized from the same base checkpoint. We fix the optimization budget to 3538 updates for every model and condition, with the same effective batch size (per-device batch size $\times$ gradient accumulation, single GPU), and resample smaller datasets with replacement to reach the same update count. Evaluation uses identical decoding settings and a shared chat template across all models. Judging uses GPT-5.2 at temperature 0 with JSON-formatted outputs and randomized response order for each judge call. All bootstrap confidence intervals use ranking-level resampling with $n = 1000$.

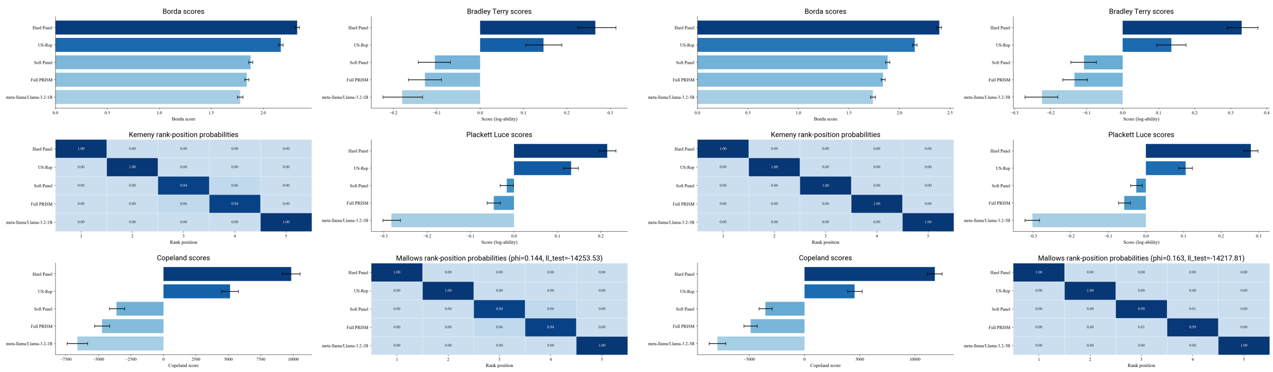

Figure 3. Score grids for Llama-3.2-1B (left) and Llama-3.2-3B (right) under the same evaluation protocol.

A.13. Model-Size Scaling

We include the 1B and 3B score grids in Figure 3. The same ordering (Hard $>$ US-Rep $>$ Soft $>$ Full $>$ Base) appears at all sizes, while effect sizes increase with model capacity (see Figure 2 in the main text).

Model	Borda avg	Copeland win rate	Copeland
Hard Panel	2.459 [2.437, 2.480]	0.615 [0.609, 0.620]	13,762 [13,100, 14,414]
US-Rep	2.118 [2.097, 2.139]	0.530 [0.524, 0.535]	3,542 [2,900, 4,174]
Soft Panel	2.006 [1.986, 2.025]	0.501 [0.497, 0.506]	167 [-416, 752]
Full PRISM	1.930 [1.909, 1.950]	0.482 [0.477, 0.487]	-2,113 [-2,720, -1,508]
Base	1.488 [1.463, 1.514]	0.372 [0.366, 0.379]	-15,358 [-16,108, -14,576]

Table 8. Raw Borda and Copeland scores computed from 15,000 listwise rankings, with 95% bootstrap confidence intervals (ranking-level bootstrap, n=1000). "Borda avg" is the average per-ranking Borda score (higher is better). "Copeland" is wins minus losses in the implied pairwise tournament (higher is better). Numeric values are reported in the released artifacts for each model size.

Method	Consensus ranking / parameters
Kemeny-Young	Hard $\succ$ US-Rep $\succ$ Soft $\succ$ Full $\succ$ Base; exact-match (bootstrap) 1.00
Mallows (Kendall)	π_0 matches Kemeny; $\phi = 0.219$; $\ell_{\text{test}} = -14107.4$

Table 9. Kemeny and Mallows consensus summaries (computed from 15,000 listwise rankings; bootstrap n=1000 where reported).

Model	BT score (mean [95% CI])	PL score (mean [95% CI])
Hard Panel	0.384 [0.345, 0.427]	0.375 [0.357, 0.393]
US-Rep	0.102 [0.062, 0.141]	0.113 [0.097, 0.129]
Soft Panel	0.017 [-0.018, 0.054]	0.071 [0.055, 0.087]
Full PRISM	-0.070 [-0.110, -0.034]	0.026 [0.009, 0.041]
Base	-0.432 [-0.484, -0.381]	-0.584 [-0.604, -0.561]

Table 10. Bradley–Terry (BT) and Plackett–Luce (PL) scores with 95% bootstrap confidence intervals (bootstrap n=1000). BT is fit to the pairwise preferences derived from listwise rankings. Scores are log-abilities up to an additive constant. Numeric values are reported in the released artifacts for each model size.

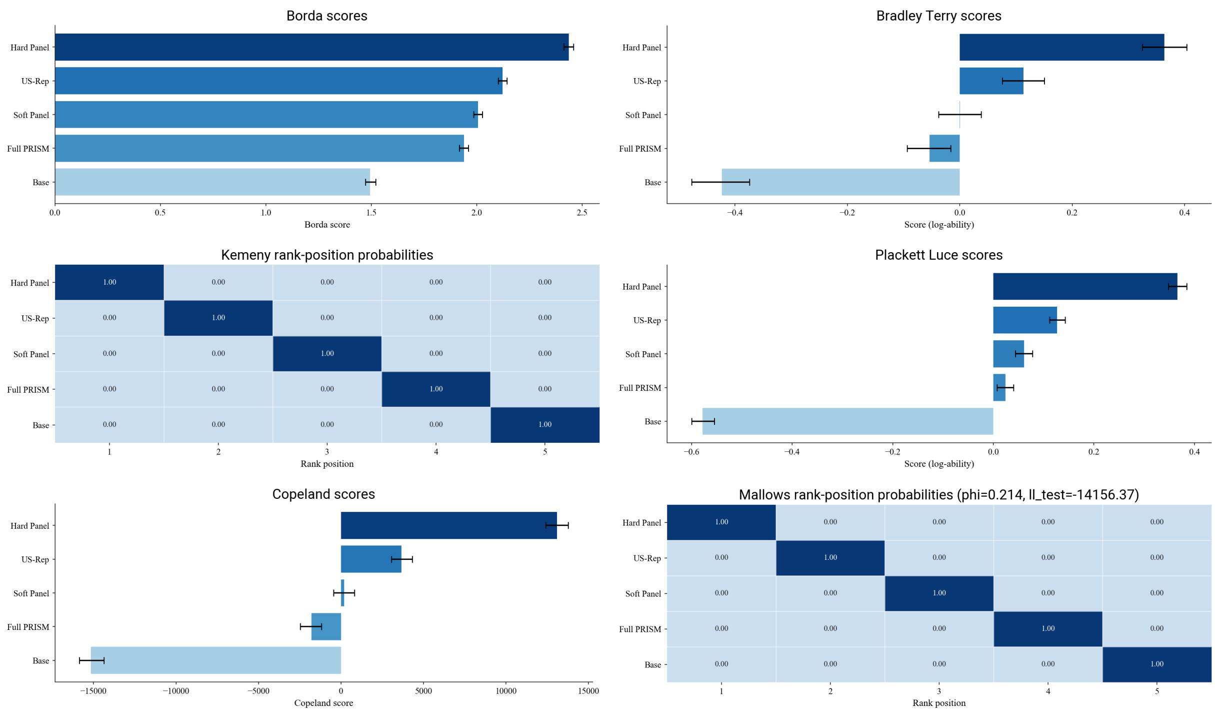

Figure 4. Normalized-step score grid for Llama-3.1-8B, where Hard Panel and US-Rep are retrained for fewer steps (Hard: 2200, US-Rep: 2100) to match the update magnitude of Full/Soft.

Model	Borda avg	Copeland win rate	Copeland
Hard Panel	2.437 [2.414, 2.459]	0.609 [0.603, 0.615]	13,110 [12,416, 13,770]
US-Rep	2.122 [2.102, 2.144]	0.531 [0.525, 0.536]	3,667 [3,052, 4,314]
Soft Panel	2.006 [1.985, 2.027]	0.502 [0.496, 0.507]	182 [-445, 818]
Full PRISM	1.940 [1.918, 1.961]	0.485 [0.479, 0.490]	-1,801 [-2,468, -1,182]
Base	1.495 [1.472, 1.521]	0.374 [0.368, 0.380]	-15,158 [-15,852, -14,358]

Table 11. Llama-3.1-8B normalized-step Borda and Copeland scores (bootstrap n=1000).

Model	BT score (mean [95% CI])	PL score (mean [95% CI])
Hard Panel	0.363 [0.325, 0.404]	0.366 [0.348, 0.384]
US-Rep	0.114 [0.076, 0.150]	0.127 [0.112, 0.143]
Soft Panel	0.000 [-0.038, 0.038]	0.061 [0.044, 0.078]
Full PRISM	-0.054 [-0.094, -0.016]	0.023 [0.007, 0.040]
Base	-0.423 [-0.477, -0.374]	-0.578 [-0.600, -0.555]

Table 12. Llama-3.1-8B normalized-step BT and PL scores (bootstrap n=1000).

Method	Consensus ranking / parameters
Kemeny-Young	Hard $\succ$ US-Rep $\succ$ Soft $\succ$ Full $\succ$ Base; exact-match (bootstrap) 1.00
Mallows (Kendall)	π_0 matches Kemeny; $\phi = 0.214$; $\ell_{\text{test}} = -14156.4$

Table 13. Llama-3.1-8B normalized-step Kemeny and Mallows summaries (bootstrap n=1000 where reported).

Llama-3.1-8B normalized-step raw scores.

Model	Borda avg (mean [95% CI])	Copeland (mean [95% CI])
Hard Panel	2.328 [2.305, 2.350]	9,836.294 [9,137.950, 10,514.050]
US-Rep	2.171 [2.149, 2.192]	5,125.712 [4,469.900, 5,768.400]
Soft Panel	1.880 [1.861, 1.900]	-3,600.814 [-4,156.050, -2,987.950]
Full PRISM	1.842 [1.823, 1.862]	-4,729.654 [-5,302.100, -4,127.950]
Base	1.779 [1.753, 1.805]	-6,631.538 [-7,416.050, -5,839.750]

Table 14. Llama-3.2-1B Borda and Copeland scores (bootstrap n=1000).

Model	BT score (mean [95% CI])	PL score (mean [95% CI])
Hard Panel	0.267 [0.225, 0.313]	0.215 [0.196, 0.235]
US-Rep	0.146 [0.105, 0.188]	0.132 [0.113, 0.148]
Soft Panel	-0.106 [-0.143, -0.069]	-0.017 [-0.032, -0.002]
Full PRISM	-0.127 [-0.166, -0.089]	-0.047 [-0.062, -0.032]
Base	-0.180 [-0.224, -0.133]	-0.283 [-0.302, -0.262]

Table 15. Llama-3.2-1B BT and PL scores (bootstrap n=1000).

Method	Consensus ranking / parameters
Kemeny-Young	Hard $\succ$ US-Rep $\succ$ Soft $\succ$ Full $\succ$ Base; exact-match 0.94
Mallows (Kendall)	π_0 matches Kemeny; $\phi = 0.144$; $\ell_{\text{test}} = -14253.5$

Table 16. Llama-3.2-1B Kemeny and Mallows summaries (bootstrap n=1000 where reported).

Llama-3.2-1B raw scores.

Model	Borda avg (mean [95% CI])	Copeland (mean [95% CI])
Hard Panel	2.393 [2.370, 2.415]	11,797.238 [11,102.000, 12,446.350]
US-Rep	2.151 [2.130, 2.174]	4,537.120 [3,905.850, 5,208.400]
Soft Panel	1.882 [1.862, 1.902]	-3,542.388 [-4,134.000, -2,931.950]
Full PRISM	1.837 [1.817, 1.857]	-4,892.736 [-5,500.100, -4,283.950]
Base	1.737 [1.712, 1.762]	-7,899.234 [-8,642.050, -7,129.950]

Table 17. Llama-3.2-3B Borda and Copeland scores (bootstrap n=1000).

Model	BT score (mean [95% CI])	PL score (mean [95% CI])
Hard Panel	0.331 [0.290, 0.375]	0.279 [0.260, 0.297]
US-Rep	0.135 [0.094, 0.175]	0.106 [0.087, 0.123]
Soft Panel	-0.108 [-0.144, -0.073]	-0.026 [-0.041, -0.010]
Full PRISM	-0.133 [-0.166, -0.098]	-0.057 [-0.073, -0.041]
Base	-0.225 [-0.271, -0.180]	-0.302 [-0.322, -0.283]

Table 18. Llama-3.2-3B BT and PL scores (bootstrap n=1000).

Method	Consensus ranking / parameters
Kemeny-Young	Hard $\succ$ US-Rep $\succ$ Soft $\succ$ Full $\succ$ Base; exact-match 1.00
Mallows (Kendall)	π_0 matches Kemeny; $\phi = 0.163$; $\ell_{\text{test}} = -14217.8$

Table 19. Llama-3.2-3B Kemeny and Mallows summaries (bootstrap n=1000 where reported).

Llama-3.2-3B raw scores.