PLURALISTIC PREFERENCE ALIGNMENT VIA SORTITION-WEIGHTED RLHF

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference-based alignment methods such as reinforcement learning from human feedback (RLHF) and Direct Preference Optimization (DPO) implicitly determine whose values a model learns through the composition of their rater pools. Because these pools are often convenience samples, the resulting preference signal may systematically overrepresent some demographic groups while marginalizing others. We introduce *sortition-weighted preference learning*, a representativeness-aware framework that incorporates algorithmic sortition—the sampling mechanism underlying citizens' assemblies—into preference-based fine-tuning. We study two implementations. *Hard Panel* trains exclusively on preferences from a quota-satisfying mini-public selected through sortition, whereas *Soft Panel* retains the full dataset while weighting raters according to their inclusion probabilities under the same sampling design. Using PRISM rater demographics and preference data, we fine-tune Llama models with DPO and evaluate their behavior against a 75-clause constitution elicited from a representative panel of U.S. residents. Across multiple preference-aggregation rules, Hard Panel achieves the strongest constitutional alignment, while Soft Panel consistently improves upon training on the full, unadjusted PRISM sample. Additional analyses of weighting functions, panel sizes, training gradients, and the Community Alignment dataset reveal an important data-quality tradeoff: hard selection performs best when representative subsamples retain sufficiently informative preferences, whereas soft weighting becomes more competitive when filtering would discard substantial amounts of high-quality data. Our results demonstrate that the demographic composition of preference feedback is an empirically consequential design choice and provide an explicit, auditable mechanism for aligning models with a specified target population. We also provide theoretical analysis of our framework.

1 INTRODUCTION

As AI systems become increasingly embedded in everyday life, supporting decision-making (Kleinberg et al., 2018), informing policy analysis (Valle-Cruz et al., 2020), and assuming responsibility for delegated tasks–it is essential that their development reflects the diverse perspectives of affected stakeholders. Preference-based alignment methods, such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) and Direct Preference Optimization (DPO) (Rafailov et al., 2023), shape model behavior by translating human judgments into optimization signals. Consequently, the composition of the rater pool implicitly determines whose preferences and values the system learns to represent. In practice, however, rater pools are often convenience samples rather than representative cross-sections of affected populations. Preference-learning pipelines may therefore systematically amplify the views of groups that are easier to recruit while underrepresenting those that are demographically, geographically, or politically scarce.

Motivated by sortition (Bouricius, 2013), a longstanding mechanism in political philosophy and democratic practice used to form citizens' assemblies that deliberate on public issues, we study *sortition weighted preference learning*, a simple framework for making participation in RLHF and DPO more explicit and auditable. Given a target population distribution and a pool of raters with demographic metadata, we use algorithmic sortition (Flanigan et al., 2021) to construct a lottery over quota feasible mini publics. This lottery induces two training schemes. **Hard Panel** draws one feasible panel and trains exclusively on the preferences of its members. **Soft Panel** retains all

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

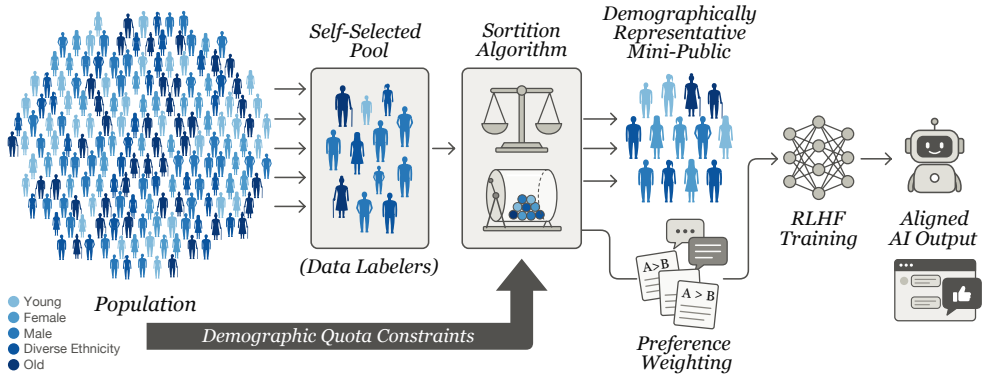

Figure 1: Sortition-weighted preference learning. A target public defines demographic quota constraints; a biased rater pool supplies preferences and metadata; algorithmic sortition induces either a sampled mini-public for Hard Panel training or inclusion-probability weights for Soft Panel training.

observed preferences but weights each rater according to their probability of inclusion under the sortition lottery.

Our framing is deliberately pluralistic rather than universalist. We do not seek to identify or recover a single set of universal values. Instead, consistent with the notion of *distributional pluralism* (Sorensen et al., 2024), we ask whether a preference learning pipeline can more faithfully represent the values and judgments of a specified target public through transparent demographic constraints and standard preference optimization objectives. This perspective is particularly relevant when a model is intended to serve a known population, when demographic representativeness is an explicit governance requirement, or when researchers wish to audit how the composition of preference data shapes downstream model behavior.

We evaluate on PRISM (Kirk et al., 2024), which pairs preference data with rater demographics, and measure downstream alignment against a 75-clause constitution elicited from a representative U.S. panel (Huang et al., 2024). Across Bradley–Terry, Plackett–Luce, Borda, Copeland, Kemeny-Young, and Mallows aggregation, Hard Panel ranks first and Soft Panel improves over Full PRISM. We additionally study model size, alternative weighting functions, panel-size sensitivity, gradient diagnostics, and transfer to Community Alignment (Zhang et al., 2026). The results support the core representativeness effect while also clarifying a practical tradeoff: concentrated representative panels can help, but hard filtering can be costly when it discards many high-quality comparisons.

Contributions. We contribute: (i) a sortition-based construction for representative preference-learning data; (ii) Hard and Soft Panel objectives for DPO/RLHF, with an identity showing that linear Soft Panel weighting recovers the expected Hard Panel objective; (iii) a PRISM evaluation against externally elicited representative constitutional criteria; (iv) robustness studies showing how weighting concentration, panel size, model capacity, gradients, and dataset regime affect the Hard/Soft tradeoff and ; (v) theoretical analysis of our framework grounded in statistical sampling theory.

What problem this does and does not solve. The method does not remove the need for better data collection: if a group is absent from the rater pool, no reweighting procedure can recover its preferences. Its contribution is instead to separate three objects that are often conflated in alignment practice: the target public, the available rater pool, and the training electorate induced by sampling or weighting. This makes representativeness assumptions explicit, exposes infeasibility when the pool is too narrow, and gives practitioners a tunable correction strength rather than leaving participation bias implicit in a convenience sample.

2 Related Work

RLHF and direct preference objectives train models from human judgments (Christiano et al., 2017; Rafailov et al., 2023), while Constitutional AI uses explicit principles to steer behavior (Bai et al., 2022). (Zhong et al., 2024) study multi-party RLHF by learning heterogeneous party-specific preferences and aggregating them through social-welfare objectives, whereas our work focuses on the upstream problem of constructing a representative training electorate from a potentially biased rater pool. Our approach is compatible with these objectives: it changes how preference data are selected or weighted before optimization. Importance weighting is standard in off-policy learning (Munos et al., 2016; Precup et al., 2000); here the weights encode demographic representativeness rather than policy mismatch.

Pluralistic alignment asks the question whose preferences are represented when human values vary between groups and contexts (Gabriel, 2020; Conitzer et al., 2024; Sorensen et al., 2024). Prior work studies demographic opinion reflection in language models (Santurkar et al., 2023), preference heterogeneity and agreement (Bakker et al., 2022), diverse feedback collection (Lanchantin et al., 2025), and annotator discretion in alignment data (Buyl et al., 2025). Community Alignment provides a multilingual preference dataset for pluralistic alignment research (Zhang et al., 2026). Our focus is biased participation in feedback collection: we enforce representativeness at the level of who contributes to the optimization signal, using the same algorithmic sortition tools used for citizens' assemblies (Flanigan et al., 2021).

Our construction admits direct reading in design-based survey sampling. The LEXIMIN lottery π_{panel} is a fixed-size sampling design with balancing constraints (quota), closely related to balanced sampling via the cube method (Deville and Tillé, 2004); the π_i of equation 1 are its first-order inclusion probabilities and the π_{ij} of Theorem 3 are its second-order inclusion probabilities. Hard Panel training is design-based estimation on a single realized probability sample, whereas Soft Panel training replaces the sample with its design expectation (Theorem 1), corresponding to the survey identity that the expected sample mean equals the π-weighted frame mean. The weights are entered as π_i rather than Horvitz–Thompson $1/\pi_i$ (Horvitz and Thompson, 1952) because the roles of the frame and population are inverted. The rating pool is the biased object and the design is constructed to be representative, so π_i acts as a calibration or pseudo-weight applied to a nonprobability sample (Deville and Särndal, 1992; Elliott and Valliant, 2017). The group-mass bounds of Theorem 1 are the calibration property that weighted demographic margins match their targets within the quota tolerance. The self-normalized minibatch mean of Section 3 is the Hájek estimator (Hájek, 1971), whereas the known-normalizer estimator of Theorem 5 is Horvitz–Thompson. The $1/N_i$ normalization treats each rater as a primary sampling unit with equal cluster weight, and the transforms $f(\pi_i)$ correspond to survey weight trimming and smoothing (Potter, 1990).

3 Method

Let $\mathcal{P}$ be a target population described by demographic attributes $A^{(1)}, \ldots, A^{(d)}$, each with marginal target distribution $p^{(t)}_{\text{pop}}$. We observe a rater pool $\mathcal{I} = \{1, \ldots, n\}$, where each rater i has attributes $a_i = (a_i^{(1)}, \ldots, a_i^{(d)})$, and a preference dataset

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

where r_j is the rater who prefers response y_j^+ to y_j^- for context x_j. We use losses of the form $\ell(\theta; x, y^+, y^-)$, instantiated as DPO in our experiments.

Sortition. A panel is a subset $S \subseteq \mathcal{I}$ of fixed size k satisfying marginal quota bounds for each attribute/category,

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

Let $\mathcal{S}$ be the feasible-panel family. Following Flanigan et al. (2021), we use LEXIMIN to construct a lottery π_{panel} over $\mathcal{S}$ whose induced inclusion-probability vector is lexicographically max-min fair. The inclusion probability for rater i is

$$\pi_i = \Pr_{S \sim \pi_{\text{panel}}}[i \in S] = \sum_{S \in \mathcal{S}: i \in S} \pi_{\text{panel}}(S). \tag{1}$$

This gives max-min protection against near-zero inclusion probabilities for scarce demographic intersections and symmetry among raters indistinguishable with respect to quota attributes.

Hard Panel. Hard Panel samples one mini-public $S \sim \pi_{\text{panel}}$ and trains only on preferences from panel members. With N_i comparisons from rater i, define $L_i(\theta) = N_i^{-1} \sum_{j:r_j=i} \ell(\theta; x_j, y_j^+, y_j^-)$. The objective is

$$\mathcal{L}_{\text{hard}}(\theta \mid S) = \frac{1}{k} \sum_{i \in S} L_i(\theta). \tag{2}$$

The $1/N_i$ normalization implements one person, one voice, rather than weighting high-volume annotators more heavily.

Soft Panel. Soft Panel keeps all comparisons but weights rater i by $w_i = f(\pi_i)$:

$$\mathcal{L}_{\text{soft}}(\theta) = \frac{1}{Z} \sum_{i=1}^{n} w_i L_i(\theta), \qquad Z = \sum_i w_i. \tag{3}$$

Our main Soft Panel uses $f(\pi_i) = \pi_i$; ablations test $\sqrt{\pi_i}$, π_i^2, and clipped weights. In minibatch training we use a self-normalized weighted mean, the Hájek estimator, with per-example weight w_{r_b}/N_{r_b}, keeping optimization scale invariant to global weight rescaling. Because $\sum_i \pi_i = k$, linear Soft Panel recovers the expected Hard Panel objective as shown in theorem 1.

Concentration and weighting choices. Inclusion probabilities need not be uniform: scarce demographic intersections can receive higher probability under LEXIMIN. In our PRISM configuration the positive $\{\pi_i\}$ vector has Gini coefficient 0.56, max/min ratio 208.9, and coefficient of variation 2.33. This concentration motivates treating $f(\pi_i)$ as a design knob. Linear weighting exactly matches the expected hard-panel objective, square-root weighting compresses the tail, squared weighting amplifies the high-inclusion raters, and clipping controls extreme weights while preserving monotonicity.

4 Theoretical Analysis

All the proofs of the results can be found in Appendix A.2

Theorem 1 *The inclusion probabilities satisfy $\sum_{i=1}^n \pi_i = k$, and for every θ,*

$$\mathbb{E}_{S \sim \pi_{\text{panel}}}\big[\mathcal{L}_{\text{hard}}(\theta \mid S)\big] = \frac{1}{k} \sum_{i=1}^{n} \pi_i L_i(\theta) = \mathcal{L}_{\text{soft}}(\theta). \tag{4}$$

If each L_i is continuously differentiable (respectively, twice continuously differentiable), the identity 3 also holds for gradients (respectively, Hessians). Moreover, if every feasible panel satisfies $L_g \le |S \cap g| \le U_g$ for some demographic attribute g, then $L_g/k \le \sum_{i \in g} q_i \le U_g/k$, where $q_i = \pi_i/k$.

Theorem 2 *Suppose the relevant infima and expectations which appear below are finite. Then*

$$\mathbb{E}_{S \sim \pi_{\text{panel}}}\left[\inf_\theta \mathcal{L}_{\text{hard}}(\theta \mid S)\right] \le \inf_\theta \mathcal{L}_{\text{soft}}(\theta). \tag{5}$$

If $\mathcal{L}_{\text{soft}}$ attains its minimum at θ^*_{soft}, equality holds if and only if all positive-probability panel objectives share a common minimizer. Moreover, any Hard-trained output $\widehat{\theta}_{\text{hard}}$ satisfies

$$\mathcal{L}_{\text{soft}}(\widehat{\theta}_{\text{hard}}) \ge \mathcal{L}_{\text{soft}}(\theta^*_{\text{soft}}).$$

Neither conclusion determines the ordering of the external risks $R(\widehat{\theta}_{\text{hard}})$ and $R(\theta^*_{\text{soft}})$.

Theorem 3 *Fix θ and condition on the dataset. Let $\pi_{ij} = \Pr(i, j \in S)$, $\pi_{ii} = \pi_i$, $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$, $L = (L_1(\theta), \ldots, L_n(\theta))^\top$, and $g_i = \nabla L_i(\theta)$. Then*

$$\operatorname{Var}_S\big(\mathcal{L}_{\text{hard}}(\theta \mid S)\big) = \frac{1}{k^2} L^\top \Delta L,$$

$$\mathrm{Cov}_S\big(\nabla\mathcal{L}_{\mathrm{hard}}(\theta \mid S)\big) = \frac{1}{k^2}\sum_{i,j}\Delta_{ij}g_i g_j^\top.$$

Moreover, $\Delta \succeq 0$ *and* $\Delta\mathbf{1} = 0$*. For a realizable inclusion vector* π^* *and* $W \succeq 0$*, define* $g_S = k^{-1}\sum_{i\in S} g_i$ *and* $\bar{g} = k^{-1}\sum_i \pi_i^* g_i$*. A lottery minimizing* $\mathrm{tr}(W\,\mathrm{Cov}(g_S))$ *while preserving* π^* *exists and solves*

$$\begin{aligned} \min_{\lambda\geq 0} \quad & \sum_{S\in\mathcal{S}} \lambda_S (g_S - \bar{g})^\top W (g_S - \bar{g}) \\ \textit{subject to} \quad & \sum_{S\in\mathcal{S}} \lambda_S = 1, \qquad \sum_{S\in\mathcal{S}} \lambda_S v_S = \pi^*. \end{aligned}$$

Exact LEXIMIN has a unique inclusion vector but generally not a unique panel lottery. Consequently, identical Soft objectives can coexist with different Hard variability. Minimizing variance at one reference parameter does not guarantee better external alignment or uniformly lower variance throughout training.

Theorem 4 *Suppose every* $\mathcal{L}_{\mathrm{hard}}(\cdot \mid S)$ *is differentiable and* μ*-strongly convex on* $\mathbb{R}^d$*, with an attained minimum, and* $\mathcal{L}_{\mathrm{soft}}$ *is* L*-smooth. Let* $\theta^*_{\mathrm{hard}}(S)$ *and* θ^*_{soft} *denote their respective minimizers, and define*

$$\Sigma = \mathrm{Cov}_S\big(\nabla\mathcal{L}_{\mathrm{hard}}(\theta^*_{\mathrm{soft}} \mid S)\big).$$

Then

$$\mathbb{E}_S\|\theta^*_{\mathrm{hard}}(S) - \theta^*_{\mathrm{soft}}\|^2 \leq \frac{\mathrm{tr}(\Sigma)}{\mu^2},$$

$$0 \leq \mathbb{E}_S\big[\mathcal{L}_{\mathrm{soft}}(\theta^*_{\mathrm{hard}}(S))\big] - \mathcal{L}_{\mathrm{soft}}(\theta^*_{\mathrm{soft}}) \leq \frac{L\,\mathrm{tr}(\Sigma)}{2\mu^2}.$$

The panel-adaptation gain

$$G_{\mathrm{adapt}} := \mathcal{L}_{\mathrm{soft}}(\theta^*_{\mathrm{soft}}) - \mathbb{E}_S\left[\min_\theta \mathcal{L}_{\mathrm{hard}}(\theta \mid S)\right]$$

satisfies $0 \leq G_{\mathrm{adapt}} \leq \mathrm{tr}(\Sigma)/(2\mu)$*. If every* $\mathcal{L}_{\mathrm{hard}}(\cdot \mid S)$ *is also* L_H*-smooth, then* $G_{\mathrm{adapt}} \geq \mathrm{tr}(\Sigma)/(2L_H)$*.*

The implemented electorate depends on normalization. For N data rows, let r_j denote the rater associated with row j, and set $a_j = w_{r_j}/N_{r_j}$ and $A = \sum_j a_j$. The intended row mass is a_j/A. Under uniform row sampling, normalizing each random batch by its own weight sum need not recover this mass.

Theorem 5 (Exact self-normalized minibatch estimand) *Consider* b *independent uniform row draws, with row counts* C_j*, positive parameter-independent weights* a_j*, and gradients* $h_j = \nabla\ell_j(\theta)$*. Then*

$$\mathbb{E}\left[\frac{\sum_j C_j a_j h_j}{\sum_j C_j a_j}\right] = \sum_j \rho_j^{(b)} h_j, \qquad \rho_j^{(b)} := \mathbb{E}\left[\frac{C_j a_j}{\sum_l C_l a_l}\right].$$

Thus, the expected gradient equals $\nabla\mathcal{L}_b(\theta)$*, where* $\mathcal{L}_b(\theta) = \sum_j \rho_j^{(b)}\ell_j(\theta)$ *and* $\rho^{(b)}$ *is a probability vector. At* $b = 1$*,* $\rho_j^{(1)} = 1/N$*, regardless of the weights; as* $b \to \infty$*,* $\rho_j^{(b)} \to a_j/A$*.*

Theorem 5 is the Horvitz–Thompson/Hájek distinction: the known-normalizer estimator is unbiased because A is known, whereas the self-normalized estimator carries ratio bias of order $1/b$ that collapses to the unweighted objective at $b = 1$. Two implementations from survey and importance-sampling practice avoid the issue: the known normalizer, or drawing rows with probability proportional to a_j and averaging without weights (probability-proportional-to-size sampling, the two-stage reading of equation 3). Because the Hard–Soft gap is not interpretable without this information, we report the microbatch size and normalization scope used.

Separately normalizing singleton microbatches and then accumulating their gradients remains unweighted, even with a large effective batch. By contrast, the known-normalizer estimator

$$\widehat{g} = \frac{N}{bA}\sum_{t=1}^{b} a_{J_t} h_{J_t}$$

is unbiased for the gradient of the intended weighted objective. Interpreting an observed Hard–Soft gap therefore requires checking microbatch sizes, gradient accumulation, the Hard baseline's inverse-volume weights, and zero-weight handling.

Theorem 6 *[Binary DPO and external reversal] For a common prompt and two responses A, B, suppose rater i prefers A with probability p_i, and let $\bar{p} = \sum_i q_i p_i \in (0,1)$. Under population DPO with $\beta > 0$, an unrestricted positive binary policy, and reference probability $p_{\mathrm{ref}}(A) = r \in (0,1)$, the Soft-optimal policy satisfies*

$$p^*_{\mathrm{soft}}(A) = \sigma\big(\mathrm{logit}(r) + \beta^{-1}\,\mathrm{logit}(\bar{p})\big)\,.$$

There exist exact LEXIMIN Hard/Soft pairs exhibiting either ordering in expected external preference satisfaction under greedy decoding.

Remark. Each result above has a counterpart in the survey literature, which also supplies standard remedies. Theorem 1 is design-unbiasedness of the Hard objective for the Soft objective; its gradient and Hessian forms state that expected estimating equations equal the weighted population estimating equations. Soft is therefore the Rao–Blackwell-type marginalization of Hard (Rao, 1945; Blackwell, 1947): equal mean and zero panel variance, hence weakly dominant in mean squared error for the expected-panel target, and for that target only, as Theorem 2,Theorem 6, and the Community Alignment reversal make explicit. The variance $k^{-2}L^\top \Delta L$ of Theorem 3 is the Horvitz–Thompson variance of a fixed-size design in Sen–Yates–Grundy form (Sen, 1953; Yates and Grundy, 1953), with $\Delta\mathbf{1} = 0$ the fixed-size identity $\sum_j \pi_{ij} = k\pi_i$ that licenses that form. The observation that LEXIMIN fixes π but not the lottery is the classical fact that first-order inclusion probabilities do not determine a design: designs sharing π differ in π_{ij} and hence in variance, and choosing among them is the problem that motivates balanced sampling. LEXIMIN selects its lottery on neither the variance criterion of Theorem 3 nor a maximum-entropy criterion, so Hard variability is an unoptimized by-product of the fairness solution. Theorem 4 is a nonasymptotic, bound-form analogue of the design-based linearization variance of survey-weighted M-estimators (Binder, 1983), with Σ the design covariance of the score at the population optimum and μ in place of the Hessian.

For example, take two raters with $p_1 = 0.2$, $p_2 = 0.7$, panel size $k = 1$, a uniform panel lottery, and a uniform reference. The Soft-optimal policy greedily selects B, whereas the Hard-optimal policy selects A or B with equal probability over panels. Against a target with preference probability $p_T = 0.8$, expected satisfaction is 0.5 for Hard and 0.2 for Soft. For $p_T = 0.2$, these scores become 0.5 and 0.8, respectively. The reversal arises from a target-population shift, with Hard performance averaged over the full panel lottery.

Note that these results do not imply universal Hard superiority, a monotone benefit from smaller panels, or a model-size scaling law. They distinguish demographic participation, the objective actually optimized, the fitted policy, and external evaluation. The observed PRISM and Community Alignment orderings are compatible with these results; identifying their empirical mechanism requires replicated panels and training seeds, verified minibatch normalization, and uncertainty estimates that account for prompt and clause dependence.

5 EXPERIMENTS

Data and target distribution. We use PRISM (Kirk et al., 2024), which contains multi-turn conversations, model responses, human ratings/preferences, and rater demographics. We reconstruct coherent multi-turn contexts and form ordered preference pairs whenever two candidate responses for the same rater/context have different scores. The resulting Full PRISM training set has 57,750 pairs from 1,394 raters. Our U.S. target quotas cover ethnicity, religion, marital status, education, employment status, gender, and age. Targets come from ACS/Census estimates for demographic attributes (U.S. Census Bureau, 2020) and Pew estimates for religion (Smith, 2021). We use $k = 145$, the largest feasible panel size under 10% relative tolerance, treating ambiguous categories as slack.

For each conversation turn, we construct the context from prior user messages and the chosen or highest-scoring assistant response at each previous turn, avoiding inconsistent branching histories. We interpret PRISM scores only within the same rater/context turn and use all ordered pairs with nonzero score difference. This high-coverage construction may include noisy small-margin com-

Table 1: Training data after PRISM preprocessing. Hard Panel and US-Rep are smaller, so all conditions are trained for matched optimizer updates by resampling with replacement.

Condition	# pairs	# raters	Training signal
Full PRISM	57,750	1,394	unweighted full pool
Soft Panel	57,750	1,394	weighted by π_i
Hard Panel	5,737	145	sampled mini-public
US-Rep	9,541	230	PRISM U.S.-representative subset

parisons, but it preserves the conversational structure and keeps the comparison rule identical across training conditions.

Models and training. We fine-tune base Llama checkpoints with DPO: Llama-3.1-8B for the main experiment and Llama-3.2-1B/3B for scaling (Grattafiori et al., 2024). We use base rather than instruction-tuned checkpoints to isolate the effect of preference-data construction. All conditions use full-parameter bf16 fine-tuning, a frozen reference model initialized from the same base checkpoint, $\beta = 0.1$, learning rate 5×10^{-6}, and the same optimization budget: 3,538 optimizer updates, corresponding to two Full PRISM epochs. Smaller datasets are resampled with replacement to match this budget.

Training conditions. We compare Base, Full PRISM, PRISM's U.S.-representative subset (US-Rep), Hard Panel, and Soft Panel. US-Rep is a pre-constructed stratified subset in PRISM; Hard/Soft use our LEXIMIN sortition lottery and the same inclusion probabilities. For robustness, we train alternative Soft transforms, smaller panels, normalized-step Hard/US-Rep variants, and Community Alignment variants on U.S. annotators.

Evaluation. We evaluate against the 75-clause constitution of Huang et al. (2024), elicited through structured public input from approximately 1,000 representative U.S. adults. For each clause we generate 40 standalone questions, yielding 3,000 prompts. Each model answers every question with greedy decoding under a shared chat template. GPT-5.2 ranks all model responses listwise for each prompt; we collect five randomized-order rankings per prompt. We then compute Borda, Copeland, Kemeny-Young, Bradley–Terry, Plackett–Luce, and Mallows summaries with 1,000 bootstrap resamples where applicable. Inter-judge agreement is substantial: average Kendall $\tau = 0.776$ and Fleiss' $\kappa = 0.710$ for the top choice.

6 Results

Sortition panels improve over the full-pool baseline. Figure 2 shows the main 8B result. Hard Panel ranks first across aggregation families, US-Rep ranks second, and Soft Panel consistently improves over Full PRISM. Pairwise win rates give the same picture: Soft beats Full with probability 0.520 (95% CI [0.506, 0.534]), Hard beats Full with probability 0.618 ([0.602, 0.634]), and Hard beats US-Rep with probability 0.576 ([0.561, 0.592]). The US-Rep result shows that demographic filtering alone helps; Hard Panel's improvement over US-Rep suggests that the selection rule also matters.

We hypothesize two mechanisms behind the Hard/US-Rep gap. First, LEXIMIN's max-min property prevents individuals in scarce demographic intersections from receiving near-zero inclusion probability. Second, its symmetry property treats demographically equivalent raters equally, reducing arbitrary dependence on response volume or subset construction. These mechanisms are not separated in the current experiments, and testing alternative sortition rules is an important next step.

Scaling with model size. The ordering Hard $>$ US-Rep $>$ Soft $>$ Full $>$ Base appears at 1B, 3B, and 8B. Figure 3 summarizes Borda effect sizes: the Soft–Full gap grows from 0.038 (1B) to 0.045 (3B) to 0.076 (8B), and the Hard–US-Rep gap also increases with scale. We treat this as evidence that larger models more strongly absorb the distributional structure of their feedback data, not as a universal scaling law.

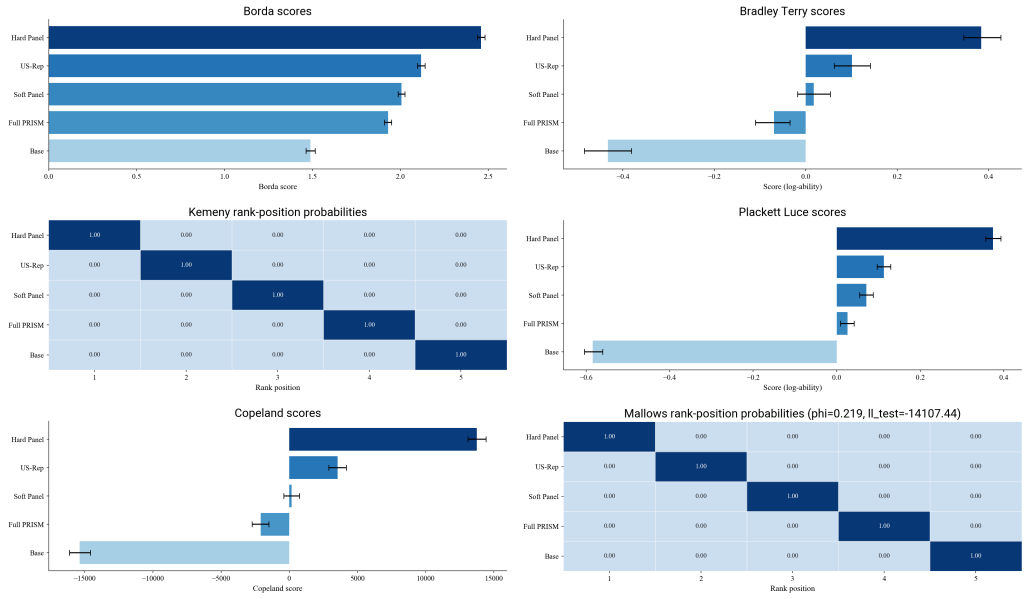

Figure 2: Llama-3.1-8B rankings under multiple aggregation methods. Across Bradley–Terry, Borda, Copeland, Kemeny, and Mallows summaries, Hard Panel ranks highest, US-Rep second, Soft Panel above Full PRISM, and Base last. Error bars are 95% bootstrap intervals with $n = 1000$.

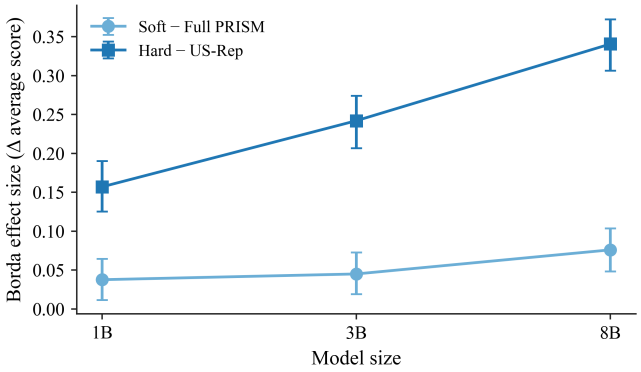

Figure 3: Panel effect sizes grow with model size. The Soft Panel–Full PRISM gap and Hard Panel–US-Rep gap are Borda average-score differences under bootstrap resampling. Error bars are 95% CIs.

Robustness and diagnostics. Table 2 summarizes additional experiments. Alternative Soft transforms show that all sortition-weighted variants improve over Full PRISM, while more concentrated transforms (π_i^2 and clipping) outperform linear weighting in this setting. Panel-size sensitivity shows different tradeoffs: smaller Soft panels can help by concentrating inclusion probability mass, whereas Hard Panel performs best at the largest feasible panel size because smaller hard panels discard more data. Gradient diagnostics show Hard and Soft do not implement the same practical update despite the expectation identity: Hard gradients are slightly more concentrated and have higher stable rank. Finally, on Community Alignment, Soft-US-$k350$ outperforms Full-US and Hard-US-$k350$, while Hard still outperforms the dataset's own balanced subset. This supports the representativeness effect but shows the Hard/Soft ordering depends on the data regime.

Cross-dataset replication on Community Alignment. Community Alignment provides a useful stress test because its preference data are collected in a different setting and its U.S. balanced subset is already curated. We filter to U.S. annotators and construct a $k = 350$ sortition target using age,

Table 2: Additional robustness checks. Scores are Bradley–Terry log-abilities within each comparison set.

Check	Result 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 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.

gender, and ethnicity marginals aligned with the dataset's balanced-subset notion. Soft-US-k350 slightly outperforms Full-US (BT 0.141 vs. 0.125), while Hard-US-k350 outperforms the dataset balanced subset but trails Full-US. This does not replicate the exact PRISM ordering; instead it supports a more nuanced claim. Representativeness correction helps, but hard filtering can lose too much data when the discarded comparisons are high quality or when the full subset is already close to the target.

Update magnitude and judge controls. Hard and US-Rep produce larger parameter deltas under the matched 3,538-step budget. We therefore train normalized-step variants (Hard 2,200 steps, US-Rep 2,100 steps) to match the Full/Soft update magnitude. The ranking persists, though gaps shrink. We also test judge ordering and verbosity biases: response presentation position has only small correlation with final rank ($\rho = 0.047$), and response length has small rank correlations ($\rho = -0.056$ by characters, -0.044 by words).

7 Limitations and Scope

Sortition-weighted preference learning requires demographic metadata, and datasets such as PRISM and Community Alignment are unusual in making such metadata available. The method should be understood as *reweighting the willing*: if a group is absent from the rater pool, no sortition procedure can recover that group's preferences. Demographic quotas are also imperfect proxies for values and do not resolve value conflicts or provide a full theory of minority rights. Our primary evaluation target is a U.S.-elicited constitution, so claims about other publics require their own target distributions and evaluation benchmarks. In the PRISM experiments, the U.S. demographic target is applied to a rater pool that includes multiple locales; this is a feasibility compromise that we report in the appendix. Finally, LLM-as-a-judge evaluation is a scalable proxy for open-ended constitutional alignment, but it should be complemented by human evaluation before high-stakes deployment.

8 Conclusion

We studied sortition-weighted preference learning as a way to make the composition of RLHF/DPO feedback explicit for a target public. On PRISM, Hard Panel and Soft Panel improve over unweighted Full PRISM under multiple aggregation methods, and the effect grows with model size. Ablations show that the strength of representativeness correction matters, while Community Alignment shows that Hard versus Soft is a data-regime tradeoff rather than a universal ordering. We also provided detailed theoretical analysis of our framework grounded in statistical sampling theory. The broader lesson is that pluralistic alignment is not only a matter of choosing an aggregation rule after data are collected: the composition of the feedback signal itself can be audited, constrained, and empirically consequential.

AI USE STATEMENT

We used generative AI tools to polish some sentences, refine word choices, and help debug minor code snippets. We did not use these tools for substantial coding tasks or other aspects of the paper. We take full responsibility for the final content, including the text and all claims made in the paper.

REPRODUCIBILITY STATEMENT

All code required to reproduce the results presented in this paper is provided in the supplementary materials.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences, 2022. URL https://arxiv.org/abs/2211.15006.

David A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3):279–292, 1983.

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18(1):105–110, 1947.

Terrill G Bouricius. Democracy through multi-body sortition: Athenian lessons for the modern day. *Journal of Deliberative Democracy*, 9(1), 2013.

Maarten Buyl, Hadi Khalaf, Claudio Mayrink Verdun, Lucas Monteiro Paes, Caio Cesar Vieira Machado, and Flavio du Pin Calmon. Ai alignment at your discretion. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 3046–3074, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: the cube method. *Biometrika*, 91 (4):893–912, 2004.

Michael R. Elliott and Richard Valliant. Inference for nonprobability samples. *Statistical Science*, 32(2):249–264, 2017.

Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 596(7873):548–552, 2021.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jaroslav Hájek. Comment on "an essay on the logical foundations of survey sampling, part one" by d. basu. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, page 236. Holt, Rinehart and Winston, Toronto, 1971.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional AI: Aligning a language model with public input, 2024. URL https://arxiv.org/abs/2406.07814.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 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.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Frank J. Potter. A study of procedures to identify and trim extreme sample weights. In *Proceedings of the Section on Survey Research Methods*, pages 225–230. American Statistical Association, 1990.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. 2000.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. URL https://arxiv.org/abs/2305.18290.

C. Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023. URL https://arxiv.org/abs/2303.17548.

A. R. Sen. On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5(2):119–127, 1953.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 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*, pages 292–314, 2025.

Gregory A. Smith. 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/.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

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

David Valle-Cruz, J Ignacio Criado, Rodrigo Sandoval-Almazán, and Edgar A Ruvalcaba-Gomez. Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation. *Government information quarterly*, 37(4):101509, 2020.

Frank Yates and P. M. Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*, 15(2):253–261, 1953.

Lily H. Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. 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.

Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.

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 Proofs of the theoretical results

Throughout this subsection, $\mathcal{I} = \{1, \dots, n\}$ is the finite rater pool, $\mathcal{S}$ is a nonempty family of feasible panels of fixed size $k \geq 1$, and π_{panel} is a probability distribution on $\mathcal{S}$. The inclusion probabilities are $\pi_i = \Pr_{S \sim \pi_{\text{panel}}}(i \in S)$, and $q_i = \pi_i / k$.

Unless stated otherwise, the dataset is fixed and randomness comes only from the sampled panel. In the results comparing Hard and Soft objectives through expectation, we use linear Soft weights $w_i = \pi_i$, so that

$$\mathcal{L}_{\text{hard}}(\theta \mid S) = \frac{1}{k} \sum_{i \in S} L_i(\theta), \qquad \mathcal{L}_{\text{soft}}(\theta) = \sum_{i=1}^{n} q_i L_i(\theta).$$

All objectives being compared have the same parameter domain.

Proof of Theorem 1:

For each rater i, define

$$v_{S,i} = \mathbf{1}\{i \in S\}.$$

Because every feasible panel contains exactly k raters,

$$\sum_{i=1}^{n} v_{S,i} = k$$

for every $S \in \mathcal{S}$. Taking expectations and using linearity of expectation gives

$$k = \mathbb{E}_S \left[\sum_{i=1}^{n} v_{S,i} \right] = \sum_{i=1}^{n} \mathbb{E}_S[v_{S,i}] = \sum_{i=1}^{n} \pi_i.$$

Consequently, $\sum_i q_i = 1$.

Fix a parameter θ. The Hard objective can be written as

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

Since the dataset and θ are fixed, each $L_i(\theta)$ is a constant with respect to panel sampling. Therefore,

$$\begin{aligned}\mathbb{E}_S[\mathcal{L}_{\text{hard}}(\theta \mid S)] &= \frac{1}{k}\sum_{i=1}^{n}\mathbb{E}_S[v_{S,i}]L_i(\theta)\\ &= \frac{1}{k}\sum_{i=1}^{n}\pi_i L_i(\theta)\\ &= \sum_{i=1}^{n} q_i L_i(\theta)\\ &= \mathcal{L}_{\text{soft}}(\theta).\end{aligned}$$

The panel lottery and inclusion probabilities do not depend on θ. If each L_i is continuously differentiable on the parameter domain, differentiating the finite sums gives

$$\nabla\mathcal{L}_{\text{hard}}(\theta \mid S) = \frac{1}{k}\sum_{i=1}^{n} v_{S,i}\nabla L_i(\theta),$$

and hence

$$\begin{aligned}\mathbb{E}_S[\nabla\mathcal{L}_{\text{hard}}(\theta \mid S)] &= \frac{1}{k}\sum_{i=1}^{n}\pi_i\nabla L_i(\theta)\\ &= \nabla\mathcal{L}_{\text{soft}}(\theta).\end{aligned}$$

If each L_i is twice continuously differentiable, the same calculation yields

$$\begin{aligned}\mathbb{E}_S[\nabla^2\mathcal{L}_{\text{hard}}(\theta \mid S)] &= \frac{1}{k}\sum_{i=1}^{n}\pi_i\nabla^2 L_i(\theta)\\ &= \nabla^2\mathcal{L}_{\text{soft}}(\theta).\end{aligned}$$

There is no interchange of differentiation with an infinite sum or integral: both the rater pool and the panel family are finite.

Let $g \subseteq \mathcal{I}$ denote a demographic group. By assumption,

$$L_g \le |S \cap g| \le U_g$$

for every feasible panel. Since $|S \cap g| = \sum_{i\in g} v_{S,i}$, taking expectations gives

$$L_g \le \sum_{i\in g}\mathbb{E}_S[v_{S,i}] = \sum_{i\in g}\pi_i \le U_g.$$

Dividing by $k > 0$ and using $q_i = \pi_i/k$ proves

$$\frac{L_g}{k} \le \sum_{i\in g} q_i \le \frac{U_g}{k}.$$

Proof of Theorem 2:

For every fixed θ and every panel S,

$$\inf_{\vartheta}\mathcal{L}_{\text{hard}}(\vartheta \mid S) \le \mathcal{L}_{\text{hard}}(\theta \mid S).$$

Here ϑ is the optimization variable, whereas θ is held fixed. Taking expectations gives

$$\mathbb{E}_S\left[\inf_{\vartheta}\mathcal{L}_{\text{hard}}(\vartheta \mid S)\right] \le \mathbb{E}_S[\mathcal{L}_{\text{hard}}(\theta \mid S)] = \mathcal{L}_{\text{soft}}(\theta),$$

where the equality follows from Theorem 1.

The left side does not depend on θ. Since the inequality holds for every θ, taking the infimum on the right gives

$$\mathbb{E}_S\left[\inf_{\vartheta} \mathcal{L}_{\text{hard}}(\vartheta \mid S)\right] \leq \inf_{\theta} \mathcal{L}_{\text{soft}}(\theta).$$

Suppose θ^*_{soft} minimizes $\mathcal{L}_{\text{soft}}$. For each panel, define

$$D_S = \mathcal{L}_{\text{hard}}(\theta^*_{\text{soft}} \mid S) - \inf_{\vartheta} \mathcal{L}_{\text{hard}}(\vartheta \mid S).$$

By the definition of an infimum, $D_S \geq 0$. Moreover,

$$\begin{aligned}\mathbb{E}_S[D_S] &= \mathbb{E}_S[\mathcal{L}_{\text{hard}}(\theta^*_{\text{soft}} \mid S)] - \mathbb{E}_S\left[\inf_{\vartheta} \mathcal{L}_{\text{hard}}(\vartheta \mid S)\right] \\ &= \mathcal{L}_{\text{soft}}(\theta^*_{\text{soft}}) - \mathbb{E}_S\left[\inf_{\vartheta} \mathcal{L}_{\text{hard}}(\vartheta \mid S)\right].\end{aligned}$$

Thus equality in equation 5 is equivalent to $\mathbb{E}_S[D_S] = 0$.

Because $\mathcal{S}$ is finite,

$$\mathbb{E}_S[D_S] = \sum_{S \in \mathcal{S}} \pi_{\text{panel}}(S) D_S$$

is a finite sum of nonnegative terms. It equals zero if and only if $D_S = 0$ for every panel with $\pi_{\text{panel}}(S) > 0$. For such a panel, $D_S = 0$ means

$$\mathcal{L}_{\text{hard}}(\theta^*_{\text{soft}} \mid S) = \inf_{\vartheta} \mathcal{L}_{\text{hard}}(\vartheta \mid S).$$

Therefore equality implies that θ^*_{soft} is a common minimizer of all panel objectives with positive probability.

Conversely, suppose a parameter θ_c minimizes every panel objective with positive probability. For any θ,

$$\mathcal{L}_{\text{hard}}(\theta_c \mid S) \leq \mathcal{L}_{\text{hard}}(\theta \mid S)$$

on the support of the lottery. Taking expectations shows

$$\mathcal{L}_{\text{soft}}(\theta_c) \leq \mathcal{L}_{\text{soft}}(\theta).$$

Thus θ_c also minimizes the Soft objective, and

$$\begin{aligned}\mathbb{E}_S\left[\inf_{\vartheta} \mathcal{L}_{\text{hard}}(\vartheta \mid S)\right] &= \mathbb{E}_S[\mathcal{L}_{\text{hard}}(\theta_c \mid S)] \\ &= \mathcal{L}_{\text{soft}}(\theta_c) \\ &= \inf_{\theta} \mathcal{L}_{\text{soft}}(\theta).\end{aligned}$$

This proves the equivalence.

Since θ^*_{soft} is a global Soft minimizer,

$$\mathcal{L}_{\text{soft}}(\theta) \geq \mathcal{L}_{\text{soft}}(\theta^*_{\text{soft}})$$

for every admissible parameter θ. In particular,

$$\mathcal{L}_{\text{soft}}(\widehat{\theta}_{\text{hard}}) \geq \mathcal{L}_{\text{soft}}(\theta^*_{\text{soft}}).$$

If $\widehat{\theta}_{\text{hard}}$ is random, this inequality holds for every realization.

These comparisons concern the training objectives. No assumption in the theorem relates an external risk R to those objectives, so the displayed inequalities imply no ordering of external risks. The construction in Theorem 6 provides explicit examples of both external orderings.

Proof of Theorem 3:

Fix θ, and abbreviate $L_i = L_i(\theta)$ and $g_i = \nabla L_i(\theta)$. Let $v_{S,i} = \mathbf{1}\{i \in S\}$. Because $\mathbb{E}_S[v_{S,i}] = \pi_i$ and $\mathbb{E}_S[v_{S,i}v_{S,j}] = \pi_{ij}$,

$$\begin{aligned}\mathbb{E}_S[(v_{S,i} - \pi_i)(v_{S,j} - \pi_j)] &= \mathbb{E}_S[v_{S,i}v_{S,j}] - \pi_j\mathbb{E}_S[v_{S,i}] - \pi_i\mathbb{E}_S[v_{S,j}] + \pi_i\pi_j \\ &= \pi_{ij} - \pi_i\pi_j \\ &= \Delta_{ij}.\end{aligned}$$

When $i = j$, the identity remains valid because $v_{S,i}^2 = v_{S,i}$ and $\pi_{ii} = \pi_i$. Thus

$$\Delta = \mathbb{E}_S[(v_S - \pi)(v_S - \pi)^\top] = \mathrm{Cov}_S(v_S),$$

where $\pi = (\pi_1, \ldots, \pi_n)^\top$.

By Theorem 1,

$$\mathcal{L}_{\mathrm{hard}}(\theta \mid S) - \mathbb{E}[\mathcal{L}_{\mathrm{hard}}(\theta \mid S)] = \mathcal{L}_{\mathrm{hard}}(\theta \mid S) - \mathcal{L}_{\mathrm{soft}}(\theta) = \frac{1}{k}\sum_i (v_{S,i} - \pi_i)L_i.$$

Expanding the square and taking expectations gives

$$\begin{aligned}\mathrm{Var}_S\big(\mathcal{L}_{\mathrm{hard}}(\theta \mid S)\big) &= \frac{1}{k^2}\mathbb{E}_S\left[\left(\sum_i (v_{S,i} - \pi_i)L_i\right)^2\right] \\ &= \frac{1}{k^2}\sum_{i,j} L_iL_j\,\mathbb{E}_S[(v_{S,i} - \pi_i)(v_{S,j} - \pi_j)] \\ &= \frac{1}{k^2}\sum_{i,j}\Delta_{ij}L_iL_j \\ &= \frac{1}{k^2}L^\top\Delta L.\end{aligned}$$

No independence between inclusion indicators is required.

The centered panel gradient is

$$\nabla\mathcal{L}_{\mathrm{hard}}(\theta \mid S) - \mathbb{E}_S[\nabla\mathcal{L}_{\mathrm{hard}}(\theta \mid S)] = \frac{1}{k}\sum_i (v_{S,i} - \pi_i)g_i.$$

Using the definition of vector covariance and expanding the outer product,

$$\begin{aligned}\mathrm{Cov}_S\big(\nabla\mathcal{L}_{\mathrm{hard}}(\theta \mid S)\big) &= \frac{1}{k^2}\mathbb{E}_S\left[\left(\sum_i (v_{S,i} - \pi_i)g_i\right)\left(\sum_j (v_{S,j} - \pi_j)g_j\right)^\top\right] \\ &= \frac{1}{k^2}\sum_{i,j}\mathbb{E}_S[(v_{S,i} - \pi_i)(v_{S,j} - \pi_j)]\,g_ig_j^\top \\ &= \frac{1}{k^2}\sum_{i,j}\Delta_{ij}g_ig_j^\top.\end{aligned}$$

Now we will prove that $\Delta \succeq 0$ and $\Delta\mathbf{1} = 0$.

For any $a \in \mathbb{R}^n$,

$$a^\top\Delta a = \mathbb{E}_S[(a^\top(v_S - \pi))^2] \geq 0,$$

so Δ is positive semidefinite.

Also, $\sum_j v_{S,j} = k$ and $\sum_j \pi_j = k$, implying $\sum_j (v_{S,j} - \pi_j) = 0$ for every panel. Therefore, for each i,

$$\begin{aligned}(\Delta \mathbf{1})_i &= \sum_j \Delta_{ij} \\ &= \mathbb{E}_S \left[(v_{S,i} - \pi_i) \sum_j (v_{S,j} - \pi_j) \right] \\ &= 0.\end{aligned}$$

Hence $\Delta \mathbf{1} = 0$.

Fix a realizable inclusion vector π^*. A lottery with probabilities λ_S preserves this vector exactly when

$$\lambda_S \geq 0, \qquad \sum_{S \in \mathcal{S}} \lambda_S = 1, \qquad \sum_{S \in \mathcal{S}} \lambda_S v_S = \pi^*.$$

For every such lottery,

$$\begin{aligned}\mathbb{E}_\lambda[g_S] &= \sum_{S \in \mathcal{S}} \lambda_S \left(\frac{1}{k} \sum_i v_{S,i} g_i \right) \\ &= \frac{1}{k} \sum_i \left(\sum_{S \in \mathcal{S}} \lambda_S v_{S,i} \right) g_i \\ &= \frac{1}{k} \sum_i \pi_i^* g_i = \bar{g}.\end{aligned}$$

Thus the mean gradient is fixed throughout the feasible lottery set.

Writing $\delta_S = g_S - \bar{g}$ and using linearity of trace,

$$\begin{aligned}\operatorname{tr}\big(W \operatorname{Cov}_\lambda(g_S)\big) &= \operatorname{tr}\left(W \mathbb{E}_\lambda[\delta_S \delta_S^\top]\right) \\ &= \mathbb{E}_\lambda[\operatorname{tr}(W \delta_S \delta_S^\top)] \\ &= \mathbb{E}_\lambda[\delta_S^\top W \delta_S] \\ &= \sum_{S \in \mathcal{S}} \lambda_S (g_S - \bar{g})^\top W (g_S - \bar{g}).\end{aligned}$$

The third equality uses the cyclic trace identity. Each coefficient $(g_S - \bar{g})^\top W (g_S - \bar{g})$ is fixed, so this objective is linear in λ. This proves its equivalence to the linear program in the theorem.

The feasible lottery set is nonempty because π^* is realizable. It is a closed subset of the finite dimensional probability simplex, since its additional constraints are linear equalities. It is therefore compact. The objective is linear and hence continuous, so it attains a minimum on this nonempty compact set.

Proof of Theorem 4:

For each panel, let

$$\theta_S = \theta^*_{\mathrm{hard}}(S), \qquad \theta_* = \theta^*_{\mathrm{soft}}, \qquad g^*_S = \nabla \mathcal{L}_{\mathrm{hard}}(\theta_* \mid S).$$

The abbreviations θ_S and θ_* are used only within this proof. All norms are Euclidean.

A convex combination of functions that are strongly convex with parameter μ is also strongly convex with parameter μ. Thus the Soft objective is strongly convex. In particular, its minimizer is unique. The panel minimizers are likewise unique. Since the domain is $\mathbb{R}^d$ and the objectives are differentiable,

$$\nabla \mathcal{L}_{\mathrm{hard}}(\theta_S \mid S) = 0, \qquad \nabla \mathcal{L}_{\mathrm{soft}}(\theta_*) = 0.$$

For any differentiable function f that is strongly convex with parameter μ,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|^2.$$

Interchanging x and y and adding the two inequalities gives

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \mu \|x - y\|^2.$$

Apply this inequality to $f = \mathcal{L}_{\mathrm{hard}}(\cdot \mid S)$, $x = \theta_*$, and $y = \theta_S$:

$$(g_S^*)^\top (\theta_* - \theta_S) \geq \mu \|\theta_* - \theta_S\|^2.$$

By the Cauchy–Schwarz inequality,

$$(g_S^*)^\top (\theta_* - \theta_S) \leq \|g_S^*\| \, \|\theta_* - \theta_S\|.$$

If $\theta_S = \theta_*$, the desired bound is immediate. Otherwise, division by $\mu \|\theta_* - \theta_S\|$ gives

$$\|\theta_S - \theta_*\| \leq \frac{\|g_S^*\|}{\mu}.$$

Consequently,

$$\mathbb{E}_S \|\theta_S - \theta_*\|^2 \leq \frac{1}{\mu^2} \mathbb{E}_S \|g_S^*\|^2.$$

The gradient identity in Theorem 1 gives

$$\mathbb{E}_S[g_S^*] = \nabla \mathcal{L}_{\mathrm{soft}}(\theta_*) = 0.$$

Therefore,

$$\Sigma = \mathbb{E}_S[g_S^* (g_S^*)^\top],$$

and

$$\mathrm{tr}(\Sigma) = \mathbb{E}_S[\mathrm{tr}(g_S^* (g_S^*)^\top)] = \mathbb{E}_S \|g_S^*\|^2.$$

Substituting into Step 1 yields

$$\mathbb{E}_S \|\theta_{\mathrm{hard}}^*(S) - \theta_{\mathrm{soft}}^*\|^2 \leq \frac{\mathrm{tr}(\Sigma)}{\mu^2}.$$

We first recall the smoothness inequality. If ∇f is Lipschitz with constant L, then, for $d = y - x$,

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\top d &= \int_0^1 (\nabla f(x + td) - \nabla f(x))^\top d \, dt \\ &\leq \int_0^1 Lt \|d\|^2 \, dt \\ &= \frac{L}{2} \|d\|^2. \end{aligned}$$

Apply this to $f = \mathcal{L}_{\mathrm{soft}}$, $x = \theta_*$, and $y = \theta_S$. Since $\nabla \mathcal{L}_{\mathrm{soft}}(\theta_*) = 0$,

$$\mathcal{L}_{\mathrm{soft}}(\theta_S) - \mathcal{L}_{\mathrm{soft}}(\theta_*) \leq \frac{L}{2} \|\theta_S - \theta_*\|^2.$$

The left side is nonnegative because θ_* is a minimizer. Taking expectations and using Step 2 gives

$$0 \leq \mathbb{E}_S[\mathcal{L}_{\mathrm{soft}}(\theta_S)] - \mathcal{L}_{\mathrm{soft}}(\theta_*) \leq \frac{L \, \mathrm{tr}(\Sigma)}{2\mu^2}.$$

The expectation identity gives

$$\mathcal{L}_{\mathrm{soft}}(\theta_*) = \mathbb{E}_S[\mathcal{L}_{\mathrm{hard}}(\theta_* \mid S)].$$

Since θ_S minimizes its panel objective,

$$G_{\mathrm{adapt}} = \mathbb{E}_S \left[\mathcal{L}_{\mathrm{hard}}(\theta_* \mid S) - \mathcal{L}_{\mathrm{hard}}(\theta_S \mid S) \right].$$

Every quantity inside the expectation is nonnegative.

Strong convexity, applied at θ_*, implies that for every θ,

$$\mathcal{L}_{\mathrm{hard}}(\theta \mid S) \geq \mathcal{L}_{\mathrm{hard}}(\theta_* \mid S) + (g_S^*)^\top(\theta - \theta_*) + \frac{\mu}{2}\|\theta - \theta_*\|^2.$$

For $d = \theta - \theta_*$, completing the square gives

$$(g_S^*)^\top d + \frac{\mu}{2}\|d\|^2 = \frac{\mu}{2}\left\|d + \frac{g_S^*}{\mu}\right\|^2 - \frac{\|g_S^*\|^2}{2\mu} \geq -\frac{\|g_S^*\|^2}{2\mu}.$$

Thus

$$\mathcal{L}_{\mathrm{hard}}(\theta_S \mid S) \geq \mathcal{L}_{\mathrm{hard}}(\theta_* \mid S) - \frac{\|g_S^*\|^2}{2\mu}.$$

Rearranging and averaging yields

$$0 \leq G_{\mathrm{adapt}} \leq \frac{\mathbb{E}_S\|g_S^*\|^2}{2\mu} = \frac{\mathrm{tr}(\Sigma)}{2\mu}.$$

Suppose each panel gradient is Lipschitz with constant L_H. Consider the parameter

$$\widetilde{\theta}_S = \theta_* - \frac{1}{L_H} g_S^*.$$

The smoothness inequality from Step 3 gives

$$\begin{aligned}\mathcal{L}_{\mathrm{hard}}(\widetilde{\theta}_S \mid S) &\leq \mathcal{L}_{\mathrm{hard}}(\theta_* \mid S) - \frac{1}{L_H}\|g_S^*\|^2 + \frac{L_H}{2}\left\|\frac{g_S^*}{L_H}\right\|^2 \\ &= \mathcal{L}_{\mathrm{hard}}(\theta_* \mid S) - \frac{\|g_S^*\|^2}{2L_H}.\end{aligned}$$

Since θ_S is a global panel minimizer,

$$\mathcal{L}_{\mathrm{hard}}(\theta_S \mid S) \leq \mathcal{L}_{\mathrm{hard}}(\widetilde{\theta}_S \mid S).$$

Combining these inequalities gives

$$\mathcal{L}_{\mathrm{hard}}(\theta_* \mid S) - \mathcal{L}_{\mathrm{hard}}(\theta_S \mid S) \geq \frac{\|g_S^*\|^2}{2L_H}.$$

Taking expectations proves

$$G_{\mathrm{adapt}} \geq \frac{\mathrm{tr}(\Sigma)}{2L_H}.$$

Proof of Theorem 5:

Fix θ and write $h_j = \nabla \ell_j(\theta)$. Let $J_1, \ldots, J_b$ be independent uniform draws from $\{1, \ldots, N\}$, and define the row counts

$$C_j = \sum_{t=1}^{b} \mathbf{1}\{J_t = j\}.$$

Thus $\sum_j C_j = b$. The weights a_j are strictly positive and do not depend on θ.

The gradient computed from a batch is

$$\widehat{h}_b = \frac{\sum_j C_j a_j h_j}{\sum_j C_j a_j} = \sum_j \left(\frac{C_j a_j}{\sum_l C_l a_l}\right) h_j.$$

Because $b \geq 1$ and every $a_j > 0$, the denominator is strictly positive for every batch. Taking expectations and using that the h_j are fixed gives

$$\mathbb{E}[\widehat{h}_b] = \sum_j \mathbb{E}\left[\frac{C_j a_j}{\sum_l C_l a_l}\right] h_j = \sum_j \rho_j^{(b)} h_j.$$

For every realized batch,

$$0 \le \frac{C_j a_j}{\sum_l C_l a_l} \le 1, \qquad \sum_j \frac{C_j a_j}{\sum_l C_l a_l} = 1.$$

Taking expectations gives

$$\rho_j^{(b)} \ge 0, \qquad \sum_j \rho_j^{(b)} = 1.$$

The distribution of the counts and the weights are independent of θ, so $\rho_j^{(b)}$ is also independent of θ. Therefore,

$$\begin{aligned}\nabla \mathcal{L}_b(\theta) &= \nabla \left(\sum_j \rho_j^{(b)} \ell_j(\theta) \right) \\ &= \sum_j \rho_j^{(b)} \nabla \ell_j(\theta) \\ &= \mathbb{E}[\widehat{h}_b].\end{aligned}$$

When $b = 1$, precisely one row J_1 is drawn. Its weight cancels from the numerator and denominator:

$$\widehat{h}_1 = \frac{a_{J_1} h_{J_1}}{a_{J_1}} = h_{J_1}.$$

Equivalently,

$$\frac{C_j a_j}{\sum_l C_l a_l} = \mathbf{1}\{J_1 = j\}.$$

Hence

$$\rho_j^{(1)} = \Pr(J_1 = j) = \frac{1}{N}.$$

Thus the expected singleton gradient is the gradient of the unweighted row objective, regardless of the positive weights.

Couple the batches using an infinite sequence of independent uniform draws $J_1, J_2, \ldots$. For each row j, the strong law of large numbers gives

$$\frac{C_j}{b} \longrightarrow \frac{1}{N} \qquad \text{almost surely.}$$

Since the number of rows is finite,

$$\frac{1}{b} \sum_l C_l a_l = \sum_l \frac{C_l}{b} a_l \longrightarrow \frac{1}{N} \sum_l a_l = \frac{A}{N} > 0 \qquad \text{almost surely.}$$

Consequently,

$$\frac{C_j a_j}{\sum_l C_l a_l} = \frac{(C_j/b) a_j}{\sum_l (C_l/b) a_l} \longrightarrow \frac{a_j/N}{A/N} = \frac{a_j}{A} \qquad \text{almost surely.}$$

The ratio lies in $[0, 1]$ for every batch. Bounded convergence therefore implies

$$\rho_j^{(b)} = \mathbb{E}\left[\frac{C_j a_j}{\sum_l C_l a_l} \right] \longrightarrow \frac{a_j}{A}.$$

We also justify the $O(1/b)$ bias statement in the discussion. Let J be a uniform row draw and define

$$Y = a_J, \qquad X = a_J h_J, \qquad \mu = \mathbb{E}[Y] = \frac{A}{N}, \qquad h^\circ = \frac{\mathbb{E}[X]}{\mu} = \frac{\sum_j a_j h_j}{A}.$$

Set $U = X - h^\circ Y$, so that $\mathbb{E}[U] = 0$. For independent copies (X_t, Y_t, U_t), write their sample means as $\overline{X}$, $\overline{Y}$, and $\overline{U}$. Then

$$\widehat{h}_b - h^\circ = \frac{\overline{X} - h^\circ \overline{Y}}{\overline{Y}} = \frac{\overline{U}}{\overline{Y}}.$$

Since $\mathbb{E}[\overline{U}] = 0$,

$$\begin{aligned}\mathbb{E}[\widehat{h}_b] - h^\circ &= \mathbb{E}\left[\overline{U}\left(\frac{1}{\overline{Y}} - \frac{1}{\mu}\right)\right] \\ &= -\mathbb{E}\left[\frac{\overline{U}(\overline{Y} - \mu)}{\mu\overline{Y}}\right].\end{aligned}$$

Let $a_{\min} = \min_j a_j > 0$. Because $\overline{Y} \geq a_{\min}$, the Cauchy–Schwarz inequality gives

$$\left\|\mathbb{E}[\widehat{h}_b] - h^\circ\right\| \leq \frac{\sqrt{\mathbb{E}\|\overline{U}\|^2\,\mathbb{E}(\overline{Y} - \mu)^2}}{\mu a_{\min}}.$$

Independence and $\mathbb{E}[U] = 0$ imply

$$\mathbb{E}\|\overline{U}\|^2 = \frac{\mathbb{E}\|U\|^2}{b}, \qquad \mathbb{E}(\overline{Y} - \mu)^2 = \frac{\mathrm{Var}(Y)}{b}.$$

Therefore,

$$\left\|\mathbb{E}[\widehat{h}_b] - h^\circ\right\| \leq \frac{\sqrt{\mathbb{E}\|U\|^2\,\mathrm{Var}(Y)}}{b\mu a_{\min}}.$$

For the fixed finite dataset and parameter, this is an $O(1/b)$ bound. The bias can be zero in special cases; the bound does not assert that it is always nonzero.

For the known normalizer estimator,

$$\widehat{g} = \frac{N}{bA}\sum_{t=1}^{b} a_{J_t} h_{J_t},$$

linearity of expectation gives

$$\begin{aligned}\mathbb{E}[\widehat{g}] &= \frac{N}{bA}\sum_{t=1}^{b}\left(\frac{1}{N}\sum_j a_j h_j\right) \\ &= \frac{1}{A}\sum_j a_j h_j = h^\circ.\end{aligned}$$

Alternatively, draw each row with probability a_j/A and average its gradient without additional weights. The expected gradient of each draw is then $\sum_j (a_j/A)h_j = h^\circ$.

When $a_j = w_{r_j}/N_{r_j}$ and every included rater has $N_i > 0$ prepared comparisons,

$$A = \sum_{i=1}^{n}\sum_{j:r_j=i}\frac{w_i}{N_i} = \sum_{i=1}^{n} w_i = Z,$$

and

$$\frac{1}{A}\sum_j a_j \ell_j(\theta) = \frac{1}{Z}\sum_i w_i L_i(\theta) = \mathcal{L}_{\mathrm{soft}}(\theta).$$

Thus these alternatives target the intended Soft objective.

Finally, if m singleton microbatches are normalized separately and their gradients are averaged at the same parameter, the accumulated gradient is

$$\frac{1}{m}\sum_{t=1}^{m} h_{J_t}.$$

Under uniform row sampling, its expectation is $N^{-1}\sum_j h_j$. Accumulation therefore does not restore weights that have already canceled within singleton microbatches.

Proof of Theorem 6:

For the common prompt, let $u = p_\theta(A) \in (0,1)$. Then $p_\theta(B) = 1 - u$. The reference probabilities are $p_{\mathrm{ref}}(A) = r$ and $p_{\mathrm{ref}}(B) = 1 - r$, where $r \in (0,1)$.

Define the DPO margin for preferring A to B by

$$\begin{aligned} z(u) &= \beta \left(\log \frac{u}{r} - \log \frac{1-u}{1-r} \right) \\ &= \beta\big(\mathrm{logit}(u) - \mathrm{logit}(r)\big), \end{aligned}$$

where $\mathrm{logit}(u) = \log(u/(1-u))$. Write $\sigma(z) = (1 + e^{-z})^{-1}$. If A is preferred, the DPO loss is $-\log \sigma(z)$; if B is preferred, it is $-\log \sigma(-z)$.

Rater i prefers A with probability p_i, so their population loss is

$$L_i(u) = -p_i \log \sigma(z(u)) - (1 - p_i) \log \sigma(-z(u)).$$

Averaging with weights q_i and using $\sum_i q_i = 1$ gives

$$\mathcal{L}_{\mathrm{soft}}(u) = -\bar{p} \log \sigma(z(u)) - (1 - \bar{p}) \log \sigma(-z(u)), \qquad \bar{p} = \sum_i q_i p_i.$$

Thus the Soft objective depends on the rater preferences through the weighted probability $\bar{p}$.

Since $\beta > 0$, the map $u \mapsto z(u)$ is a bijection from $(0,1)$ to $\mathbb{R}$. We may therefore minimize over z instead of u.

Define

$$\phi(z) = -\bar{p} \log \sigma(z) - (1 - \bar{p}) \log \sigma(-z).$$

Using

$$-\log \sigma(z) = \log(1 + e^z) - z, \qquad -\log \sigma(-z) = \log(1 + e^z),$$

we obtain

$$\phi(z) = \log(1 + e^z) - \bar{p} z.$$

Its derivatives are

$$\phi'(z) = \sigma(z) - \bar{p}, \qquad \phi''(z) = \sigma(z)(1 - \sigma(z)) > 0.$$

The function is strictly convex. Since $\bar{p} \in (0,1)$, its unique stationary point satisfies

$$\sigma(z^*) = \bar{p}, \qquad z^* = \mathrm{logit}(\bar{p}).$$

This stationary point is the unique global minimizer.

Substituting into the definition of $z(u)$ gives

$$\beta \left(\mathrm{logit}(u^*) - \mathrm{logit}(r) \right) = \mathrm{logit}(\bar{p}).$$

Solving for u^* yields

$$p^*_{\mathrm{soft}}(A) = u^* = \sigma\left(\mathrm{logit}(r) + \beta^{-1} \mathrm{logit}(\bar{p}) \right).$$

Take two raters, panel size $k = 1$, and feasible panels $\{1\}$ and $\{2\}$. If the first panel is selected with probability λ, the inclusion vector is $(\lambda, 1 - \lambda)$. Its smallest coordinate is at most $1/2$, with equality only when $\lambda = 1/2$. Thus the exact LEXIMIN lottery selects each singleton with probability $1/2$, giving

$$\pi_1 = \pi_2 = \frac{1}{2}, \qquad q_1 = q_2 = \frac{1}{2}.$$

Set

$$p_1 = 0.2, \qquad p_2 = 0.7, \qquad r = \frac{1}{2}.$$

Then $\bar{p} = 0.45$ and $\mathrm{logit}(r) = 0$. For any $\beta > 0$,

$$p^*_{\mathrm{soft}}(A) = \sigma\big(\beta^{-1} \mathrm{logit}(0.45)\big) < \frac{1}{2}.$$

Therefore the Soft policy selects B under greedy decoding.

For a singleton Hard panel $\{i\}$, the same optimization calculation applies with preference probability p_i. Since $p_1 < 1/2$ and $p_2 > 1/2$, the optimal Hard policy selects B when $S = \{1\}$ and A when $S = \{2\}$. Thus Hard selects each response with probability $1/2$ over the panel lottery.

Let an external target population prefer A with probability p_T. The satisfaction of choosing A is p_T, whereas that of choosing B is $1 - p_T$.

The expected satisfaction of Hard is consequently

$$\frac{1}{2}p_T + \frac{1}{2}(1 - p_T) = \frac{1}{2}.$$

Soft always selects B, so its satisfaction is $1 - p_T$.

For $p_T = 0.8$, Hard has expected satisfaction 0.5, while Soft has satisfaction 0.2. For $p_T = 0.2$, Hard still has expected satisfaction 0.5, while Soft has satisfaction 0.8. This establishes both possible orderings under an exact LEXIMIN lottery.

Equivalently, defining external risk as one minus preference satisfaction gives expected Hard risk 0.5 and Soft risk p_T, again producing either ordering.

A.3 Training Dataset Sizes

Table 3 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 3: 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 the exact selection probabilities $\{\pi_i\}$ from the LEXIMIN-induced lottery and verify the 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 2. 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 4: Soft weighting-function ablation. Scores are Bradley–Terry log-abilities fit to listwise-judge-derived pairwise preferences.

Panel-size sensitivity. We vary the panel size for the 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 5: 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 6: 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 7: Community Alignment replication on the U.S. subset. Scores are Bradley–Terry logabilities 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 fig. 2).

A.10 Raw Scoring Outputs

For reproducibility and to make the plots in fig. 2 numerically explicit, Tables 9, 10, and 11 report the raw scores computed from our evaluation artifacts. We additionally report the corresponding raw scores for the normalized-step 8B ablation in fig. 5 and Tables 12–14. For completeness we also report the corresponding raw-score tables for Llama-3.2-1B (Tables 15–17) and Llama-3.2-3B (Tables 18–20). 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).

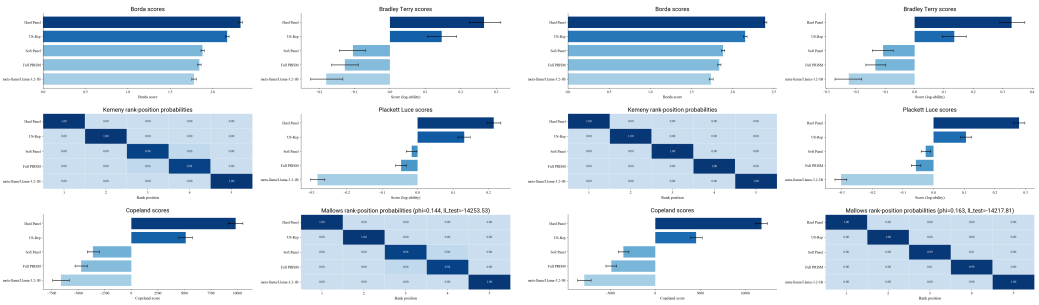

Figure 4: Score grids for Llama-3.2-1B (left) and Llama-3.2-3B (right) under the same evaluation protocol.

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 fig. 5), 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 8: 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 $n = 1000$.

A.13 Model-Size Scaling

We include the 1B and 3B score grids in Appendix Figure 4. The same ordering (Hard $>$ US-Rep $>$ Soft $>$ Full $>$ Base) appears at all sizes, while effect sizes increase with model capacity (see Figure 3 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 9: 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 10: 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 11: 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.

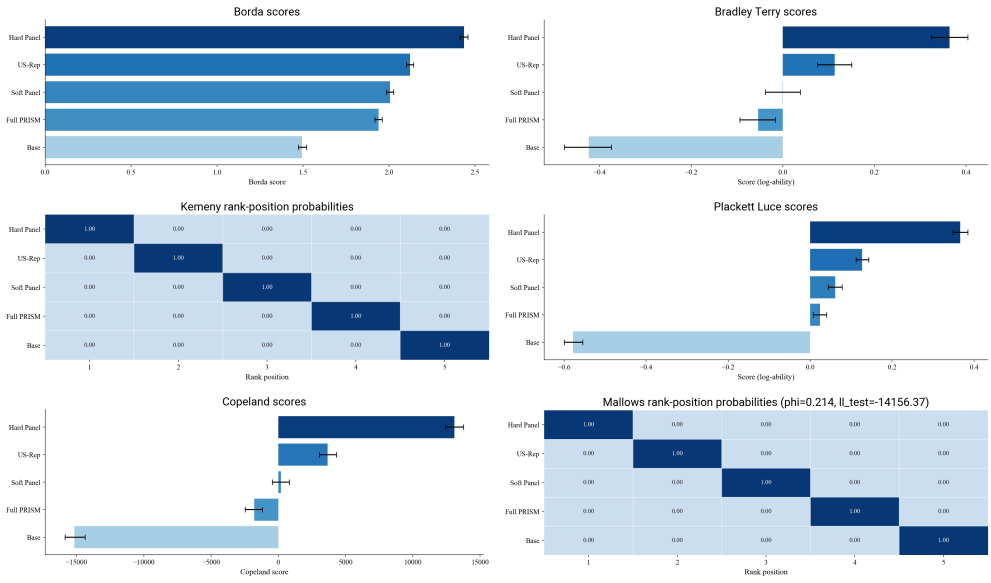

Figure 5: Normalized-step score grid for Llama-3.1-8B, where Hard Panel and US-Rep are retrained for fewer steps (Hard: 2200, US-Rep: 2100) to match the update magnitude of Full/Soft.

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 12: 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 13: 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 14: 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 15: 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 16: 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 17: 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 18: 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 19: 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 20: Llama-3.2-3B Kemeny and Mallows summaries (bootstrap n=1000 where reported).

Llama-3.2-3B raw scores.