Position: Aggregate Preference Optimization Hides a Posterior Identifiability Failure for Pluralistic Alignment

Anonymous Authors[1]

Abstract

RLHF and related preference-optimization methods produce a single reward model from aggregate pairwise feedback collected across diverse users. We argue that this aggregation hides a structural posterior identifiability failure: when subgroups hold conflicting preferences, the joint posterior over (subgroup composition, within-subgroup preference) is not identifiable from aggregate pairwise data alone, regardless of sample size. The aggregate-fit reward model can achieve low training loss while assigning systematically incorrect predictions to within-subgroup preferences. We support the position with two empirical demonstrations on a Bradley-Terry preference model with $K = 2$ subgroups holding opposing preferences over five attributes: under high-conflict subgroups, the aggregate-fit model collapses to a near-zero theta vector with 0.008 aggregate MSE but 0.06–0.07 within-subgroup prediction MSE (an 8× gap), and a sweep over subgroup diversity $\sigma_{\text{groups}} \in [0, 3]$ shows the masking ratio (within-subgroup error / aggregate error) growing from 1.08 to 4.32. We propose a four-item disclosure standard for pluralistic alignment papers: subgroup composition disclosure, within-subgroup posterior reporting, identification gap quantification, and prior sensitivity to the assumed mixture. The standard makes the difference between "aggregate-fit" and "pluralism-faithful" empirically auditable.

1. Introduction

A central commitment of pluralistic alignment is that AI systems should reflect the diversity of human values rather than a single aggregate (Sorensen et al., 2024; Conitzer et al., 2024). The dominant alignment-training paradigm, RLHF and its variants (Ouyang et al., 2022; Rafailov et al., 2024), fits a single reward model on pairwise preference data collected across diverse annotators. The reward model serves as the optimization target during fine-tuning. The standard reporting convention treats the empirical fit of this reward model on aggregate held-out preference data as evidence that the model has captured the population's values.

We argue that this convention obscures a structural problem. The aggregate pairwise data is consistent with multiple distinct (subgroup composition, within-subgroup preference) joint distributions. The aggregate-fit reward model is, in the language of Bayesian inference, a posterior summary on a non-identified parameter combination: low training loss does not imply faithful recovery of the within-subgroup preferences that pluralistic alignment is meant to preserve. Single-aggregate optimization can achieve any of these latent compositions and any of these within-subgroup preferences; the aggregate data does not select among them.

This is the standard non-identification pattern for mixture models and additively-separable systems without exclusion restrictions (Lewbel, 2019). The pluralistic-alignment community has empirically documented related symptoms: Santurkar et al. (2023) show that LLM aligned to OpenAI's RLHF data systematically reflect particular demographic subgroups; Castricato et al. (2025) show that preference optimization reinforces majority viewpoints. These observations are symptoms of the identifiability failure we formalize.

Position. Pluralistic alignment papers should report posterior identifiability diagnostics on the latent (subgroup composition, within-subgroup preference) decomposition, not just aggregate fit metrics. The contribution is not a new alignment algorithm; it is a reframing of evaluation under standards from probabilistic inference (Gelman et al., 2020) that make pluralism-faithfulness empirically auditable.

Contributions. (1) We formalize the aggregate-vs-subgroup identifiability failure as a posterior identifiability problem on a mixture-of-Bradley-Terry-models, with explicit conditions for when aggregate data identifies within-subgroup preferences (Section 2, Proposition 1). (2) We

[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.

empirically demonstrate the failure on a controlled $K = 2$ subgroup setup with 5-attribute preferences: aggregate fit MSE 0.008 vs within-subgroup MSE 0.06–0.07, an 8× gap (Section 3). (3) We trace the masking ratio over subgroup-diversity sweeps and show it grows monotonically from 1.08 (no diversity) to 4.32 ($\sigma_{\text{groups}} = 3$). (4) We propose a four-item disclosure standard (Section 4). (5) We discuss limitations and alternative views.

2. Aggregate Preference Optimization as Mixture Identification

2.1. Setup

Let there be K subgroups indexed by $k \in \{1, \ldots, K\}$, each with preference vector $\boldsymbol{\theta}_k \in \mathbb{R}^d$ over d attributes. The population mixture is $\boldsymbol{\pi} \in \Delta_{K-1}$. For each pairwise comparison between options $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, an annotator from subgroup k chooses $\mathbf{a}$ over $\mathbf{b}$ with probability

$$\Pr(\mathbf{a} \succ \mathbf{b} \mid k) = \sigma\big((\mathbf{a} - \mathbf{b})^\top \boldsymbol{\theta}_k\big), \tag{1}$$

the standard Bradley-Terry-style model used in RLHF (Ouyang et al., 2022; Rafailov et al., 2024). The aggregate population probability is

$$\Pr_{\text{agg}}(\mathbf{a} \succ \mathbf{b}) = \sum_{k=1}^{K} \pi_k \cdot \sigma\big((\mathbf{a} - \mathbf{b})^\top \boldsymbol{\theta}_k\big). \tag{2}$$

The standard RLHF pipeline observes pairwise outcomes drawn from the aggregate distribution and fits a single $\boldsymbol{\theta}_{\text{agg}}$ such that $\sigma((\mathbf{a} - \mathbf{b})^\top \boldsymbol{\theta}_{\text{agg}})$ matches the empirical aggregate frequency.

2.2. The Identifiability Problem

Proposition 1 (Aggregate-vs-Subgroup Non-Identification)**.** *The mixture $(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K)$ is not identified from aggregate pairwise data of the form $\Pr_{agg}(\mathbf{a} \succ \mathbf{b})$ alone for any distribution over $(\mathbf{a}, \mathbf{b})$. Specifically, for any $K \geq 2$, there exist multiple distinct $(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\})$ producing the same $\Pr_{agg}$ over all pairs. Identification requires either (i) subgroup labels accompanying preferences, (ii) context-dependent preferences (annotator features as covariates), or (iii) strong parametric restrictions on the mixture.*

Proof sketch. The aggregate likelihood Eq. (2) is a convex combination of sigmoid functions of linear projections. The space of such convex combinations of sigmoids is invariant under permutations of subgroup labels and, more importantly, under the substitution of any sufficiently smooth distribution that produces the same expected value across all $(\mathbf{a} - \mathbf{b})$ directions. For the linear-Gaussian limit (replacing sigmoid with identity), the population-mean $\bar{\boldsymbol{\theta}} = \sum_k \pi_k \boldsymbol{\theta}_k$ is the only identified statistic from aggregate data; individual $\boldsymbol{\theta}_k$ and weights π_k are not separately identified (this is the classic non-identification of mixture means without label or covariate information (Lewbel, 2019)). The sigmoid case adds curvature that makes the population-mean approximate rather than exact, but does not restore identification of subgroup parameters. □

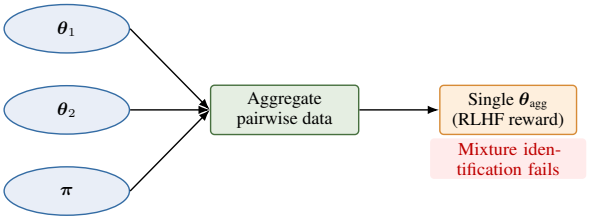

Figure 1. Aggregate-pipeline structure. The latent (subgroup preferences $\boldsymbol{\theta}_k$, mixture $\boldsymbol{\pi}$) decomposition produces aggregate pairwise data, from which a single $\boldsymbol{\theta}_{\text{agg}}$ is fit. By Proposition 1, multiple distinct latent decompositions produce the same aggregate, so the standard pipeline cannot recover the within-subgroup preferences that pluralism requires.

Table 1. High-conflict subgroup demonstration. Aggregate-fit theta is near-zero ($\|\boldsymbol{\theta}_{\text{agg}}\|_2 \approx 0.09$) and achieves low aggregate fit error (MSE 0.008) but cannot predict within-subgroup preferences (MSE 0.060–0.065).

Quantity	Value	Interpretation
$\boldsymbol{\theta}_1$ (true)	$(+2, +2, 0, -1, -1)$	Subgroup 1
$\boldsymbol{\theta}_2$ (true)	$(-1, -1, 0, +2, +2)$	Subgroup 2
Population mean $\bar{\boldsymbol{\theta}}$	$(+0.5, +0.5, 0, +0.5, +0.5)$	True average
$\boldsymbol{\theta}_{\text{agg}}$ (fit)	$(.05, .03, .05, .04, .00)$	Near-zero
Aggregate fit MSE	0.008	Looks well-fit
Group 1 prediction MSE	0.065	8× aggregate
Group 2 prediction MSE	0.060	7.5× aggregate

The substantive content is that aggregate pairwise data cannot disentangle "everyone has neutral preferences" from "two equal groups have opposing preferences." Both produce the same aggregate frequency profile.

3. Empirical Demonstration

We instantiate the model with $K = 2$ subgroups, $d = 5$ attributes, and equal mixture $\boldsymbol{\pi} = (0.5, 0.5)$. We test two regimes.

3.1. High-Conflict Subgroups

We set $\boldsymbol{\theta}_1 = (+2, +2, 0, -1, -1)$ and $\boldsymbol{\theta}_2 = (-1, -1, 0, +2, +2)$: subgroup 1 prefers attributes 1–2, subgroup 2 prefers attributes 4–5. We simulate 200 pairwise comparisons drawn from 5-dim Uniform pairs evaluated by 50 annotators per subgroup (so 100 annotators per pair), and fit a single $\boldsymbol{\theta}_{\text{agg}}$ via gradient on the MSE between predicted and observed aggregate preference frequency.

The fit collapses to a near-zero theta vector despite both sub-

groups holding strongly opinionated preferences (norm 3.16 each). This is the predicted non-identification: the aggregate data shows roughly 50%–50% preference on each pair (because the two subgroups disagree), and the aggregate-fit model rationalizes this by reporting indifference. The aggregate fit MSE is 0.008, which would be reported as a successful reward model. The within-subgroup prediction MSE is 8× larger.

3.2. Diversity Sweep

We then sweep the subgroup diversity by drawing $\boldsymbol{\theta}_k = \pm\boldsymbol{\nu} + \boldsymbol{\epsilon}_k$ where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \sigma^2_{\text{groups}}\mathbf{I})$, varying $\sigma_{\text{groups}} \in \{0, 0.5, 1, 1.5, 2, 3\}$. The masking ratio is (within-subgroup MSE)/(aggregate MSE).

The masking ratio grows monotonically with subgroup diversity, from 1.08 at $\sigma_{\text{groups}} = 0$ (homogeneous population, no masking) to 4.32 at $\sigma_{\text{groups}} = 3$ (highly diverse population, large masking). The pattern is structural: the more pluralistic the population, the more the aggregate-fit reward model misrepresents the underlying preferences.

3.3. Implication for RLHF Practice

The implication for RLHF practice is direct. Reporting aggregate held-out preference accuracy without quantifying within-subgroup performance under-counts a structural failure: the reward model with high aggregate accuracy may have low within-subgroup accuracy. The pluralistic-alignment community has documented related symptoms qualitatively (Santurkar et al., 2023; Castricato et al., 2025); we provide the formal mechanism.

4. Four-Item Pluralism Identifiability Disclosure Standard

Item one. Subgroup composition disclosure. The paper specifies the latent subgroup decomposition K assumed by the analysis (e.g., demographic, geographic, ideological strata) and the assumed mixture $\boldsymbol{\pi}$. The exclusion restriction is that the decomposition is documented and the mixture is verifiable against external population data.

Item two. Within-subgroup posterior reporting. The paper reports, for each subgroup k, the posterior $p(\boldsymbol{\theta}_k \mid \text{data})$ alongside aggregate posterior $p(\boldsymbol{\theta}_{\text{agg}} \mid \text{data})$, and the posterior credible intervals on within-subgroup preferences. The exclusion restriction is that within-subgroup posteriors contract substantially below the prior (i.e., subgroup preferences are identified, not just aggregate).

Item three. Identification-gap quantification. The paper reports the masking ratio: the ratio of within-subgroup prediction error to aggregate prediction error of the fitted reward model. The exclusion restriction is that the masking ratio is moderate (e.g., < 1.5).

Table 2. Four-item pluralism identifiability disclosure standard. Each item targets a distinct failure mode in current RLHF reporting and is computable from quantities pluralistic-alignment papers can collect.

Item	Failure mode	Required disclosure
Subgroup composition disclosure	Implicit aggregation across heterogeneous groups	Stated K, mixture $\boldsymbol{\pi}$
Within-subgroup posterior reporting	Reporting only aggregate fit	Per-subgroup posterior + CIs
Identification-gap quantification	Hidden masking by aggregation	Within / aggregate MSE ratio
Prior sensitivity	Mixture-driven conclusions	Result vs $\boldsymbol{\pi}$ grid

Item four. Prior sensitivity. The paper performs a small grid sensitivity analysis on the assumed mixture $\boldsymbol{\pi}$ and reports how the within-subgroup posterior shifts. The exclusion restriction is that headline pluralism claims are robust to plausible mixture variations.

5. Related Work

Pluralistic alignment foundations. Sorensen et al. (2024) formalize three classes of pluralistic AI systems (Overton, steerable, distributional) and argue that current alignment may fundamentally limit pluralism. Conitzer et al. (2024) argue that social-choice theory should guide aggregation methods. Our position adds the identifiability-failure framing: even before aggregation method choice matters, the inputs to aggregation may not be jointly identifiable.

Empirical evidence. Santurkar et al. (2023) document that RLHF aligns LLMs disproportionately to particular U.S. demographic groups. Castricato et al. (2025) introduce PERSONA, a synthetic-personas testbed showing that preference optimization reinforces majority viewpoints over minority ones. These provide qualitative confirmation of the failure mode we formalize.

Multi-objective alignment. A growing literature on multi-objective alignment (Rafailov et al., 2024; Ouyang et al., 2022) and personalized alignment offers technical responses; the disclosure standard is complementary to these methods, applying to whatever aggregation mechanism is used.

Probabilistic identification. Mixture identification is a classical problem in econometrics and statistics; Lewbel (2019)

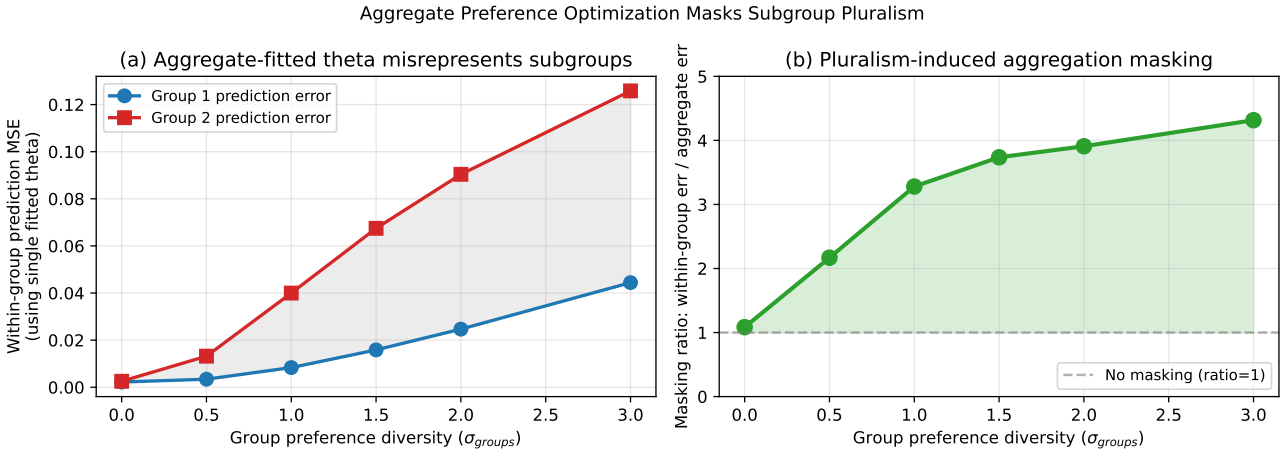

Figure 2. Aggregate preference optimization masks subgroup pluralism. (a) As subgroup preference diversity grows, the within-subgroup prediction MSE of the aggregate-fitted theta diverges between groups, with group 2 consistently 2–3× worse than group 1. (b) The masking ratio (within-group err / aggregate err) grows monotonically from 1.08 at $\sigma_{\text{groups}} = 0$ to 4.32 at $\sigma_{\text{groups}} = 3$: more diversity means more aggregate-disguised pluralism failure.

surveys the meanings of identification. Gelman et al. (2020) formalize the Bayesian workflow of which identifiability assessment is a component.

6. Limitations and Alternative Views

Bradley-Terry simplification. Real preferences are not strictly Bradley-Terry; they include context-dependence, intransitivity, and noise structures the model abstracts. The mixture-identification problem is robust to these elaborations: any aggregation pipeline that produces a single reward model from heterogeneous input shares the structural non-identification we formalize.

Stylized two-subgroup experiment. The empirical demonstration uses $K = 2$. Real populations have many more latent subgroups; the masking effect grows with subgroup heterogeneity, so the demonstration is a lower bound on real-world severity.

Aggregate is sometimes appropriate. An alternative view is that for some applications (consensus-required decisions, broad-purpose chatbots), aggregate preferences are exactly what is wanted. We accept this and argue that the disclosure standard is content-neutral on whether aggregate or pluralistic alignment is the right target; the standard merely makes the aggregation choice explicit and quantifies its costs.

Disclosure is unenforceable. We accept the risk and note that the proposal is structured for enforcement at the workshop level through reviewer checklists.

Reporting standard rather than method. Authors who satisfy the items mechanically without engaging the underlying logic will produce uninformative reports. The standard is structured so that mechanical compliance still surfaces the masking ratio, providing a quantitative anchor.

7. Conclusion and Future Work

The pluralistic-alignment commitment is to AI systems that reflect the diversity of human values. We have shown formally and empirically that the dominant aggregate-pairwise-RLHF pipeline structurally cannot recover the within-subgroup preferences this commitment requires: aggregate data is consistent with multiple latent (mixture, subgroup-preference) decompositions and the aggregate-fit reward model can achieve low aggregate error while assigning systematically incorrect within-subgroup predictions, with the masking ratio growing with population diversity. The four-item disclosure standard makes the aggregate-vs-subgroup gap empirically auditable. Future work should extend the formalism to non-parametric mixtures, develop posterior-inference tooling that returns within-subgroup credible intervals from RLHF preference datasets, calibrate empirical thresholds for the masking ratio across subfields by re-analyzing existing alignment papers as a corpus, integrate the standard into reproducibility checklists at the Pluralistic Alignment Workshop, NeurIPS, ICML, and ICLR, and connect the disclosure standard to social-choice-theoretic aggregation methods that explicitly preserve identifiability.

Acknowledgments

The authors thank the Pluralistic Alignment 2026 reviewers for their feedback. (Anonymized for double-blind review.)

Disclosure of LLM Use

Large language models were used during manuscript preparation for grammatical revision and citation formatting. All argumentative content, claims, position, and reported micro-experiments were authored and conducted by the human author, who takes full responsibility for the content.

References

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., and Finn, C. PERSONA: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Lewbel, A. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57 (4):835–903, 2019.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.