GROUPWISE DISTORTION GUARANTEES FOR PREFERENCE-BASED ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference-based alignment methods such as RLHF and NLHF aggregate ordinal pairwise comparisons to learn an LLM policy, but a natural goal is maximizing *social welfare* (average cardinal utility), which comparisons do not identify. Gölz, Haghtalab, and Yang (GHY) measure the gap by *distortion*: the worst-case ratio between the welfare of the best fixed lottery (distribution over responses) and of the learned lottery. They show NLHF is optimal when every user receives the same lottery. Account-based LLMs, however, have information about their users and can serve different lotteries to different people. We give an efficient algorithm, GLHF, that learns a single group-conditioned policy from one comparison per user. This policy attains GHY's optimal distortion bound simultaneously on every group in a prespecified, possibly overlapping collection, with sample complexity growing logarithmically in the number of groups and inversely with the smallest group mass. When the members of a group share a feasible favorite response, that group's distortion approaches one. On human coffee ratings and synthetic LLM-generated ratings, GLHF lowers distortion in every group relative to NLHF and substantially reduces worst-group distortion relative to group-agnostic baselines.

1 INTRODUCTION

Preference-based alignment methods such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and Nash learning from human feedback (NLHF) (Munos et al., 2024) train LLM policies from pairwise comparisons between candidate outputs. A natural goal is to maximize *social welfare*—the average cardinal utility they obtain from their outputs (Gölz et al., 2025). But when users differ in their preferences over outputs (Santurkar et al., 2023; Kirk et al., 2024), this problem is not well identified because the comparisons observed are only ordinal.

Gölz, Haghtalab, and Yang (GHY) (Gölz et al., 2025) study this gap through *distortion*: the worst-case ratio of the welfare of the best fixed *lottery* (a distribution over responses), chosen with full knowledge of utilities, to the welfare of a lottery learned only from sampled comparisons. Under individual Bradley–Terry preferences and data supplying one comparison per sampled user, they prove a worst-case lower bound on distortion and show that NLHF matches it asymptotically, even when constrained to lie in a KL-ball around a reference lottery. Their learned lottery is shared by all users, so the guarantee controls population welfare but not welfare within particular subgroups.

Account-based LLMs, however, observe user context such as custom instructions, settings, and persistent memory (OpenAI, 2023; 2024), so a *policy* can serve different lotteries to different users. We ask whether GHY's optimal distortion guarantee can be obtained simultaneously on prespecified, overlapping groups of users, at sample and computational cost growing only modestly with the number of groups. For disjoint groups this is easy: separately solve an NLHF problem on each group. With overlapping groups, a user belongs to several groups but can receive only one lottery, and so this approach fails. Learning a separate lottery for each group intersection pattern recovers disjointness but would require exponential sample and computational complexity.

1.1 OUR MODEL AND CONTRIBUTION

A fixed prompt has m candidate completions. Each user's observable context reveals membership in K possibly overlapping groups, and we learn a group-conditioned policy from T independently

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.

sampled users, each giving one pairwise comparison among completions. The queried completion pair is either uniformly random (*passive* sampling) or consists of one completion drawn from the learner's current policy and one drawn uniformly from the rest (*active* sampling). Following GHY, every served lottery and every comparator must lie in a KL ball of radius τ around a reference lottery π_{ref}; a uniform π_{ref} with $\tau = \log m$ recovers the unconstrained social-choice setting. Each group's benchmark is its own best feasible fixed lottery.

Groupwise maximal-lottery certificates. We introduce Groupwise Learning from Human Feedback (GLHF), which learns a single feasible group-conditioned policy π_{GLHF} such that, simultaneously for every group, no feasible lottery is substantially preferred to it on average over the group. This is an approximate groupwise analogue of the maximal-lottery property (Fishburn, 1984) underlying NLHF. Let $\mathrm{Expl}_k(\pi)$ be the average preference margin (probability of winning minus probability of losing) of the best challenger lottery over π on group k. Under active sampling, with probability at least $1-\delta$, simultaneously for every group k,

$$\mathrm{Expl}_k(\pi_{\mathrm{GLHF}}) \le \varepsilon_{T,k}, \qquad \varepsilon_{T,k} = O\left(\sqrt{\frac{(m-1)\big(\tau + \log(K/\delta)\big)}{T p_k}}\right) \quad \text{as } T p_k \to \infty,$$

where p_k is the population mass of group k. Passive sampling replaces $m-1$ by $\binom{m}{2}$.

Groupwise welfare. Under individual Bradley–Terry preferences with inverse temperature β, these certificates imply that, up to a welfare error vanishing with $\varepsilon_{T,k}$, GLHF attains the distortion bound L/ℓ_β (defined in Section 3)—shown to be optimal in the single lottery setting by Gölz et al. (2025)—simultaneously for every group. Theorem 12 shows this factor is worst-case optimal in our setting as well.

Preference-aligned groups. The guarantee strengthens when membership is informative about preferences. For each group, GLHF's welfare falls below the group's benchmark by at most a finite-sample error plus a constant multiple of the difference between the welfare of providing each group member their most preferred alternative and the welfare of the group's best fixed feasible lottery (Section 3.1). If almost every member shares a feasible favorite completion, this difference is zero and the group's distortion converges to one.

Experiments. Using human ratings of four coffees, with training limited to one simulated comparison per respondent and held-out ratings used to measure welfare directly, multigroup GLHF achieves mean worst-group distortion 1.08 versus 1.28 for NLHF. Appendix G.3 reports a parallel exploratory study using synthetic, LLM-generated ratings of six answers to one question.

1.2 Related work

The closest work is Gölz et al. (2025), as discussed above. We adopt their distortion benchmark and give an algorithm that achieves the same optimal distortion guarantee not only on average over the population, but simultaneously for arbitrarily overlapping groups.

Two other lines of work have close connections. Multigroup learning competes with each group's best predictor, even when groups overlap (Blum & Lykouris, 2020; Tosh & Hsu, 2022; Globus-Harris et al., 2022; Acharya et al., 2024). GLHF uses related techniques for a different objective: satisfying maximal-lottery constraints groupwise, which yields the distortion guarantees.

Several other works take different approaches to learning group conditioned policies. Group Preference Optimization learns a preference predictor that adapts to a target group from a few examples, including previously unseen groups (Zhao et al., 2024). Group Robust Preference Optimization instead trains a group-conditioned policy by minimizing the largest group-specific preference-training loss, an explicit minimax objective (Ramesh et al., 2024). Our work gives a stronger benchmark guarantee: simultaneously for every prespecified, possibly overlapping group, our policy competes with that group's own best feasible lottery and consequently obtains a cardinal-welfare and distortion guarantee. We give a more extensive discussion of related work in Appendix H.

2 PRELIMINARIES

Fix a prompt and $m \geq 2$ candidate responses, called *alternatives*, indexed by $\mathcal{A} := \{1, \ldots, m\}$. A user is drawn from a fixed population distribution over an observable context X in a measurable space $\mathcal{X}$ and an unobserved utility vector $U \in [0,1]^m$. Context and utility may be arbitrarily correlated. All population expectations below refer to this joint distribution.

Lotteries, policies, and groups. For clarity, we distinguish a *policy* from a *lottery*. Both specify distributions over $\mathcal{A}$, but a policy may depend on context, whereas a lottery is fixed. Formally, a lottery is a vector $q \in \Delta(\mathcal{A}) := \{q \in \mathbb{R}^m : q \geq 0,\ \sum_a q(a) = 1\}$, while a policy is a map $\pi : \mathcal{X} \to \Delta(\mathcal{A})$. Write $\pi(a \mid x)$ for the probability the policy serves a given a context x.

Before training, the learner is given $K \geq 2$ possibly overlapping groups. Group k has an indicator $G_k : \mathcal{X} \to \{0,1\}$ and known population mass $p_k := \Pr(G_k(X) = 1) > 0$. The groups cover the context space: $\sum_{k=1}^{K} G_k(x) \geq 1$ for every x. This can always be ensured by including the universal group $G \equiv 1$, which also gives a population-wide guarantee. The learned policy is *group-conditioned*: it depends on x only through $(G_1(x), \ldots, G_K(x))$. Group membership is available during both training and deployment.

A common feasible set. Before training, fix a full-support reference lottery π_{ref}, meaning $\pi_{\mathrm{ref}}(a) > 0$ for every a, and a radius $\tau \geq 0$. For a full-support lottery v, the Kullback–Leibler (KL) divergence is $\mathrm{KL}(q\|v) := \sum_a q(a) \log(q(a)/v(a))$, with zero summands when $q(a) = 0$. All logarithms are natural. The common feasible set is

$$\mathcal{B}_\tau(\pi_{\mathrm{ref}}) := \{q \in \Delta(\mathcal{A}) : \mathrm{KL}(q\|\pi_{\mathrm{ref}}) \leq \tau\}. \tag{1}$$

We abbreviate the set by $\mathcal{B}_\tau$. A policy is feasible if $\pi(x) \in \mathcal{B}_\tau$ for every x. Following GHY's terminology, the *alignment setting* imposes this KL constraint, while the *social choice setting* allows every lottery. The latter is recovered once $\tau \geq \log(1/\min_a \pi_{\mathrm{ref}}(a))$, so we cap τ at this value without changing the feasible set. At $\tau = 0$, only π_{ref} is feasible and no learning is needed.

Utilities and welfare. For integrable Z, write $\mathbb{E}_k[Z] := \mathbb{E}[Z \mid G_k(X) = 1] = \mathbb{E}[G_k(X)Z]/p_k$. Group welfare and its best fixed-lottery benchmark are

$$W_k(\pi) := \mathbb{E}_k[\pi(X)^\top U], \qquad W_k^\star := \max_{q \in \mathcal{B}_\tau} W_k(q) \tag{2}$$

Here $W_k(q) = \mathbb{E}_k[q^\top U]$ treats q as a constant policy. The benchmark may differ between groups, but the learned mapping is the same for all of them.

Following GHY's welfare-ratio notion (Gölz et al., 2025), for $W_k(\pi) > 0$ define the *constrained group distortion*

$$\mathrm{Dist}_k(\pi) := \frac{W_k^\star}{W_k(\pi)}. \tag{3}$$

Pairwise feedback and margins. A comparison presents a user with two distinct alternatives $\{a, b\}$ and records which one they choose. Write $a \succ b$ for the event that the user chooses a when this pair is queried; exactly one of $a \succ b$ and $b \succ a$ occurs. Conditional on the user's (X, U) and the queried pair, the choice follows a fixed, possibly stochastic comparison law. The comparison guarantee allows this law to depend arbitrarily on (X, U); the welfare guarantee below will impose a Bradley–Terry law. In probabilities involving $a \succ b$, the query $\{a, b\}$ is implicit. Define the signed preference margin at context x by

$$M_{ab}(x) := \Pr(a \succ b \mid X = x) - \Pr(b \succ a \mid X = x) = 2\Pr(a \succ b \mid X = x) - 1, \tag{4}$$

and set $M_{aa}(x) = 0$. The matrix $M(x)$ is antisymmetric with entries in $[-1, 1]$. For independent draws from lotteries q, p, independent of the user's latent utility given x, the *advantage* $q^\top M(x) p$ is the expected win-minus-loss margin of q against p. The algorithm observes comparisons, not $M(x)$.

Definition 1 (Signed groupwise exploitability)**.** For a policy π, its exploitability on group k is

$$\mathrm{Expl}_k(\pi) := \sup_{q \in \mathcal{B}_\tau} \mathbb{E}_k[q^\top M(X)\pi(X)]. \tag{5}$$

This is the largest group-average advantage of a fixed feasible lottery over π. We say π has an ε_k comparison certificate for group k if $\mathrm{Expl}_k(\pi) \leq \varepsilon_k$. A group-conditioned policy can outperform every fixed feasible lottery on average, so its exploitability can be negative.

Training data and sampling protocols. Training uses $T \geq 2$ independent users (X_t, U_t) from the deployment population. Before round t, the learner fixes a current policy map $x \mapsto \widehat{\pi}_t(x)$ using previous observations. It then observes X_t, queries one pair of distinct alternatives, and observes the winner, never U_t. Pair selection is independent of U_t conditional on the history and X_t; the winner is then drawn from the comparison law above, independently of the history given (X_t, U_t) and the pair. We study two ways to choose the pair:

- *Passive uniform sampling* draws an unordered pair uniformly, independently of the current policy.
- *Policy-adaptive active sampling* draws $A \sim \widehat{\pi}_t(X_t)$, then draws B uniformly from $\mathcal{A} \setminus \{A\}$, and queries the unordered pair $\{A, B\}$.

Let $\rho_t(a, b)$ denote the probability of querying $\{a, b\}$ conditional on the history and X_t. Thus, for $a \neq b$,

$$\rho_t^{\mathrm{pas}}(a, b) = \binom{m}{2}^{-1}, \qquad \rho_t^{\mathrm{act}}(a, b) = \frac{\widehat{\pi}_t(a \mid X_t) + \widehat{\pi}_t(b \mid X_t)}{m - 1}. \tag{6}$$

The active protocol concentrates queries on pairs with a response the current policy is likely to serve.

A utility model for welfare. The margins above are defined for any pairwise response law. To relate them to welfare, we assume the following link between comparison probabilities and utilities.

Assumption 2 (Individual Bradley–Terry comparisons)**.** There is a common inverse temperature $\beta > 0$ such that, conditional on X, U, and a queried pair $\{a, b\}$, the probability of $a \succ b$ is $\sigma(\beta(U(a) - U(b)))$, where $\sigma(z) = (1 + e^{-z})^{-1}$ (Bradley & Terry, 1952).

Consequently, $M_{ab}(x) = \mathbb{E}[2\sigma(\beta(U(a) - U(b))) - 1 \mid X = x]$. The latent utilities may vary arbitrarily across users.

3 Simultaneous welfare guarantees

GLHF learns one feasible contextual policy such that, within every named group, no fixed feasible lottery has much average pairwise advantage over it. In each training round, GLHF combines the lotteries maintained for the sampled user's groups and updates them from that user's single comparison; Section 4 gives the procedure. We first bound this advantage from sampled comparisons, then relate it to each group's welfare relative to its own best fixed feasible lottery.

Theorem 3 (Simultaneous comparison guarantees)**.** *Under either sampling protocol in Section 2, define its score normalization factor*

$$R := \begin{cases} m - 1, & \textit{policy-adaptive active sampling}, \\ \binom{m}{2}, & \textit{passive uniform sampling} \end{cases}. \tag{7}$$

For $\delta \in (0, 1)$, define

$$H := \tau + \log \frac{2K^2}{\delta}, \qquad \varepsilon_{T,k} := 6\sqrt{\frac{RH}{Tp_k}} + 6\frac{RH}{Tp_k} \quad (\tau > 0). \tag{8}$$

For $\tau > 0$, GLHF returns a feasible group-conditioned policy π_{GLHF} that, with probability at least $1 - \delta$, satisfies simultaneously for every $k \in [K]$,

$$\mathrm{Expl}_k(\pi_{\mathrm{GLHF}}) \leq \varepsilon_{T,k}. \tag{9}$$

Why the certificate bounds welfare. Set $L := \sigma'(0) = 1/4$ and $\ell_\beta := (\sigma(\beta) - 1/2)/\beta$. The Bradley–Terry link satisfies the following pointwise inequality (Gölz et al., 2025, Lemma 1):

$$2\sigma(\beta(u - v)) - 1 \leq 2\beta\big(Lu - \ell_\beta v\big), \qquad u, v \in [0, 1]. \tag{10}$$

The left side is a comparison margin; the right side is linear in the two utilities. Averaging within a group therefore relates its comparison margin to two group welfares. For each group, we apply this inequality with the group's best fixed feasible lottery chosen as the comparator.

Proposition 4 (Simultaneous welfare guarantees)**.** *Under Assumption 2, on the* $1-\delta$ *success event of Theorem 3, the GLHF policy satisfies simultaneously for every* $k \in [K]$*,*

$$W_k(\pi_{\mathrm{GLHF}}) \geq \frac{\ell_\beta}{L} W_k^\star - \frac{\varepsilon_{T,k}}{2\beta L}. \tag{11}$$

Proof. Fix a group k and a welfare-maximizing feasible lottery $q_k^\star \in \arg\max_{q \in \mathcal{B}_\tau} W_k(q)$. Draw a fresh user (X, U) from group k, then draw $A \sim \pi_{\mathrm{GLHF}}(X)$ and $B \sim q_k^\star$ independently of each other and of U given X. All expectations below include these response draws. Theorem 3 and antisymmetry give $-\varepsilon_{T,k} \leq \mathbb{E}_k[\pi_{\mathrm{GLHF}}(X)^\top M(X) q_k^\star]$. Thus

$$\begin{aligned} -\varepsilon_{T,k} &\leq \mathbb{E}_k[\pi_{\mathrm{GLHF}}(X)^\top M(X) q_k^\star] = \mathbb{E}_k[M_{AB}(X)] \\ &= \mathbb{E}_k[2\sigma(\beta(U(A) - U(B))) - 1] \\ &\leq 2\beta\, \mathbb{E}_k[LU(A) - \ell_\beta U(B)] \\ &= 2\beta\big(L W_k(\pi_{\mathrm{GLHF}}) - \ell_\beta W_k^\star\big). \end{aligned} \tag{12}$$

The second line uses the definition of M, the third applies (10), and the last uses linearity of expectation: $\mathbb{E}_k[U(A)] = W_k(\pi_{\mathrm{GLHF}})$ and $\mathbb{E}_k[U(B)] = W_k^\star$. Rearranging gives (11). □

For every group with $W_k^\star > 0$, rearranging Proposition 4 gives a distortion upper bound that approaches L/ℓ_β as the certificate error vanishes. This factor is worst-case optimal over the full instance class with one comparison per sampled user: Theorem 12 adapts the lower-bound construction of Gölz et al. (2025, Theorem 3) to contextual policies.

3.1 Stronger guarantees for preference-aligned groups

The worst-case factor in Proposition 4 permits arbitrary preference heterogeneity within every group. A sharper groupwise bound is available when a group's best fixed feasible lottery nearly attains its members' individual-best utilities. For each group k, fix a welfare-maximizing feasible lottery $q_k^\star \in \arg\max_{q \in \mathcal{B}_\tau} W_k(q)$. Different groups may have different maximizers. We quantify the residual welfare loss within group k by $\mathrm{Res}_k := \mathbb{E}_k[\max_a U(a)] - W_k^\star \geq 0$. This residual welfare loss decomposes as

$$\mathrm{Res}_k = \left(\mathbb{E}_k[\max_a U(a)] - \max_{q \in \Delta(\mathcal{A})} W_k(q)\right) + \left(\max_{q \in \Delta(\mathcal{A})} W_k(q) - W_k^\star\right). \tag{13}$$

The first term is the cost of requiring one fixed lottery to serve all members of the group; the second is the additional cost of constraining that lottery to $\mathcal{B}_\tau$. Informative groupings can leave a small first term, but agreement alone does not eliminate the feasibility cost.

Proposition 5 (Welfare for preference-aligned groups)**.** *Under Assumption 2, any policy* π *satisfying* $\mathrm{Expl}_k(\pi) \leq \varepsilon$ *obeys*

$$W_k(\pi) \geq W_k^\star - \frac{\varepsilon}{2\beta\ell_\beta} - \left(\frac{L}{\ell_\beta} - 1\right) \mathrm{Res}_k. \tag{14}$$

In particular, GLHF's output satisfies this bound simultaneously for every group with $\varepsilon = \varepsilon_{T,k}$ *on the event of Theorem 3. Appendix F.2 gives the proof.*

When Res_k is small, this bound can substantially improve the worst-case guarantee in Proposition 4, although its statistical error has the larger coefficient $1/(2\beta\ell_\beta)$.

For $W_k^\star > 0$ and fixed $\beta > 0$, Proposition 5 implies that distortion approaches one when both $\varepsilon_{T,k}/W_k^\star$ and $\mathrm{Res}_k/W_k^\star$ vanish. The case $\mathrm{Res}_k = 0$ occurs when a fixed feasible lottery attains each group member's highest utility almost surely.

4 Groupwise Learning from Human Feedback

GLHF maintains one *specialist* per group: a feasible lottery $\pi_{t,k} \in \mathcal{B}_\tau$ and a positive outer weight $w_{t,k}$. Before each training round, these states define a policy map for every possible context. The learner then observes one user's context and comparison, updates all specialists whose groups contain

that user, and saves the pre-round map. The deployed policy averages the saved maps. Algorithm 1 shows this algorithm template; Appendix A gives the full procedure and learning-rate choices. Neither utilities nor the Bradley–Terry parameter β are inputs.

Policy and comparison score. For $\tau > 0$, the outer and inner learning rates $\alpha_k, \eta_k \in (0, 1/2]$ are fixed before training from T, p_k, R, τ, and δ (Appendix A). Given the state at round t, the current policy serves context x according to

$$\widehat{\pi}_t(x) := \frac{\sum_{k:G_k(x)=1} \alpha_k w_{t,k} \pi_{t,k}}{\sum_{k:G_k(x)=1} \alpha_k w_{t,k}}. \tag{15}$$

The covering assumption makes this a feasible lottery even when groups overlap. After observing X_t, let a and b be the winner and loser of the queried pair, and let e_c be the cth coordinate vector. With pair probability $\rho_t(a, b)$ from (6) and protocol scale R from (7), form

$$\widehat{g}_t := \frac{\widehat{\pi}_t(b \mid X_t) e_a - \widehat{\pi}_t(a \mid X_t) e_b}{R \rho_t(a, b)}. \tag{16}$$

The pair probability corrects for how the query was selected; R normalizes the score. Conditioning on the pre-round history and X_t, this score satisfies

$$\widehat{\pi}_t(X_t)^\top \widehat{g}_t = 0, \qquad \mathbb{E}[\widehat{g}_t] = \frac{M(X_t)\widehat{\pi}_t(X_t)}{R}, \qquad |\widehat{g}_{t,c}| \le 1, \quad \mathbb{E}[\widehat{g}_{t,c}^2] \le \frac{1}{R}. \tag{17}$$

Thus the current policy has exactly zero estimated advantage over itself, while a specialist's score estimates its normalized pairwise advantage over that policy. The $1/R$ second moment yields the different active and passive rates in Theorem 3; Appendix B verifies all four properties.

Two updates and one output. Each active group k, with $G_k(X_t) = 1$, makes an outer multiplicative update

$$w_{t+1,k} = w_{t,k} \left(1 + \alpha_k \pi_{t,k}^\top \widehat{g}_t\right) \tag{18}$$

and an inner exponentiated-gradient step followed by KL projection:

$$\pi_{t+1/2,k}(c) = \frac{\pi_{t,k}(c) e^{\eta_k \widehat{g}_{t,c}}}{\sum_d \pi_{t,k}(d) e^{\eta_k \widehat{g}_{t,d}}}, \qquad \pi_{t+1,k} = \operatorname*{arg\,min}_{q \in \mathcal{B}_\tau} \mathrm{KL}(q \| \pi_{t+1/2,k}). \tag{19}$$

Inactive states do not change. The projection keeps every specialist feasible; it reduces to a one-dimensional search described in Appendix C. The outer update is a second-order expert update (Gaillard et al., 2014), and the inner update is entropic mirror descent for linear rewards (Kivinen & Warmuth, 1997; Beck & Teboulle, 2003).

Algorithm 1 GLHF training and deployment sketch

if $\tau = 0$ **then**
  return $\pi_{\mathrm{GLHF}}(x) = \pi_{\mathrm{ref}}$ for every x.
Fix α_k, η_k as in Appendix A; initialize $w_{1,k} = 1/K$ and $\pi_{1,k} = \pi_{\mathrm{ref}}$ for all k.
for $t = 1, \dots, T$ **do**
  Save the pre-round map $x \mapsto \widehat{\pi}_t(x)$ from (15).
  Observe X_t; query one pair by the selected protocol and form $\widehat{g}_t$ from (16).
  For each k with $G_k(X_t) = 1$, apply (18)–(19); leave other states unchanged.
return $\pi_{\mathrm{GLHF}}(x) = \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} \widehat{\pi}_t(x)$.

The saved maps and their average lie pointwise in $\mathcal{B}_\tau$. At deployment, a new user's group memberships determine the response lottery $\pi_{\mathrm{GLHF}}(x)$; no further comparisons are needed.

5 WHY THE COMPARISON GUARANTEE HOLDS

Suppose $\tau > 0$ below. One observed comparison supplies a score to every group containing its user. The proof separates two issues: can a group's specialist outperform the current policy by much, and can a fixed feasible lottery outperform that specialist? Fix a group k and a comparator $q \in \mathcal{B}_\tau$. Their cumulative sampled advantage decomposes as

$$A_{T,k}(q) := \sum_{t:G_k(X_t)=1} q^\top \widehat{g}_t = \sum_{t:G_k(X_t)=1} \pi_{t,k}^\top \widehat{g}_t + \sum_{t:G_k(X_t)=1} (q - \pi_{t,k})^\top \widehat{g}_t. \tag{20}$$

The outer weights control the first sum; the inner KL projection controls the second.

Regret on overlapping groups. Both regret bounds depend on the same cumulative squared-score quantity,

$$V_{T,k} := \sum_{t:G_k(X_t)=1} \sum_a \pi_{t,k}(a) \widehat{g}_{t,a}^2. \tag{21}$$

The following lemma summarizes their deterministic guarantees. The log-weight and KL calculations are in Appendix E.

Lemma 6 (Outer and inner regret)**.** *All outer weights remain positive and sum to one. All specialist and current policy lotteries have full support and lie in $\mathcal{B}_\tau$. For every group k and every $q \in \mathcal{B}_\tau$,*

$$\sum_{t:G_k(X_t)=1} \pi_{t,k}^\top \widehat{g}_t \le \frac{\log K}{\alpha_k} + \alpha_k V_{T,k}, \tag{22}$$

$$\sum_{t:G_k(X_t)=1} (q - \pi_{t,k})^\top \widehat{g}_t \le \frac{\tau}{\eta_k} + \eta_k V_{T,k}. \tag{23}$$

The key invariant is that the outer update conserves total weight:

$$\sum_j (w_{t+1,j} - w_{t,j}) = \sum_{j:G_j(X_t)=1} \alpha_j w_{t,j} \pi_{t,j}^\top \widehat{g}_t = 0. \tag{24}$$

The last equality holds because the weighted mixture in (15) has zero score by (17). Thus even when one user activates several groups, their outer updates cannot collectively increase weight. Since $w_{1,k} = 1/K$ and $w_{T+1,k} \le 1$, a logarithmic potential bounds the first sum in Lemma 6. The second follows by telescoping the KL divergence to q, initially at most τ. Together with (20), the lemma bounds $A_{T,k}(q)$ by $\log(K)/\alpha_k + \tau/\eta_k + (\alpha_k + \eta_k)V_{T,k}$.

From sampled scores to population margins. Write $\mathcal{F}_{t-1}$ for the history before round t and $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. The history fixes the complete policy $x \mapsto \widehat{\pi}_t(x)$ before the independent context X_t arrives. Score unbiasedness therefore gives $R\mathbb{E}_{t-1}[G_k(X_t) q^\top \widehat{g}_t] = p_k \mathbb{E}_k[q^\top M(X) \widehat{\pi}_t(X)]$, where the right side evaluates this fixed policy on a fresh group-k user. The discrepancy between sampled and conditional-expected score vectors is

$$S_k := \sum_{t=1}^{T} \big(\mathbb{E}_{t-1}[G_k(X_t) \widehat{g}_t] - G_k(X_t) \widehat{g}_t \big). \tag{25}$$

The output policy is the pointwise average $\pi_{\mathrm{GLHF}}(x) = T^{-1} \sum_{t=1}^T \widehat{\pi}_t(x)$. Summing the conditional score identity over rounds and using the definitions of $A_{T,k}(q)$ and S_k gives the exact population identity

$$T p_k \mathbb{E}_k[q^\top M(X) \pi_{\mathrm{GLHF}}(X)] = R\big(A_{T,k}(q) + q^\top S_k \big). \tag{26}$$

Appendix D.1 controls both $V_{T,k}$ and $\sup_{q \in \mathcal{B}_\tau} q^\top S_k$ simultaneously for all groups. The learning rates in Appendix A balance each initial regret budget ($\log K$ or τ) against its squared-score cost. Substituting the regret and concentration bounds into (26) gives, on one event of probability at least $1 - \delta$,

$$\mathbb{E}_k[q^\top M(X) \pi_{\mathrm{GLHF}}(X)] \le 6\sqrt{\frac{RH}{T p_k}} + 6\frac{RH}{T p_k} = \varepsilon_{T,k} \qquad \text{for every } k \text{ and } q \in \mathcal{B}_\tau.$$

Taking the supremum over q proves Theorem 3; Appendix E.4 contains the details.

6 EMPIRICAL EVALUATION

We test whether learning a group-conditioned policy from pairwise comparisons can improve welfare relative to learning a single lottery for all users. We run two exploratory studies, one using human-provided coffee ratings and the other using LLM-provided synthetic ratings of candidate prompt completions. Rescaled ratings are treated as cardinal utilities: they determine Bradley-Terry comparisons over alternatives and policy welfare.

6.1 EXPERIMENTAL SETUP

Our primary comparison is *multigroup GLHF* against *batch NLHF*, which estimates a preference margin matrix and fits a constrained maximal lottery to serve the population. We also compare against *pooled GLHF*—the GLHF learner run with only the universal group which serves one lottery to everyone. Both GLHF methods use active sampling, whereas NLHF learns a lottery from a single batch (passive). Within each run, all three learners use the same training respondents but may train on different pairwise comparisons due to their sampling methods.

We set both the learner and fixed-lottery benchmark KL radii to 0.5 around the uniform reference. In the main experiment, comparisons follow rating order with random tie-breaking ($\beta = \infty$), reducing preference noise to isolate learning dynamics. Additionally, the experiments use adaptive learning rates, specified in Appendix G.7.1; Theorem 3 does not directly cover this variant.

Although the worst-case welfare factor L/ℓ_β is infinite in this limit, empirical distortion remains comparably low to the finite-β setting; in Appendix G.5 we explore $\beta \in \{\frac{1}{2}, 1, 2, 4\}$ and several learner radii τ. Multigroup GLHF achieves lower mean worst-group distortion than the group-agnostic learners in every setting we test, including passive sampling settings.

For each held-out group, empirical distortion is

$$\widehat{\mathrm{Dist}}_k(\pi) = \frac{\text{welfare of best fixed feasible lottery for group } k}{\text{welfare of deployed policy } \pi \text{ on group } k}. \tag{27}$$

Every group evaluates the deployed policy against its own best fixed feasible lottery. A distortion of one means the policy matches the welfare of the best fixed lottery. Since a group-conditioned policy can serve different lotteries to different users, it may achieve distortion below one.

Both studies reserve about 10% of users for group selection. We chose to showcase groups that would likely be disadvantaged by group-agnostic methods, as well as groups with preferences similar to the population average (Appendix G.1). Groups remain fixed during ten repeats of ten-fold cross-validation on the remaining data, with 50 learner seeds per split. Distortion is calculated on each held-out fold. Worst-group distortion takes the maximum across groups within each run before averaging, while per-group distortion is computed separately for each group; Appendix G.2.2 describes the paired inference procedure.

6.2 COFFEE-RATING STUDY

The Great American Coffee Taste Test (TidyTuesday, 2024) has ratings of four coffees from 3,369 respondents after setting aside the discovery sample. We evaluate six overlapping groups, including the universal group and groups defined by age, coffee expertise, gender, having children, and working from home (Appendix G.1).

Multigroup GLHF has lower point distortion than all three comparators in every group of coffee-rating respondents (Figure 1). Mean within-run worst-group distortion is 1.08 versus 1.28 for batch NLHF, a paired reduction of 0.202 (corrected 95% interval $[0.151, 0.253]$). Pooled GLHF scores 1.25. The uniform reference scores 1.12, better than either group-agnostic learner: population-wide learning raises population welfare but lowers welfare for several subgroups. Multigroup GLHF also improves on the reference, by 0.042 $[0.032, 0.052]$.

6.3 COMPLETION-RATING STUDY

Claude Haiku 4.5 rated six fixed answers to *Why is the sky blue?* under sixteen persona prompts crossing age, education, technical background, and preferred writing style. Each answer was rated

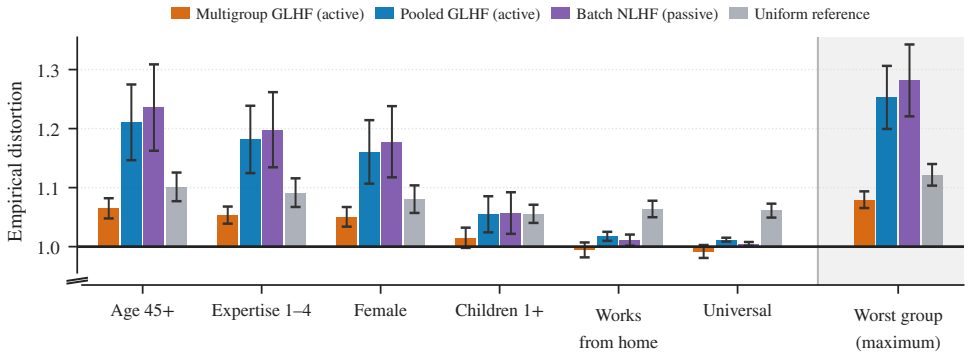

Figure 1: Empirical distortion in the coffee-rating study at $\beta = \infty$ and KL radius 0.5: active multigroup and pooled GLHF, passive batch NLHF, and the uniform reference. The line at one marks each group's best fixed feasible lottery; lower is better. The rightmost cluster averages the within-run group maximum. Whiskers are approximate 95% confidence intervals corrected for overlapping cross-validation folds.

separately; independently drawn ratings within personas were combined to form 1,497 *pseudo-respondents*. Appendix G.1 details this construction and group selection, and Appendix G.7.4 gives the prompts and answers.

Mean worst-group distortion is 1.07 for multigroup GLHF versus 1.36 for batch NLHF, a paired reduction of 0.295 $[0.282, 0.308]$. Pooled GLHF scores 1.35. The uniform reference scores 1.74, far worse than in the coffee-rating study, in part because two factually incorrect alternatives were rarely preferred. The groupwise results are in Appendix G.3.

6.4 Active and passive sampling

Active sampling performs somewhat better than passive for multigroup GLHF, reducing worst-group distortion by 0.012 $[0.007, 0.017]$ in the coffee-rating study and 0.029 $[0.026, 0.033]$ in the completion-rating study. The advantage of active sampling is further described in Appendix G.4.

7 Discussion

GLHF learns one context-dependent policy from one comparison per sampled user. Simultaneously for every prespecified, possibly overlapping group, no feasible fixed lottery has more than a vanishing average pairwise advantage over that policy. Under individual Bradley–Terry preferences, the resulting groupwise distortion bound approaches the worst-case optimal factor L/ℓ_β as the comparison error vanishes. For a fixed target comparison error, policy-adaptive sampling reduces the leading alternative-count factor in sample complexity from quadratic to linear, while retaining the same comparison budget per user. The guarantee strengthens for any group with similar member preferences, with distortion approaching one when a common feasible lottery nearly maximizes its individual users' utilities and the comparison error vanishes. In the exploratory coffee-rating and completion-rating studies, GLHF reduces worst-group empirical distortion relative to NLHF and other group-agnostic baselines.

The theory addresses one prompt, finitely many candidate responses, and prespecified groups. Extending it to multiple prompts or estimated group memberships would broaden its applicability.

Reproducibility statement

All theoretical assumptions, sampling rules, algorithms, and proofs are specified in the text and appendices. Appendix C gives the KL projection, and Appendix A gives the computational cost. Section 6 and Appendix G describe the coffee-rating and completion-rating studies, including group selection, held-out evaluation, sensitivity checks, and collection materials. The supplementary ZIP contains the code, both input datasets, and saved evaluation results; Appendix G.7.5 describes its contents and the command for reproducing the experiments and figures.

Use of generative AI

Generative AI was used to assist with coding, proof writing, manuscript writing, and synthetic data generation. The final paper is written by the authors, who take full responsibility for clarity and correctness.

References

Krishna Acharya, Eshwar Ram Arunachaleswaran, Sampath Kannan, Aaron Roth, and Juba Ziani. Oracle efficient algorithms for groupwise regret. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1c9ee8c9e1a0c7de49e15720c4ce6f78-Abstract-Conference.html.

Elliot Anshelevich, Aris Filos-Ratsikas, Nisarg Shah, and Alexandros A. Voudouris. Distortion in social choice problems: The first 15 years and beyond. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pp. 4294–4301, 2021. doi: 10.24963/ijcai.2021/589.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. doi: 10.1016/S0167-6377(02)00231-6.

Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In *11th Innovations in Theoretical Computer Science Conference*, volume 151 of *Leibniz International Proceedings in Informatics*, pp. 55:1–55:24, 2020. doi: 10.4230/LIPIcs.ITCS.2020.55. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2020.55.

Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D. Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, 227:190–213, 2015. doi: 10.1016/j.artint.2015.06.003.

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

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9346–9360, 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Imre Csiszár and Paul C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004. doi: 10.1561/0100000004.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 563–587, 2015. URL https://proceedings.mlr.press/v40/Dudik15.html.

Peter C. Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984. doi: 10.2307/2297786.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 176–196, 2014. URL https://proceedings.mlr.press/v35/gaillard14.html.

Ira Globus-Harris, Michael Kearns, and Aaron Roth. An algorithmic framework for bias bounties. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1106–1124. Association for Computing Machinery, 2022. doi: 10.1145/3531146.3533172. URL https://arxiv.org/abs/2201.10408.

Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems 38*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2526c5e8110bc6bc8b462ba95198161e-Abstract-Conference.html.

Daniel Halpern, Evi Micha, Ariel D. Procaccia, and Itai Shapira. Pairwise calibrated rewards for pluralistic alignment. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-1940. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/53dbd7e34fab703a639964e2d3ee9e84-Abstract-Conference.html.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Kihyun Kim, Jiawei Zhang, Asuman Ozdaglar, and Pablo A. Parrilo. Beyond RLHF and NLHF: Population-proportional alignment under an axiomatic framework. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4a05804445b94eb76108bc54f16f73ca-Abstract-Conference.html.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew M. 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. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track*, pp. 105236–105344, 2024. doi: 10.52202/079017-3342. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997. doi: 10.1006/inco.1996.2612.

Brenden Latham and Mehrdad Moharrami. Offline constrained RLHF with multiple preference oracles. *arXiv preprint arXiv:2604.00200*, 2026. URL https://arxiv.org/abs/2604.00200.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proceedings of the 41st International*

Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768, 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3): 239–281, 2003. doi: 10.1023/A:1024068626366.

Kazusato Oko, Annie Ulichney, Nika Haghtalab, and Han Bao. Distortion of AI alignment revisited: RLHF is a decent utilitarian aligner. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2609.12651.

OpenAI. Custom instructions for ChatGPT. https://openai.com/index/custom-instructions-for-chatgpt/, 2023.

OpenAI. Memory and new controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. RLHF from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024. URL https://arxiv.org/abs/2405.00254.

Ariel D. Procaccia and Jeffrey S. Rosenschein. The distortion of cardinal preferences in voting. In *Cooperative Information Agents X*, volume 4149 of *Lecture Notes in Computer Science*, pp. 317–331, 2006. doi: 10.1007/11839354_23.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free RLHF. In *Advances in Neural Information Processing Systems 37*, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/4147dfaa46cd7e20a2aecb91097ae8cc-Abstract-Conference.html.

Guy N. Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9107–9115, 2021. URL https://proceedings.mlr.press/v139/rothblum21a.html.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 968–994, 2022. URL https://proceedings.mlr.press/v167/saha22a.html.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 29971–30004, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-bayes-bernstein inequality for martingales and its application to multiarmed bandits. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pp. 98–111, 2012. URL https://proceedings.mlr.press/v26/seldin12a.html.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

TidyTuesday. The great american coffee taste test. Dataset release, May 14, 2024. URL https://github.com/rfordatascience/tidytuesday/tree/main/data/2024/2024-05-14.

Christopher J. Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21633–21657, 2022. URL https://proceedings.mlr.press/v162/tosh22a.html.

Nuoya Xiong and Aarti Singh. Projection optimization: A general framework for multi-objective and multi-group RLHF. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69017–69058, 2025. URL https://proceedings.mlr.press/v267/xiong25c.html.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ffd4f5a2ea6b93e9bf5af9264d568cf2-Abstract-Conference.html.

Huiying Zhong, Tianwei Gao, Zhiwei Steven Wu, Linjun Zhang, Weijie J. Su, and Zhun Deng. Provable pluralistic alignment: Multi-party RLHF under offline human feedback. *arXiv preprint arXiv:2403.05006v2*, 2026. URL https://arxiv.org/abs/2403.05006v2. First circulated in 2024.

A GLHF PSEUDOCODE

Algorithm 2 gives the full procedure described in Section 4. For $\tau > 0$, the predetermined bound on the squared-score sum $V_{T,k}$ from (21) is

$$V_k^{\mathrm{ub}} := \frac{2Tp_k}{R} + \log\frac{2K}{\delta}. \tag{28}$$

The score's conditional second moment is at most $1/R$ by (17); Appendix D.1 proves $V_{T,k} \le V_k^{\mathrm{ub}}$ simultaneously for all groups on the high-probability event. Balance the outer and inner regret budgets against this bound by setting

$$\alpha_k := \min\left\{\frac{1}{2}, \sqrt{\frac{\log K}{V_k^{\mathrm{ub}}}}\right\}, \qquad \eta_k := \min\left\{\frac{1}{2}, \sqrt{\frac{\tau}{V_k^{\mathrm{ub}}}}\right\}. \tag{29}$$

These rates use only the known group masses, horizon, radius, and confidence level. Appendix E.5 checks the clipped cases.

Algorithm 2 Groupwise Learning from Human Feedback (GLHF)

Require: $T \ge 2$, $\delta \in (0,1)$, covering indicators G_k and masses $p_k > 0$, full-support π_{ref}, radius $\tau \ge 0$, passive or active comparison sampling protocol

if $\tau = 0$ **then**
  return $\pi_{\mathrm{GLHF}}(x) = \pi_{\mathrm{ref}}$ for every x
Set R by (7) and rates by (29)
Initialize $w_{1,k} = 1/K$ and $\pi_{1,k} = \pi_{\mathrm{ref}}$ for every k
for $t = 1, \dots, T$ **do**
  Save the pre-round state and fix $x \mapsto \widehat{\pi}_t(x)$ by (15)
  Observe X_t; sample a pair by the chosen protocol in Section 2
  Observe the winning and losing alternatives to form $\widehat{g}_t$ by (16)
  for each group k with $G_k(X_t) = 1$ **do**
    Update $w_{t+1,k}$ and $\pi_{t+1,k}$ from the pre-update state by (18)–(19)
  Leave inactive groups unchanged
return $\pi_{\mathrm{GLHF}}(x) = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \widehat{\pi}_t(x)$

Computation. On each round, GLHF updates the weight and lottery for each group containing the sampled user. Appendix C reduces each KL projection to a one-dimensional search. If a numerical implementation uses at most J search steps per projection, each costing $O(m)$, the updates take $O(Km(1+J))$ arithmetic operations per round in the worst case. The current weights and lotteries use $O(Km)$ space; saving all T pre-round states uses $O(TKm)$ space, and evaluating the returned policy at a new context costs $O(TKm)$ operations.

B THE COMPARISON SCORE

This appendix verifies the comparison-score bounds used in Section 4 for the sampling protocols of Section 2.

Proposition 7 (Comparison-score properties)**.** *Under either sampling protocol, the score $\widehat{g}_t$ in* (16) *satisfies* (17)*, with R given by* (7)*.*

Proof. Condition throughout on $(\mathcal{F}_{t-1}, X_t)$, and write $\pi = \widehat{\pi}_t(X_t)$ and $\rho(c,d) = \rho_t(c,d)$. The lottery π is fixed. By Appendix E.1, it has full support, so $\rho(c,d) > 0$ for every pair of distinct alternatives under either protocol. Since $\rho(c,d)$ is the probability of the unordered pair $\{c,d\}$, we also have $\rho(c,d) = \rho(d,c)$. Label the observed winner a and loser b, as in (16). The two nonzero coordinates of $\widehat{g}_t$ give the exact cancellation

$$\pi^\top \widehat{g}_t = \frac{\pi(a)\pi(b) - \pi(b)\pi(a)}{R\rho(a,b)} = 0.$$

Fix a coordinate c. If the sampled pair is $\{c, d\}$, then $\widehat{g}_{t,c} = \pi(d)/(R\rho(c,d))$ when c wins and $\widehat{g}_{t,c} = -\pi(d)/(R\rho(c,d))$ when d wins. The pair is selected independently of U_t given the history and X_t. Hence, conditional on sampling $\{c, d\}$, the probability that c wins minus the probability that d wins is $M_{cd}(X_t)$, and

$$\begin{aligned}\mathbb{E}[\widehat{g}_{t,c}] &= \sum_{d \neq c} \rho(c,d) \frac{\pi(d)}{R\rho(c,d)} M_{cd}(X_t) \\ &= \frac{1}{R} \sum_{d \neq c} M_{cd}(X_t)\pi(d) = \frac{(M(X_t)\pi)_c}{R}.\end{aligned}$$

The coordinate $\widehat{g}_{t,c}$ is zero unless c is in the sampled pair. If the sampled pair is $\{c, d\}$, then (6) gives $|\widehat{g}_{t,c}| = \pi(d)/(\pi(c) + \pi(d))$ under active sampling and $|\widehat{g}_{t,c}| = \pi(d)$ under passive sampling. Both are at most one.

Averaging over the sampled pairs gives

$$\begin{aligned}\mathbb{E}[\widehat{g}_{t,c}^2] &= \frac{1}{R^2} \sum_{d \neq c} \frac{\pi(d)^2}{\rho(c,d)} \\ &= \begin{cases} \dfrac{1}{R} \displaystyle\sum_{d \neq c} \dfrac{\pi(d)^2}{\pi(c) + \pi(d)}, & \text{active,} \\ \dfrac{1}{R} \displaystyle\sum_{d \neq c} \pi(d)^2, & \text{passive} \end{cases} \\ &\leq \frac{1}{R} \sum_{d \neq c} \pi(d) \leq \frac{1}{R}.\end{aligned}$$

□

C KL PROJECTION

An inner learner's exponentiated step in (19) may leave $\mathcal{B}_\tau$. Here we show how to project it back with a one-dimensional search and prove the KL projection inequality used in the regret analysis.

The map $q \mapsto \mathrm{KL}(q\|\pi_{\mathrm{ref}})$ is continuous and convex on the simplex, so $\mathcal{B}_\tau$ is compact, convex, and contains π_{ref}. Thus mixtures of feasible lotteries remain feasible. The radius cap in Section 2 follows from

$$\mathrm{KL}(q\|\pi_{\mathrm{ref}}) = \sum_a q(a) \log q(a) - \sum_a q(a) \log \pi_{\mathrm{ref}}(a) \leq \log \frac{1}{\min_a \pi_{\mathrm{ref}}(a)},$$

with equality at a point mass on an alternative of minimum reference probability. The KL projection inequality is standard (Csiszár & Shields, 2004, Theorem 3.1). We give a direct proof and a one-dimensional way to find the projection.

For a full-support $v \in \Delta(\mathcal{A})$, consider the geometric path

$$q_s(a) := \frac{\pi_{\mathrm{ref}}(a)^{1-s} v(a)^s}{Z_s}, \qquad Z_s := \sum_b \pi_{\mathrm{ref}}(b)^{1-s} v(b)^s, \qquad s \in [0,1]. \tag{30}$$

At $s = 0$ this is π_{ref}, at $s = 1$ it is v, and every q_s has full support.

Proposition 8 (Exact projection)**.** *If $v \in \mathcal{B}_\tau$, its KL projection onto $\mathcal{B}_\tau$ is v. If $v \notin \mathcal{B}_\tau$ and $\tau > 0$, there is a unique $s_\star \in (0,1)$ satisfying $\mathrm{KL}(q_{s_\star}\|\pi_{\mathrm{ref}}) = \tau$, and*

$$\arg\min_{q \in \mathcal{B}_\tau} \mathrm{KL}(q\|v) = q_{s_\star}. \tag{31}$$

At $\tau = 0$, the projection is π_{ref}. Writing p for this projection, every feasible q satisfies

$$\mathrm{KL}(q\|p) + \mathrm{KL}(p\|v) \leq \mathrm{KL}(q\|v). \tag{32}$$

Proof. If $v \in \mathcal{B}_\tau$, its projection is v; if $\tau = 0$, the feasible set is $\{\pi_{\mathrm{ref}}\}$. In either case (32) holds with equality. Otherwise, write $h_a = \log(v(a)/\pi_{\mathrm{ref}}(a))$ and $A(s) = \log \sum_a \pi_{\mathrm{ref}}(a) e^{s h_a}$. Then

$q_s(a) = \pi_{\mathrm{ref}}(a)e^{sh_a - A(s)}$. Direct differentiation gives $A'(s) = \mathbb{E}_{q_s}[h]$ and $A''(s) = \mathrm{Var}_{q_s}(h)$. Moreover

$$\mathrm{KL}(q_s\|\pi_{\mathrm{ref}}) = sA'(s) - A(s), \qquad \frac{\mathrm{d}}{\mathrm{d}s}\mathrm{KL}(q_s\|\pi_{\mathrm{ref}}) = s\,\mathrm{Var}_{a\sim q_s}(h_a).$$

Since $v \neq \pi_{\mathrm{ref}}$, h is nonconstant and $\mathrm{Var}_{q_s}(h) > 0$ by full support. Thus the derivative is positive for $s > 0$, and $\mathrm{KL}(q_s\|\pi_{\mathrm{ref}})$ increases continuously from zero to $\mathrm{KL}(v\|\pi_{\mathrm{ref}}) > \tau$, giving a unique $s_\star \in (0, 1)$.

To identify this boundary point as the projection, use $\log q_s(a) = (1-s)\log \pi_{\mathrm{ref}}(a) + s\log v(a) - \log Z_s$ from (30). Expanding the KL terms gives, for any $q \in \Delta(\mathcal{A})$ and $s \in [0, 1]$,

$$\mathrm{KL}(q\|q_s) = s\big[\mathrm{KL}(q\|v) - \mathrm{KL}(q_s\|v)\big] + (1-s)\big[\mathrm{KL}(q\|\pi_{\mathrm{ref}}) - \mathrm{KL}(q_s\|\pi_{\mathrm{ref}})\big]. \tag{33}$$

Set $p = q_{s_\star}$. For feasible q, the second bracket is nonpositive because $\mathrm{KL}(p\|\pi_{\mathrm{ref}}) = \tau$. Since $s_\star < 1$,

$$\mathrm{KL}(q\|v) - \mathrm{KL}(p\|v) \geq \frac{1}{s_\star}\mathrm{KL}(q\|p) \geq \mathrm{KL}(q\|p).$$

For $q \neq p$, $\mathrm{KL}(q\|p) > 0$, so the first inequality proves uniqueness; the second gives (32). □

D Concentration for groupwise scores

The population conversion in (26) requires bounds on $V_{T,k}$ and on $q^\top S_k$ for every fixed feasible lottery q. We work at the fixed horizon T, with $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ as in Section 5.

Lemma 9 (A bounded nonnegative sum)**.** *Let $b > 0$ and let $Z_t \in [0, b]$ depend only on the history through round t. Suppose $\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t] \leq u$ almost surely for a deterministic $u \geq 0$. Then, for every $\gamma \in (0, 1)$,*

$$\Pr\left(\sum_{t=1}^T Z_t > 2u + b\log\frac{1}{\gamma}\right) \leq \gamma.$$

Proof. Put $\mu_t = \mathbb{E}_{t-1}[Z_t]$. The chord bound $e^z \leq 1 + (e-1)z$ for $z \in [0, 1]$ gives $\mathbb{E}_{t-1}[e^{Z_t/b}] \leq 1 + (e-1)\mu_t/b \leq e^{(e-1)\mu_t/b}$. For each running round $s = 0, \ldots, T$, define

$$\mathcal{L}_s := \exp\left(\frac{1}{b}\sum_{t=1}^s [Z_t - (e-1)\mu_t]\right), \qquad \mathcal{L}_0 = 1.$$

The process is positive and adapted to $(\mathcal{F}_s)$. Since μ_s is $\mathcal{F}_{s-1}$-measurable, the preceding bound implies

$$\mathbb{E}[\mathcal{L}_s \mid \mathcal{F}_{s-1}] = \mathcal{L}_{s-1}e^{-(e-1)\mu_s/b}\mathbb{E}[e^{Z_s/b} \mid \mathcal{F}_{s-1}] \leq \mathcal{L}_{s-1}.$$

Thus $(\mathcal{L}_s)_{s=0}^T$ is a nonnegative supermartingale with respect to $(\mathcal{F}_s)$, and $\mathbb{E}[\mathcal{L}_T] \leq 1$. Since $e - 1 < 2$ and $\sum_t \mu_t \leq u$, $e^{(\sum_t Z_t - 2u)/b} \leq \mathcal{L}_T$, so $\mathbb{E}[e^{(\sum_t Z_t - 2u)/b}] \leq 1$. Markov's inequality gives the claim. □

The next lemma combines the Bernstein exponential argument for bounded martingales (Freedman, 1975) with the finite KL change-of-measure inequality (Seldin et al., 2012, Lemma 6). We give the fixed-horizon form and constants used here.

Lemma 10 (One martingale event for every feasible lottery)**.** *Let $Z_t \in \mathbb{R}^m$ be martingale differences with $|Z_{t,a}| \leq b$ and $\sum_{t=1}^T \mathbb{E}_{t-1}[Z_{t,a}^2] \leq v$ almost surely for every coordinate a, where $b > 0$ and $v \geq 0$ are deterministic. Let π_{ref} be the full-support reference, fixed before the sequence is observed. For $\tau \geq 0$ and $\gamma \in (0, 1)$, with probability at least $1 - \gamma$,*

$$\sup_{q\in\mathcal{B}_\tau} q^\top \sum_{t=1}^T Z_t \leq 2\sqrt{v\left(\tau + \log\frac{1}{\gamma}\right)} + \frac{4}{3}b\left(\tau + \log\frac{1}{\gamma}\right). \tag{34}$$

Proof. If $v = 0$, all increments vanish almost surely. Otherwise put $S = \sum_t Z_t$, $h = \tau + \log(1/\gamma)$, and choose the deterministic $\zeta = \min\{3/(2b), \sqrt{h/v}\}$. For $\sigma_{t,a}^2 = \mathbb{E}_{t-1}[Z_{t,a}^2]$, the bounded-martingale Bernstein moment bound gives

$$\mathbb{E}_{t-1}[e^{\zeta Z_{t,a}}] \le \exp\left(\frac{\zeta^2 \sigma_{t,a}^2}{2(1 - b\zeta/3)}\right) \le e^{\zeta^2 \sigma_{t,a}^2}.$$

The first inequality follows by expanding the exponential, using $\mathbb{E}_{t-1}[Z_{t,a}] = 0$, $|Z_{t,a}|^j \le b^{j-2} Z_{t,a}^2$, and $j! \ge 2 \cdot 3^{j-2}$ for $j \ge 2$. The second inequality uses $b\zeta \le 3/2$. For each a, $\mathcal{L}_{s,a} = \exp\big(\zeta \sum_{t=1}^s Z_{t,a} - \zeta^2 \sum_{t=1}^s \sigma_{t,a}^2\big)$ is therefore a nonnegative supermartingale with $\mathcal{L}_{0,a} = 1$. Since $\sum_t \sigma_{t,a}^2 \le v$, $\mathbb{E}[e^{\zeta S_a - \zeta^2 v}] \le \mathbb{E}[\mathcal{L}_{T,a}] \le 1$ for each a. Hence the fixed mixture $\mathcal{Z}_\zeta = \sum_a \pi_{\mathrm{ref}}(a) e^{\zeta S_a}$ obeys $\mathbb{E}[e^{-\zeta^2 v} \mathcal{Z}_\zeta] \le 1$. By Markov's inequality, with probability at least $1 - \gamma$,

$$\log \mathcal{Z}_\zeta \le \zeta^2 v + \log(1/\gamma).$$

This is one event for the entire mixture; no union bound over the m coordinates is needed. On this event, define $r_\zeta(a) = \pi_{\mathrm{ref}}(a) e^{\zeta S_a} / \mathcal{Z}_\zeta$. For every $q \in \Delta(\mathcal{A})$,

$$0 \le \mathrm{KL}(q \| r_\zeta) = \mathrm{KL}(q \| \pi_{\mathrm{ref}}) - \zeta q^\top S + \log \mathcal{Z}_\zeta.$$

Thus, simultaneously for every $q \in \mathcal{B}_\tau$,

$$q^\top S \le \frac{\mathrm{KL}(q \| \pi_{\mathrm{ref}}) + \log \mathcal{Z}_\zeta}{\zeta} \le \frac{h}{\zeta} + \zeta v.$$

If $\zeta = \sqrt{h/v}$, the last expression is $2\sqrt{vh}$. Otherwise $\zeta = 3/(2b)$, so $v \le 4b^2 h/9$ and the bound is at most $2bh/3 + 3v/(2b) \le 4bh/3$. These two cases imply (34). □

D.1 The simultaneous concentration event

Recall the realized second moment $V_{T,k}$ from (21), its predetermined bound V_k^{ub} from (28), and the sampling-error martingale S_k from (25).

Lemma 11 (One event for all groups and comparators)**.** *With probability at least $1 - \delta$, simultaneously for every group $k \in [K]$,*

$$\begin{aligned} V_{T,k} &\le V_k^{\mathrm{ub}}, \\ \sup_{q \in \mathcal{B}_\tau} q^\top S_k &\le 2\sqrt{\frac{T p_k}{R}\left(\tau + \log\frac{2K}{\delta}\right)} + \frac{8}{3}\left(\tau + \log\frac{2K}{\delta}\right). \end{aligned} \tag{35}$$

Proof. Fix k. Set $v_{t,k} = G_k(X_t) \sum_a \pi_{t,k}(a) \widehat{g}_{t,a}^2$, so that $V_{T,k} = \sum_t v_{t,k}$. The score bound $|\widehat{g}_{t,a}| \le 1$ and the fact that $\pi_{t,k}$ is a lottery give $v_{t,k} \in [0,1]$. The lottery $\pi_{t,k}$ is fixed by $\mathcal{F}_{t-1}$. Conditioning first on $(\mathcal{F}_{t-1}, X_t)$ and using (17) gives

$$\begin{aligned} \mathbb{E}_{t-1}[v_{t,k}] &= \mathbb{E}_{t-1}\left[G_k(X_t) \sum_a \pi_{t,k}(a) \mathbb{E}[\widehat{g}_{t,a}^2 \mid \mathcal{F}_{t-1}, X_t]\right] \\ &\le \frac{1}{R} \mathbb{E}_{t-1}\left[G_k(X_t) \sum_a \pi_{t,k}(a)\right] \\ &= \frac{1}{R} \mathbb{E}_{t-1}[G_k(X_t)] = \frac{p_k}{R}. \end{aligned}$$

The last equality uses independence of the fresh context from the history. Lemma 9 with $b = 1$, $u = Tp_k/R$, and $\gamma = \delta/(2K)$ therefore gives $V_{T,k} \le 2Tp_k/R + \log(2K/\delta) = V_k^{\mathrm{ub}}$ except with probability $\delta/(2K)$.

For the sampling error, set $Y_{t,k} = G_k(X_t)\widehat{g}_t$. The martingale differences $Z_t = \mathbb{E}_{t-1}[Y_{t,k}] - Y_{t,k}$ sum to S_k and have $|Z_{t,a}| \le 2$, since both $Y_{t,k,a}$ and its conditional mean lie in $[-1, 1]$. For each

coordinate, the same conditioning argument gives

$$\begin{aligned}\mathbb{E}_{t-1}[Z_{t,a}^2] &= \mathrm{Var}(Y_{t,k,a} \mid \mathcal{F}_{t-1}) \\ &\le \mathbb{E}_{t-1}[Y_{t,k,a}^2] \\ &= \mathbb{E}_{t-1}\big[G_k(X_t)\mathbb{E}[\widehat{g}_{t,a}^2 \mid \mathcal{F}_{t-1}, X_t]\big] \\ &\le \frac{1}{R}\mathbb{E}_{t-1}[G_k(X_t)] = \frac{p_k}{R}.\end{aligned}$$

Thus $\sum_t \mathbb{E}_{t-1}[Z_{t,a}^2] \le Tp_k/R$ for every coordinate a. Lemma 10 with $b = 2$, $v = Tp_k/R$, and $\gamma = \delta/(2K)$ gives the second bound except with probability $\delta/(2K)$. A union bound over these two events for each of the K groups proves the claim. □

E Proof details for the comparison guarantee

E.1 Positivity and pointwise feasibility

Initially, the outer weights are positive and each specialist lottery is the full-support feasible reference. The positivity of the outer weights and the full support and feasibility of the lotteries in Lemma 6 follow by induction. If these properties hold at round t, then because every context belongs to at least one group, deployment is a well-defined, full-support convex combination of feasible lotteries. The queried-pair probabilities are positive, and Appendix B shows that the comparison score satisfies $|\widehat{g}_{t,a}| \le 1$ for every a. Each active outer weight is therefore multiplied by a factor in $[1/2, 3/2]$. Exponentiation preserves full support, and the KL projection in Proposition 8 preserves full support and restores feasibility. Inactive states are unchanged, completing the induction.

E.2 The outer log-weight bound

Equation (24) shows that the outer weights sum to one at every round. Together with positivity, this gives $w_{T+1,k} \le 1$, which will bound the log-weight ratio below.

Fix a group k and write $r_{t,k} := \pi_{t,k}^\top \widehat{g}_t$ on active rounds. The outer update multiplies $w_{t,k}$ by $1 + \alpha_k r_{t,k}$ on these rounds and leaves it unchanged otherwise. Since $|\alpha_k r_{t,k}| \le 1/2$, the inequality $\log(1+z) \ge z - z^2$ gives

$$\begin{aligned}\log \frac{w_{T+1,k}}{w_{1,k}} &= \sum_{t:G_k(X_t)=1} \log(1 + \alpha_k r_{t,k}) \\ &\ge \alpha_k \sum_{t:G_k(X_t)=1} r_{t,k} - \alpha_k^2 \sum_{t:G_k(X_t)=1} r_{t,k}^2 \\ &\ge \alpha_k \sum_{t:G_k(X_t)=1} r_{t,k} - \alpha_k^2 \sum_{t:G_k(X_t)=1} \sum_a \pi_{t,k}(a)\widehat{g}_{t,a}^2 \\ &= \alpha_k \sum_{t:G_k(X_t)=1} r_{t,k} - \alpha_k^2 V_{T,k}.\end{aligned}$$

The second inequality is Jensen's inequality. Combining this with $\log(w_{T+1,k}/w_{1,k}) \le \log K$ proves (22).

E.3 The inner one-step inequality

Fix $q \in \mathcal{B}_\tau$ and let $D_t := \mathrm{KL}(q\|\pi_{t,k})$. On an active round, write $Z_t := \sum_a \pi_{t,k}(a)e^{\eta_k \widehat{g}_{t,a}}$, so the unprojected update in (19) is $\pi_{t+1/2,k}(a) = \pi_{t,k}(a)e^{\eta_k \widehat{g}_{t,a}}/Z_t$. Since $|\eta_k \widehat{g}_{t,a}| \le 1/2$, $e^z \le 1 + z + z^2$ and $\log u \le u - 1$ imply

$$\log Z_t \le \eta_k \pi_{t,k}^\top \widehat{g}_t + \eta_k^2 \sum_a \pi_{t,k}(a)\widehat{g}_{t,a}^2.$$

The KL projection inequality in Proposition 8 gives

$$\begin{aligned} D_{t+1} &\le \mathrm{KL}(q\|\pi_{t+1/2,k}) = D_t - \eta_k q^\top \widehat{g}_t + \log Z_t \\ &\le D_t - \eta_k (q-\pi_{t,k})^\top \widehat{g}_t + \eta_k^2 \sum_a \pi_{t,k}(a)\widehat{g}_{t,a}^2. \end{aligned}$$

Rearranging gives the one-step inequality

$$(q-\pi_{t,k})^\top \widehat{g}_t \le \frac{D_t - D_{t+1}}{\eta_k} + \eta_k \sum_a \pi_{t,k}(a)\widehat{g}_{t,a}^2. \tag{36}$$

Inactive rounds leave D_t unchanged, so summing over active rounds yields

$$\sum_{t:G_k(X_t)=1} (q-\pi_{t,k})^\top \widehat{g}_t \le \frac{D_1 - D_{T+1}}{\eta_k} + \eta_k V_{T,k} \le \frac{\tau}{\eta_k} + \eta_k V_{T,k},$$

using $D_1 = \mathrm{KL}(q\|\pi_{\mathrm{ref}}) \le \tau$ and $D_{T+1} \ge 0$.

E.4 Detailed certificate calculation

On the event of Lemma 11, substitute the deterministic regret bound and both concentration bounds into (26). For every k and $q \in \mathcal{B}_\tau$, this gives

$$\begin{aligned} Tp_k \mathbb{E}_k[q^\top M(X)\pi_{\mathrm{GLHF}}(X)] \le R\left[\frac{\log K}{\alpha_k} + \frac{\tau}{\eta_k} + (\alpha_k+\eta_k)V_k^{\mathrm{ub}}\right] \\ + 2\sqrt{RTp_k\left(\tau + \log\frac{2K}{\delta}\right)} + \frac{8}{3}R\left(\tau + \log\frac{2K}{\delta}\right). \end{aligned} \tag{37}$$

The bracket bounds the outer and inner regret terms in Lemma 6; the remaining terms come from concentration. Appendix E.5 substitutes the fixed rates and bounds the display by $6\sqrt{HRTp_k} + 6RH$. Dividing by Tp_k gives

$$\mathbb{E}_k[q^\top M(X)\pi_{\mathrm{GLHF}}(X)] \le 6\sqrt{\frac{RH}{Tp_k}} + 6\frac{RH}{Tp_k} = \varepsilon_{T,k}.$$

The event is uniform in q, so taking the supremum gives (9).

E.5 The fixed-rate calculation

For $c, V > 0$ and $h = \min\{1/2, \sqrt{c/V}\}$,

$$\frac{c}{h} + hV \le 2\sqrt{cV} + 2c. \tag{38}$$

Apply (38) with $V = V_k^{\mathrm{ub}}$ in both cases: first with $c = \log K$, $h = \alpha_k$, and then with $c = \tau$, $h = \eta_k$. Since $V_{T,k} \le V_k^{\mathrm{ub}}$ on the event in Lemma 11, the optimization bracket in (37), after multiplication by R, is at most

$$\begin{aligned} &2R\sqrt{V_k^{\mathrm{ub}}}\left(\sqrt{\log K} + \sqrt{\tau}\right) + 2R(\tau + \log K) \\ &\le 2R\sqrt{2(\tau+\log K)\left(\frac{2Tp_k}{R} + \log\frac{2K}{\delta}\right)} + 2R(\tau+\log K) \\ &\le 4\sqrt{(\tau+\log K)RTp_k} + R\left[3(\tau+\log K) + 2\log\frac{2K}{\delta}\right]. \end{aligned}$$

The last step uses

$$2\sqrt{2(\tau+\log K)\log(2K/\delta)} \le \tau + \log K + 2\log(2K/\delta).$$

Because both $\tau + \log K$ and $\tau + \log(2K/\delta)$ are at most H, the two square-root terms in (37) sum to at most $6\sqrt{HRTp_k}$. The remaining terms obey

$$3\log K + \frac{17}{3}\tau + \frac{14}{3}\log\frac{2K}{\delta} \le 6H.$$

This proves the simplification used in Theorem 3.

F WELFARE REDUCTIONS AND THE DISTORTION IMPOSSIBILITY

F.1 THE INHERITED LINK INEQUALITY

The pointwise inequality (10) is the signed-margin form of GHY's linearization lemma (Gölz et al., 2025, Lemma 1 and Appendix B). We also record their chord and tangent bounds for use below: for $d \in [0, 1]$,

$$2\beta\ell_\beta d \le 2\sigma(\beta d) - 1 \le 2\beta L d. \tag{39}$$

F.2 PROOF FOR PREFERENCE-ALIGNED GROUPS

Proof of Proposition 5. Draw (X, U) from the group-conditional distribution, then draw $A \sim \pi(X)$ and $B \sim q_k^\star$, independently of each other and of U given X. Write $D := U(A) - U(B) = D^+ - D^-$, where $D^+ := \max\{D, 0\}$ and $D^- := \max\{-D, 0\}$. Its expectation is $\mathbb{E}_k[D] = W_k(\pi) - W_k^\star$.

Let $f(d) := 2\sigma(\beta d) - 1$. Since f is odd and $f(0) = 0$, we have $f(D) = f(D^+) - f(D^-)$. The inherited tangent and chord bounds in (39) give $f(D^+) \le 2\beta L D^+$ and $f(D^-) \ge 2\beta\ell_\beta D^-$. Since $q_k^\star \in \mathcal{B}_\tau$, the antecedent gives $\mathbb{E}_k[(q_k^\star)^\top M(X)\pi(X)] \le \mathrm{Expl}_k(\pi) \le \varepsilon$. Together with antisymmetry, these inequalities imply

$$\begin{aligned}
-\varepsilon &\le \mathbb{E}_k[\pi(X)^\top M(X) q_k^\star] \\
&= \mathbb{E}_k[f(D)] \\
&= \mathbb{E}_k[f(D^+) - f(D^-)] \\
&\le 2\beta L \mathbb{E}_k[D^+] - 2\beta\ell_\beta \mathbb{E}_k[D^-] \\
&= 2\beta\ell_\beta(W_k(\pi) - W_k^\star) + 2\beta(L - \ell_\beta)\mathbb{E}_k[D^+].
\end{aligned}$$

The positive utility difference between A and B cannot exceed B's shortfall from the user's best alternative. Therefore,

$$\begin{aligned}
\mathbb{E}_k[D^+] &= \mathbb{E}_k[\max\{U(A) - U(B), 0\}] \\
&\le \mathbb{E}_k[\max_a U(a) - U(B)] \\
&= \mathbb{E}_k[\max_a U(a)] - W_k^\star = \mathrm{Res}_k.
\end{aligned}$$

Substituting this bound, using $L \ge \ell_\beta$, and rearranging yields

$$W_k(\pi) \ge W_k^\star - \frac{\varepsilon}{2\beta\ell_\beta} - \left(\frac{L}{\ell_\beta} - 1\right)\mathrm{Res}_k. \qquad \square$$

For fixed $\beta > 0$ and $W_k^\star > 0$, as the two error terms in (14) vanish relative to $W_k^\star$, the distortion of π_{GLHF} tends to one.

F.3 FROM GROUP DISTORTION TO POPULATION DISTORTION

For this subsection, write $W(\pi) := \mathbb{E}[\pi(X)^\top U]$ for population welfare and $W^\star := \max_{q \in \mathcal{B}_\tau} W(q)$ for its best fixed feasible lottery benchmark. When $W(\pi) > 0$, define population distortion by $\mathrm{Dist}_{\mathrm{pop}}(\pi) := W^\star / W(\pi)$.

Group distortion bounds need not hold at the same factor for the population when groups overlap, even if they cover every context, none is universal, and every group has zero residual loss. For example, take three equally likely contexts with identical utilities $(1, 0)$ for alternatives A, B, a uniform reference, and $\tau = \log 2$, so both pure actions are feasible. Let group 1 contain contexts $\{1, 2\}$ and group 2 contain contexts $\{2, 3\}$. A policy serving A only at context 2 and B elsewhere has

$$\begin{aligned}
W_1(\pi) = W_2(\pi) &= \tfrac{1}{2}, & W(\pi) &= \tfrac{1}{3}, \\
\mathrm{Dist}_1(\pi) = \mathrm{Dist}_2(\pi) &= 2, & \mathrm{Dist}_{\mathrm{pop}}(\pi) &= 3.
\end{aligned}$$

Every benchmark welfare is one and both residual losses are zero: the well-served intersection contributes to both group averages. Thus a group distortion bound of two does not imply the same bound for the population.

However, a group distortion bound of one does transfer when every group has zero residual loss. This holds for overlapping groups that cover the population even if none is universal. Indeed, let $W_{\mathrm{ind}} := \mathbb{E}[\max_a U(a)]$ and define the nonnegative individual shortfall $Z := \max_a U(a) - \pi(X)^\top U$. Since every user belongs to at least one group and $\mathrm{Res}_k = 0$ for every k, group distortion at most one gives $W_k(\pi) \geq W_k^\star$; groups with zero benchmark welfare satisfy this inequality automatically. Therefore,

$$\begin{aligned} 0 \leq W_{\mathrm{ind}} - W(\pi) &= \mathbb{E}[Z] \\ &\leq \mathbb{E}\left[\sum_k G_k(X) Z\right] \\ &= \sum_k p_k \mathbb{E}_k[Z] \\ &= \sum_k p_k \big(W_k^\star - W_k(\pi)\big) \leq 0. \end{aligned}$$

Thus $W(\pi) = W_{\mathrm{ind}} \geq W^\star$, so population distortion is at most one whenever $W^\star > 0$. Overlap can count a positive shortfall more than once, but a sum of zero shortfalls remains zero.

For disjoint groups, any common distortion bound transfers without the zero-residual condition. If $W_k(\pi) \geq W_k^\star / D$ for every group and $q^\star \in \arg\max_{q \in \mathcal{B}_\tau} W(q)$, then

$$W(\pi) = \sum_k p_k W_k(\pi) \geq \frac{1}{D} \sum_k p_k W_k^\star \geq \frac{1}{D} \sum_k p_k W_k(q^\star) = \frac{W^\star}{D}.$$

F.4 Adapting the GHY impossibility to contextual deployment

Theorem 12 (Worst-case optimality; specialization of GHY)**.** *Fix $\beta > 0$. For every $c > \ell_\beta / L$, there are a response-set size m, a uniform reference, and a nonbinding radius $\tau = \log m$ such that no algorithm can guarantee*

$$\mathbb{E}[W_k(\widehat{\pi})] \geq c W_k^\star \tag{40}$$

on every instance for any specified positive-mass group k. This holds at every sample size, with adaptive pair selection, randomized contextual output, uninformative contexts, and one comparison from each independently sampled user. The expectation is over training and algorithmic randomness.

Proof of Theorem 12. Take the one-comparison instances from GHY's proof of Theorem 3 (Gölz et al., 2025, Appendix E.3). The instances differ only in which alternative a° is welfare-optimal, and every pairwise comparison has probability exactly $1/2$ in each of them. With a uniform reference and $\tau = \log m$, every lottery is feasible, and GHY show that any lottery placing mass at most $1/m$ on a° attains a welfare fraction below c, for m large enough given c and β.

Give every instance the same context distribution, drawn independently of the user's utility. Contexts then carry no information about a°: each label is a fair coin conditional on the history, context, and queried pair. Hence the transcript and the output policy $\widehat{\pi}$ have the same law in every instance, at every sample size and under adaptive pair selection.

Fix a positive-mass group k and let $\bar{\pi}_k := \mathbb{E}[\mathbb{E}_k[\widehat{\pi}(X)]]$ be the output averaged over the group's contexts, the training data, and the learner's randomness. By independence of contexts and utilities, $\mathbb{E}[W_k(\widehat{\pi})]$ is the welfare of the fixed lottery $\bar{\pi}_k$, and $W_k^\star = \max_{q \in \Delta(\mathcal{A})} \mathbb{E}[q^\top U]$. Since $\bar{\pi}_k$ does not depend on the instance, some alternative receives mass at most $1/m$ under it, and the instance with that alternative as a° gives $\mathbb{E}[W_k(\widehat{\pi})] < c W_k^\star$. The instances are fixed across horizons, so for a sequence of algorithms one instance violates the fraction at infinitely many horizons, ruling out a limiting-inferior guarantee of c as well. □

The one-comparison condition is substantive. Repeated answers from the same utility draw can expose correlations that are absent from a single fresh-user observation. GHY obtain extensions to multiple comparisons under an additional Condorcet-loser condition; our impossibility statement uses only their one-comparison setting.

For $\tau > 0$, setting $\delta_T = 1/T$ makes $\varepsilon_{T,k} \to 0$ for each fixed instance and group k. Proposition 4 then gives, for GLHF's output π_T, $\liminf_{T\to\infty} \mathbb{E}[W_k(\pi_T)] \geq (\ell_\beta/L)W_k^\star$, matching the lower bound above; at $\tau = 0$, the reference is the only feasible policy and attains every group benchmark.

G Experimental details and additional results

Our primary comparison is multigroup GLHF against NLHF, which serves everyone the same lottery. Both studies use ratings to generate training comparisons and measure held-out welfare. After describing the data and evaluation procedure, we report the main comparisons, active sampling, and sensitivity checks. Implementation and collection details appear last.

G.1 Data and group selection

Coffee ratings. The Great American Coffee Taste Test (TidyTuesday, 2024) contains ratings of four coffees on a 1–5 scale. We retain the 3,744 respondents who rated all four and set $U = (r-1)/4$, treating equal rating increments as equal utility increments. We reserve a random discovery sample of 375 respondents for choosing groups, leaving 3,369 for analysis.

Synthetic language ratings. The completion-rating study uses six fixed answers to *Why is the sky blue?*: four correct answers vary in technical detail and writing style, and two are incorrect controls. We construct sixteen personas using binary characteristics: age group, education, technical background, and preferred writing style. For each persona and candidate completion, 104 independent instances of Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) were asked to assume the persona and provide a rating of the completion on a 1–10 scale. This gives 9,984 completion ratings, which we rescaled as $U = (r-1)/9$. Within each persona, we independently shuffled the ratings for each answer, then combined one rating for each of the six answers into a synthetic user's utility vector. This produced 104 synthetic users per persona and 1,664 in total. Each collected rating was used once. We reserve 167 synthetic users for group selection (10 or 11 per persona), leaving 1,497 for analysis. Appendix G.7.4 specifies the synthetic data collection procedure.

Group definitions. Each study includes the universal group (everyone in the analysis sample) and five subgroups that may overlap. Tables 1 and 2 give definitions and analysis-sample counts. Coffee respondents who did not report a demographic attribute remain in the universal group and are not included in subgroups defined by that attribute.

Group	n	Definition
Universal	3,369	Every analysis respondent.
Age 45+	483	Reported age 45–54, 55–64, or over 65.
Female	759	Reported gender is female.
Expertise 1–4	815	Self-rated coffee expertise is 1, 2, 3, or 4.
Children 1+	767	Reported one, two, three, or more than three children.
Works from home	1,424	Selected "I primarily work from home."

Table 1: Groups and membership counts in the coffee-rating analysis sample.

Group	n	Definition
Universal	1,497	Every analysis pseudo-respondent.
Not technical	750	Persona specifies a nontechnical background.
Expressive wording	747	Persona prefers expressive writing.
Direct wording	750	Persona prefers direct writing.
Very technical	747	Persona specifies a very technical background.
University education	748	Persona specifies university education.

Table 2: Groups and analysis-sample counts in the completion-rating study.

Selecting groups for the coffee-rating study. We chose the five subgroups in Table 1 by judgment, using preference correlations in the discovery sample to guide selection. For each of the six coffee pairs, we computed the fraction preferring the first coffee minus the fraction preferring the second, with ties contributing zero. We calculated these six margins within each candidate group and within the full discovery sample, then compared the two vectors using Pearson correlation. The selected groups included three with negative correlations (Age 45+, Female, and Expertise 1–4) and two with positive correlations (Children 1+ and Works from home). We fixed these groups before training and evaluation.

Selecting groups for the completion-rating study. We considered eight candidate subgroups, corresponding to the two levels of each persona attribute. For each of the 15 answer pairs, we calculated the average rescaled rating of the first answer minus the average rescaled rating of the second. We calculated these differences within each candidate group and within the full discovery sample, then compared them using Pearson correlation. Lower correlations indicate that the group's preferences across answer pairs are less aligned with those of the full discovery sample.

Candidate groups must contain at least 30 discovery rows and have a defined correlation. We select the four lowest-correlation candidates and the lower median of the remaining candidates ordered by correlation. The fifth group adds a more typical candidate to the four least aligned groups. Table 3 reports the selection, run once on the discovery sample. We fixed these groups before training and evaluation.

Persona attribute	Discovery n	Rating-difference correlation	Selected
Not technical	82	-0.115	yes
Expressive wording	85	0.511	yes
Direct wording	82	0.643	yes
Very technical	85	0.777	yes
High school or less	83	0.983	no
University education	84	0.988	yes
Age 22–32	82	0.996	no
Age 62–72	85	0.996	no

Table 3: Group selection in the completion-rating study on the 167 reserved discovery rows. Correlations compare each candidate's 15 mean-rating differences with those of the full discovery sample. The universal group was included separately.

G.2 Experimental design and evaluation

Methods and main setting. Table 4 summarizes the four methods. Multigroup GLHF uses memberships both to update group specialists and to choose the lottery served to a user. Batch NLHF is the primary population-wide comparator. Pooled GLHF is the same implementation as multigroup GLHF with only the universal group. The uniform reference shows how learning compares to serving the starting lottery.

Method	Deployed output	Comparisons at main setting
Multigroup GLHF	Lottery depending on group memberships	Active
Batch NLHF	One learned lottery for everyone	Passive
Pooled GLHF	One learned lottery for everyone	Active
Uniform reference	Equal probability for every alternative	None

Table 4: Methods at the main setting. Each learner receives one comparison per training user. Passive versions of both GLHF methods are also evaluated. Batch NLHF always uses passive comparisons.

The reference lottery is uniform over the four coffees or six answers. The learner radius τ_L limits a served lottery's KL divergence from this reference; the evaluation radius τ_E constrains the fixed-lottery welfare benchmark. Both radii are 0.5 at the main setting. The sensitivity checks in Appendix G.5 vary τ_L while keeping the benchmark fixed.

Generating training comparisons. Each training respondent supplies exactly one comparison per fit. Passive sampling draws an unordered pair uniformly. Active sampling draws the first alternative from the learner's current policy for that respondent and the second uniformly from the remaining alternatives. At the main setting, $\beta = \infty$, the higher-rated alternative wins and a seeded fair coin resolves ties. At finite β, the probability that a wins against b is $\sigma(\beta[U(a) - U(b)])$, as in Assumption 2. Smaller β makes labels less reliably follow rating order. The learner observes the queried pair and winner, together with group memberships, but not the ratings used to generate the label.

Within each training split and learner seed, methods share training respondents and their order. Passive multigroup GLHF, passive pooled GLHF, and NLHF share the same comparison log. Active multigroup and pooled GLHF each generate their own policy-dependent log, so equal comparison budgets do not imply identical queried pairs.

Fitting and deploying policies. Multigroup GLHF makes one pass through the training users, updating the specialists for every group containing the current user. It saves the policy before each comparison. To serve a new user, it averages the lotteries those saved policies assign to that user's memberships. Every group evaluates this same mapping. Pooled GLHF averages a single specialist's lotteries. The experiments use adaptive learning rates; Theorem 3 analyzes fixed rates and does not directly cover this variant. Appendix G.7.1 specifies the updates.

Batch NLHF estimates each pair's win probability from the complete passive comparison log and fits a constrained maximal lottery to the estimated margin matrix game. It uses no additional comparisons or held-out ratings. Appendix G.7.3 gives the estimator and numerical checks.

G.2.1 Held-out welfare

Discovery users are excluded from all policy fitting and evaluation. We randomly shuffle the remaining users and divide them into ten nearly equal folds. We train on nine folds and test on the tenth, rotating through all ten test folds. We repeat this procedure with ten different random partitions, giving 100 train/test splits. Each split uses 50 learner seeds to vary the randomized training procedure.

For each held-out group, empirical distortion divides the welfare of its best fixed feasible lottery by the welfare of the deployed policy. For group k in split f, let I_{kf} be its test respondents, $n_{kf} = |I_{kf}|$, and $\bar{U}_{kf}$ their mean utility vector. The benchmark welfare is

$$\widehat{W}^*_{kf} = \max_{q:\, \mathrm{KL}(q\|\pi_{\mathrm{ref}})\le\tau_E} q^\top \bar{U}_{kf}. \tag{41}$$

For method j and learner seed s, define

$$\widehat{\mathrm{Dist}}^{(j)}_{kfs} = \frac{\widehat{W}^*_{kf}}{n_{kf}^{-1}\sum_{i\in I_{kf}} U_i^\top \pi^{(j)}_{fs}(X_i)}, \qquad \widehat{\mathrm{Dist}}^{(j)}_{fs} = \max_k \widehat{\mathrm{Dist}}^{(j)}_{kfs}. \tag{42}$$

The benchmark is optimized on the test fold using full ratings; it is not another trained method. This optimization can make benchmark welfare optimistic relative to the population optimum on average. Held-out utilities never influence policy training, and the reported ratios describe welfare on these test folds.

G.2.2 Averaging and uncertainty

For worst-group results, we take the maximum across groups within each split and seed, average the 50 seeds within each split, and then average the 100 splits. Thus the reported worst-group value is not the maximum of the displayed per-group means. For method comparisons, we first subtract the two methods' distortions within each split and seed, then follow the same averaging procedure. Positive NLHF-minus-multigroup, pooled-minus-multigroup, and reference-minus-multigroup differences favor multigroup GLHF; positive passive-minus-active differences favor active sampling.

The train/test splits overlap, so neither the 100 split scores nor the 5,000 seed–split values are independent replicates. We use the Nadeau–Bengio correction (Nadeau & Bengio, 2003) to form

approximate 95% t intervals from the split scores, with 99 degrees of freedom and standard error

$$\widehat{\mathrm{SE}} = \sqrt{\left(\frac{1}{100} + \overline{n_{\mathrm{test}}/n_{\mathrm{train}}}\right) S^2_{\mathrm{split}}}.$$

Here S^2_{split} is the sample variance of the 100 seed-averaged scores, and the ratio is averaged over the actual split sizes. For a method comparison, the scores are paired differences. Overlap between two methods' individual intervals is not a test of their paired difference. The correction is approximate and does not establish exact coverage for the nonlinear worst-group ratio. We also use two-sided tests for the per-group NLHF-minus-multigroup differences, with Holm adjustment (Holm, 1979) across the six groups in each study.

G.3 Results at the main setting

At $\beta = \infty$ and $\tau_L = \tau_E = 0.5$, multigroup GLHF has lower mean worst-group distortion than NLHF in both studies: 1.08 versus 1.28 in the coffee-rating study and 1.07 versus 1.36 in the completion-rating study. Table 5 gives the precise levels and paired differences. Both GLHF methods use active sampling here; NLHF remains passive-trained.

Study	MG	Batch NLHF	Pooled	Ref.	Batch NLHF − MG	Pooled − MG	Ref. − MG
Coffee-rating					0.2021	0.1732	0.0421
study	1.0795	1.2816	1.2528	1.1217	$[0.1507, 0.2535]$	$[0.1293, 0.2171]$	$[0.0323, 0.0519]$
Completion-rating					0.2951	0.2807	0.6662
study	1.0695	1.3646	1.3501	1.7357	$[0.2823, 0.3080]$	$[0.2700, 0.2913]$	$[0.6540, 0.6784]$

Table 5: Worst-group distortion at $\beta = \infty$ and $\tau_L = \tau_E = 0.5$ (lower is better). GLHF uses active sampling; batch NLHF uses passive sampling. MG denotes multigroup GLHF and Ref. the uniform reference. Method columns are means; difference columns give paired estimates and corrected 95% intervals over 100 splits, each averaging 50 learner seeds.

Groupwise welfare. In the coffee-rating study, multigroup GLHF has lower point distortion than NLHF and pooled GLHF in every selected group. The largest subgroup gaps are for respondents aged 45+, those with coffee expertise 1–4, and female respondents. It also has the lowest universal-group point distortion: 0.992, compared with 1.005 for NLHF and 1.012 for pooled GLHF (Figure 1). The figure includes the untrained uniform lottery as a reference.

In the completion-rating study, multigroup GLHF has lower point distortion than NLHF and pooled GLHF in every selected group (Figure 2). It exceeds the fixed-lottery benchmark for expressive wording, university education, and the universal group. These values below one reflect its ability to vary the served lottery with memberships.

Paired comparisons within groups. Figure 3 compares multigroup GLHF with NLHF on each group. Multigroup distortion is lower in all six groups in each study, with every two-sided test significant at the 0.05 level after Holm adjustment. This conclusion concerns NLHF at the main setting; the group-bar point estimates alone do not establish significance against every comparator.

G.4 Active and passive sampling

Active sampling gives a smaller additional improvement than the comparison of multigroup GLHF with NLHF. At $\beta = \infty$ and radius 0.5, multigroup worst-group distortion falls from 1.0913 under passive sampling to 1.0795 under active sampling in the coffee-rating study, and from 1.0986 to 1.0695 in the completion-rating study. The paired reductions are 0.0118 (corrected 95% interval $[0.0071, 0.0165]$) and 0.0291 ($[0.0255, 0.0327]$), respectively.

The implemented protocols differ in both pair selection and score normalization. In (16), active sampling uses $R = m-1$ and passive sampling uses $R = \binom{m}{2}$, where m is the number of alternatives. The comparison measures the effect of these complete protocols, rather than isolating pair selection.

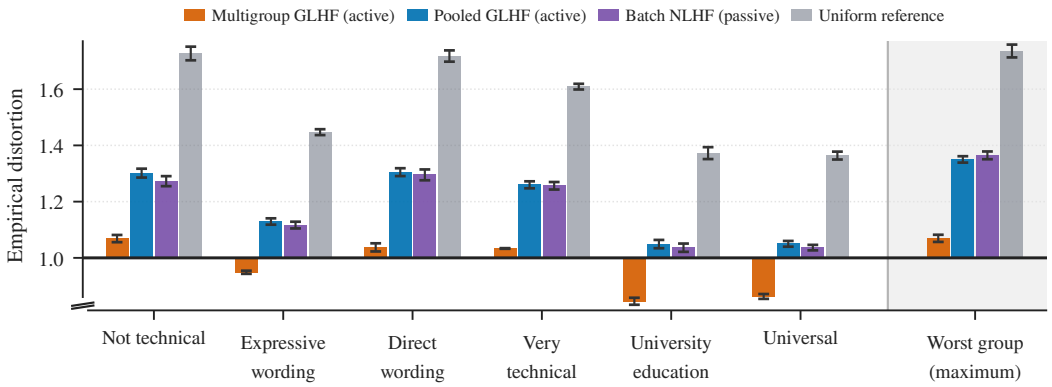

Figure 2: Empirical distortion in the completion-rating study at $\beta = \infty$ and $\tau_L = \tau_E = 0.5$: active multigroup GLHF, active pooled GLHF, passive batch NLHF, and the uniform reference. Lower is better. Bars below one indicate welfare above the group's best fixed feasible lottery. The rightmost cluster averages the maximum across groups within each run. Whiskers are corrected 95% confidence intervals for individual means.

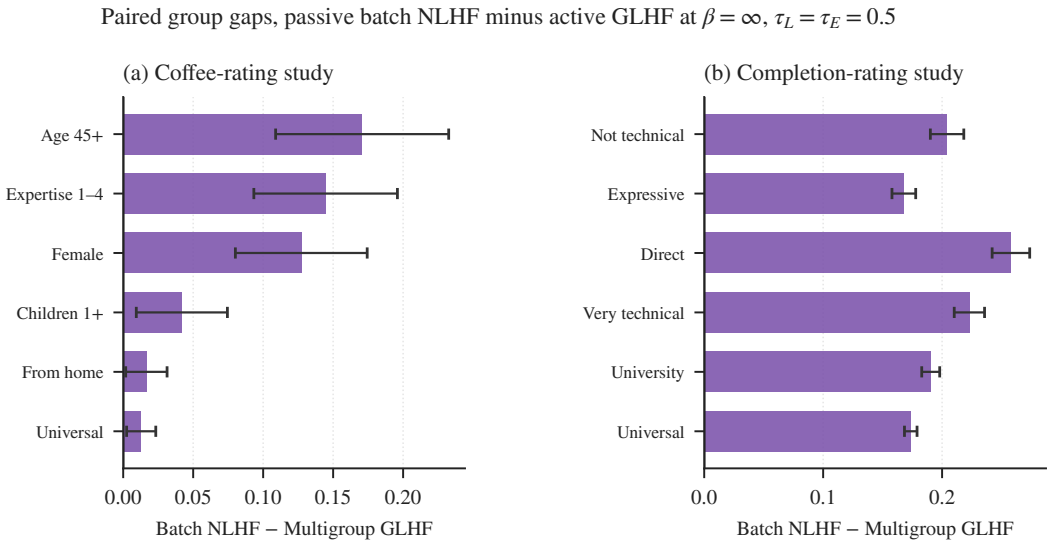

Figure 3: Per-group distortion differences: passive batch NLHF minus active multigroup GLHF at $\beta = \infty$, $\tau_L = \tau_E = 0.5$. Positive values favor multigroup GLHF. Corrected 95% intervals apply to each difference separately, not to all differences jointly. The accompanying two-sided tests use Holm adjustment over the six groups within each study.

Figure 4 evaluates the average of the first t saved policies as comparisons accumulate, using the same test fold and benchmark throughout. Learning rates retain the original planned number of comparisons; they are not retuned for each checkpoint. The zero-comparison policy is uniform. Group maxima are taken within each seed before averaging.

Early checkpoints in the coffee-rating study do not consistently favor active sampling. In the completion-rating study, the active checkpoint at 674 comparisons has mean distortion 1.0976, close to passive GLHF's final 1.0986. The plotted intervals apply to individual checkpoints and do not establish a uniform advantage along either curve.

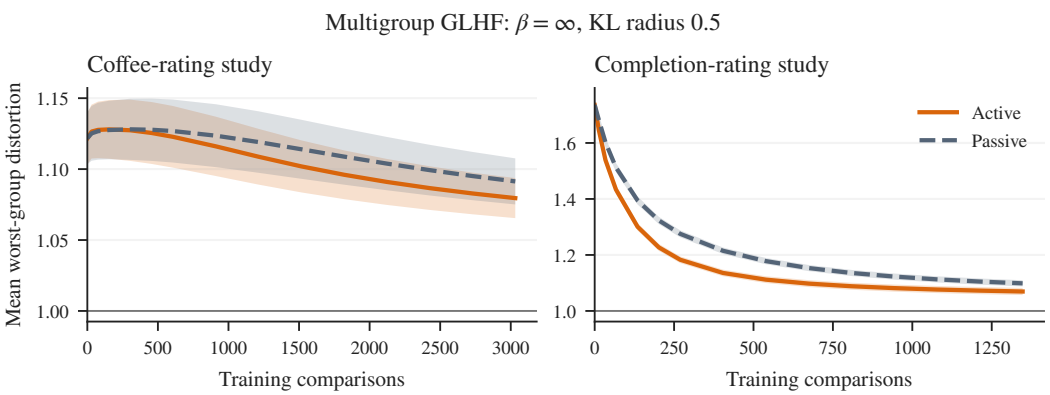

Figure 4: Active and passive multigroup GLHF at $\beta = \infty$ and $\tau_L = \tau_E = 0.5$. Checkpoints average the first t policies of one full-length adaptive run, each saved before its comparison. The line at one marks the fixed-lottery benchmark. Shading gives corrected 95% intervals at individual checkpoints. Intervals for the completion-rating study are narrow at this scale; vertical scales differ.

G.5 Sensitivity checks

Learner radius. A larger τ_L allows the learner to move farther from the uniform reference. Figures 5 and 6 vary τ_L over $\{0.05, 0.2, 0.35, 0.5, 0.75, 1\}$ while keeping the evaluation benchmark fixed at $\tau_E = 0.5$ and feedback at $\beta = \infty$. Both GLHF methods use active sampling; NLHF remains passive-trained.

Multigroup GLHF has the lowest mean worst-group distortion among the trained methods at every tested radius in both studies. In the coffee-rating study, its distortion decreases from 1.110 to 1.072 as the radius grows, while pooled GLHF and NLHF worsen. For NLHF, universal-group distortion falls with the larger radius, but distortion rises for respondents aged 45+, with coffee expertise 1–4, and female respondents; these groups' pairwise preference margins were negatively correlated with the population margins in the discovery sample. This pattern is consistent with a single lottery moving toward coffees favored by the population overall while serving groups with less aligned preferences less well. In the completion-rating study, multigroup distortion falls from 1.452 to 0.953. Because the benchmark radius stays fixed, these curves show how additional training flexibility changes welfare against the same target. When $\tau_L > \tau_E$, distortion below one can also reflect the learner's larger feasible set.

Comparison noise. Figure 7 varies $\beta \in \{1/2, 1, 2, 4, \infty\}$ at $\tau_L = \tau_E = 0.5$. Passive GLHF and NLHF cover the full grid; active GLHF is evaluated only at 4 and ∞. Multigroup GLHF has lower mean worst-group distortion than the trained population-wide comparators in every tested setting. This does not extend to the uniform reference: in the coffee-rating study at $\beta = 1/2$ and 1, passive multigroup's point estimates are slightly worse than uniform's.

From the smallest β to ∞, passive multigroup distortion decreases from 1.125 to 1.091 in the coffee-rating study and from 1.530 to 1.099 in the completion-rating study. At $\beta = 4$, active sampling also improves on passive multigroup, by 0.0113 and 0.0431, respectively. Less label noise need not always improve performance: the active completion-rating estimate is 1.0595 at $\beta = 4$, slightly below its 1.0695 at ∞.

Constructing synthetic respondents. We repeated the completion-rating analysis with four additional shuffles of the same ratings, using passive sampling at $\beta = \infty$ and $\tau_L = \tau_E = 0.5$. We repeated group selection for each shuffle and selected the same five subgroups every time. The pooled-minus-multigroup differences in worst-group distortion ranged from 0.2747 to 0.2786, close to 0.2760 for the original construction. The active results use only the original construction.

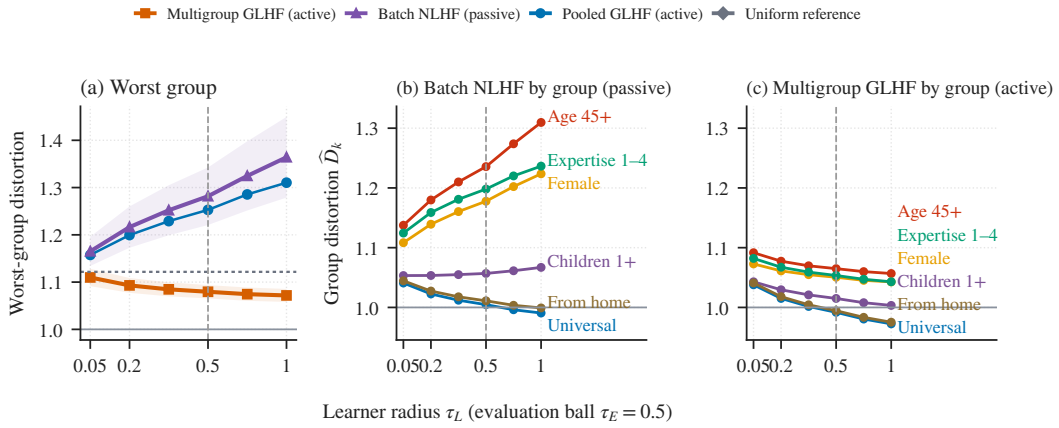

Figure 5: Distortion in the coffee-rating study as the learner radius varies, with $\beta = \infty$ and evaluation radius fixed at $\tau_E = 0.5$. Both GLHF methods use active sampling; batch NLHF uses passive sampling. Panel (a) shows worst-group distortion for all four methods; panels (b) and (c) show groupwise batch NLHF and multigroup GLHF distortion. The dashed line marks the main learner radius, 0.5. Bands in (a) are corrected 95% intervals for individual means; groupwise curves show point estimates.

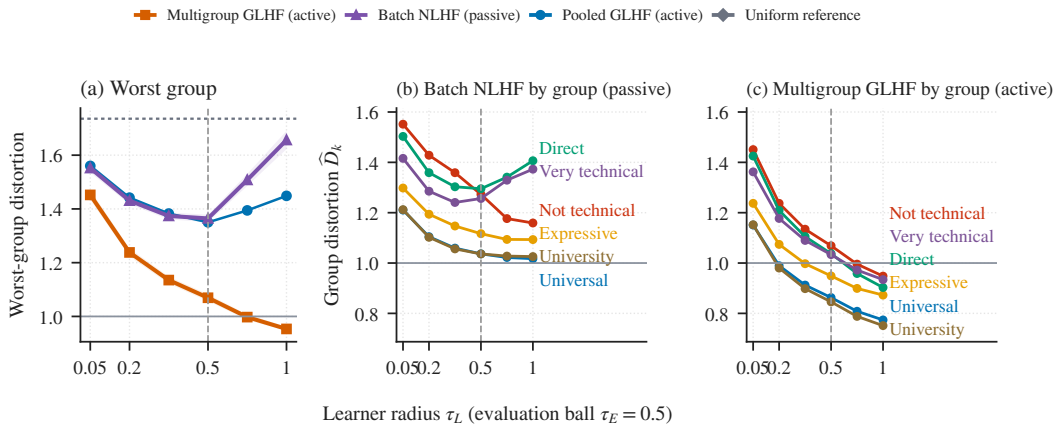

Figure 6: Distortion in the completion-rating study as the learner radius varies, with $\beta = \infty$ and evaluation radius fixed at $\tau_E = 0.5$. Both GLHF methods use active sampling; batch NLHF uses passive sampling. Panel (a) shows worst-group distortion for all four methods; panels (b) and (c) show groupwise batch NLHF and multigroup GLHF distortion. Bands in (a) are corrected 95% intervals for individual means; groupwise curves show point estimates.

G.6 Limitations

These studies evaluate welfare under specific rating-based utility models and selected group definitions. Both assume that equal rating increments represent equal, interpersonally comparable utility increments; the reported distortion also depends on assigning zero utility to the lowest rating. The coffee study captures heterogeneous preferences over shared alternatives, but choosing among four coffees does not test the complexity of LLM responses or variation across prompts. The completion study uses one model to rate six answers to one question, with separately elicited ratings combined into synthetic respondents; it does not measure human preferences. Groups were chosen partly to capture disagreement with population preferences, and the confidence intervals condition on these choices rather than accounting for variability in group selection. The baseline comparisons therefore demonstrate gains from conditioning on these memberships, but do not establish an advantage over other group-aware learning methods.

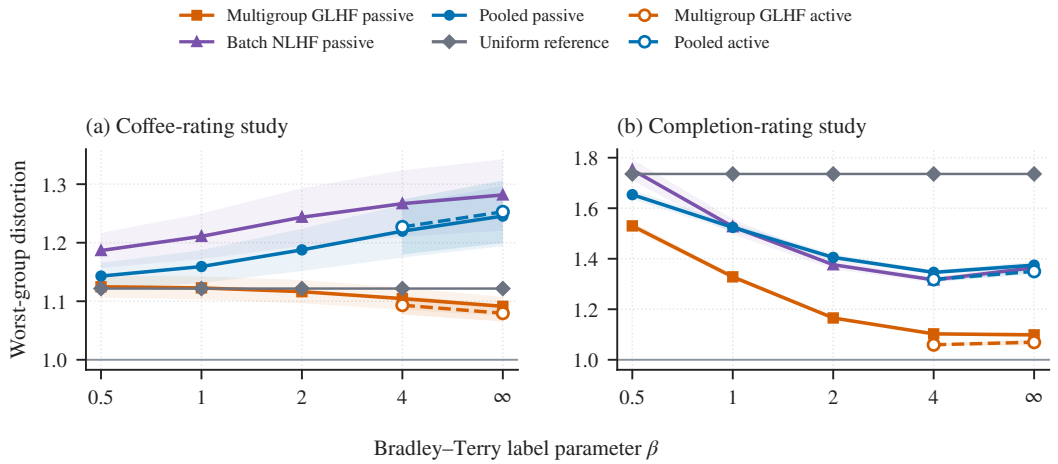

Figure 7: Sensitivity to comparison noise in the coffee-rating (left) and completion-rating (right) studies, with $\tau_L = \tau_E = 0.5$. Larger β makes comparisons more reliably follow rating order; at ∞, only ties are randomized. Passive multigroup GLHF, pooled GLHF, and batch NLHF cover the full grid; active GLHF covers 4 and ∞. Bands are corrected 95% intervals for individual means.

The experiments also differ from the theoretical setting in two respects. The main setting uses deterministic comparisons except at ties ($\beta = \infty$), where the worst-case welfare factor L/ℓ_β diverges; low empirical distortion in this setting does not validate a finite theoretical welfare bound. The implementation uses adaptive learning rates, which are not directly covered by the fixed-rate comparison certificate. The results thus provide empirical evidence for the implemented method on these studies, without establishing the same performance for the algorithm analyzed by the theorem.

G.7 Implementation and Reproducibility

G.7.1 Adaptive Learning Rates

The implementation chooses rates from observed comparison scores, without using group masses or the confidence level. The updates below specify the adaptive variant; Theorem 3 analyzes fixed rates.

For $\tau_L > 0$, initialize $w_{1,k} = 1/K$ and $\pi_{1,k} = \pi_{\mathrm{ref}}$. Let T be the number of training comparisons, and use the normalized score $\widehat{g}_t$ from (16). Write $r_{t,k} = \pi_{t,k}^\top \widehat{g}_t$. Before round t, the accumulated squared specialist scores and within-lottery score variances are, respectively,

$$Q_{t,k} = \sum_{\substack{s<t \\ G_k(X_s)=1}} r_{s,k}^2, \qquad C_{t,k} = \sum_{\substack{s<t \\ G_k(X_s)=1}} \sum_a \pi_{s,k}(a)\big(\widehat{g}_{s,a} - r_{s,k}\big)^2.$$

With $L_{\mathrm{ad}} = \max\{1, \log[K(1 + \frac{1}{2}\log(T+1))]\}$ and $\bar{\tau} = \min\{\tau_L, \log(1/\min_a \pi_{\mathrm{ref}}(a))\}$, the rates are

$$\alpha_{t,k} = \min\left\{\frac{1}{2}, \sqrt{\frac{L_{\mathrm{ad}}}{2(1+Q_{t,k})}}\right\}, \qquad \eta_{t,k} = \min\left\{1, \sqrt{\frac{\bar{\tau}}{2(1/4+C_{t,k})}}\right\}.$$

Deployment weights are proportional to $\alpha_{t,k} w_{t,k}$ among groups containing the current user. Larger accumulated scores or variances reduce the corresponding learning rate. The empirical inner cap is 1, whereas the theoretical fixed-rate algorithm uses a cap of $1/2$.

After adding round t to the accumulators, write $\gamma_{\alpha,k} = \alpha_{t+1,k}/\alpha_{t,k}$ and $\gamma_{\eta,k} = \eta_{t+1,k}/\eta_{t,k}$. For groups containing the current user, a decrease in the rates mixes the previous log weights with their

initial values:

$$\begin{aligned}\log w_{t+1,k} &= (1-\gamma_{\alpha,k})(-\log K) + \gamma_{\alpha,k}\big[\log w_{t,k} + \log(1+\alpha_{t,k}r_{t,k})\big],\\ z_{t,k}(a) &\propto \exp\big\{\gamma_{\eta,k}\log \pi_{t,k}(a) + (1-\gamma_{\eta,k})\log \pi_{\mathrm{ref}}(a) + \eta_{t+1,k}\widehat{g}_{t,a}\big\},\\ \pi_{t+1,k} &= \operatorname*{arg\,min}_{q:\,\mathrm{KL}(q\|\pi_{\mathrm{ref}})\le\tau_L} \mathrm{KL}(q\|z_{t,k}).\end{aligned}$$

States for other groups remain unchanged. The outer tentative update uses the rate at the current step, while the inner exponential step uses the rate at the next step. Constant rates make both ratios one and recover the fixed-rate recurrence for the same rates.

The outer update follows the adaptive multiple-rate Prod construction (Gaillard et al., 2014). The inner update uses the changing rate shown above and projects back into the learner's KL ball. If $\tau_L = 0$, the output is the reference lottery and no updates are needed.

Computing the KL projection. We compute the projection of $z_{t,k}$ using Appendix C, with $v = z_{t,k}$ and $\tau = \tau_L$. If the candidate lies within the learner's KL ball, we retain it. Otherwise, we use Brent's method to find the point on the geometric path from π_{ref} to $z_{t,k}$ whose KL divergence from π_{ref} equals τ_L.

Pooled implementation. For pooled GLHF, $K = 1$ and the only specialist serves everyone: $\widehat{\pi}_t(x) = \pi_{t,1}$. Score cancellation gives $\pi_{t,1}^\top \widehat{g}_t = 0$, so the outer weight stays at one. Unlike the prescribed fixed-rate schedule, the adaptive schedule is well defined at $K = 1$ because $L_{\mathrm{ad}} \ge 1$.

G.7.2 Computing the welfare benchmark

To compute the benchmark in (41), we maximize $q^\top \bar{U}_{kf}$ over $\mathcal{B}_{\tau_E}$. At zero radius or when all alternatives have equal mean utility, we use π_{ref}. If the reference lottery conditioned on the highest-utility alternatives is feasible, we use that lottery. Otherwise, the KL constraint binds, and the solution has the exponential-tilt form

$$q_\lambda(a) = \frac{\pi_{\mathrm{ref}}(a)\exp(\bar{U}_{kf}(a)/\lambda)}{\sum_b \pi_{\mathrm{ref}}(b)\exp(\bar{U}_{kf}(b)/\lambda)}, \qquad \lambda > 0.$$

Here λ is the multiplier for the KL constraint. We use a one-dimensional root search to choose it so that $\mathrm{KL}(q_\lambda\|\pi_{\mathrm{ref}}) = \tau_E$. Benchmark and deployed welfare are calculated as expectations under their respective lotteries, without sampling an alternative.

G.7.3 Batch NLHF

For each pair (a, b), NLHF estimates $\Pr(a \succ b)$ as the observed fraction of wins for a among comparisons of that pair. An unobserved pair is assigned probability $1/2$ and recorded as unobserved. These estimates define the antisymmetric margin matrix $\widehat{M}_{ab} = 2\widehat{\Pr}(a \succ b) - 1$ for $a \ne b$, with $\widehat{M}_{aa} = 0$. This estimates the population margins $\mathbb{E}[M(X)]$ from the passive comparison log. Using the feasible set from (1) with radius τ_L, we seek a fixed lottery p minimizing the largest advantage of any feasible challenger:

$$\min_{p\in\mathcal{B}_{\tau_L}} \max_{q\in\mathcal{B}_{\tau_L}} q^\top \widehat{M} p.$$

The fitted policy serves this same lottery to every user. Antisymmetry and the common feasible set imply that the optimal value is zero.

For a fixed lottery p, the dual formulation is

$$\max_{q\in\mathcal{B}_{\tau_L}} q^\top \widehat{M} p = \inf_{\lambda>0}\left\{\lambda\tau_L + \lambda\log\left[\sum_a \pi_{\mathrm{ref}}(a)\exp\left(\frac{(\widehat{M}p)_a}{\lambda}\right)\right]\right\}.$$

Here λ is the multiplier for the challenger's KL constraint. We minimize the right-hand expression jointly over $p \in \mathcal{B}_{\tau_L}$ and the auxiliary variable $\log\lambda$ using SciPy's SLSQP solver, with finite-difference derivatives, at most 500 iterations, and `ftol`$= 10^{-12}$. Each fit uses one optimization call initialized at $p = \pi_{\mathrm{ref}}$ and $\lambda = \max\{\tau_L, 10^{-3}\}$, without restarts or a secondary equilibrium-selection step.

We normalize the returned lottery and, if needed, project it back into the learner's KL ball. We then check its maximum feasible challenger advantage using the exponential-tilt procedure described in Appendix G.7.2, replacing $\bar{U}_{kf}$ by $\widehat{Mp}$ and τ_E by τ_L. We accept a fit only if this advantage is at most 2×10^{-5}; all reported fits passed. This check is separate from the optimizer's convergence flag and does not concern held-out welfare.

G.7.4 SYNTHETIC COLLECTION MATERIALS

The collection requests specified Claude Haiku 4.5, temperature 1.0, a maximum output length of 64 tokens, and extended thinking disabled. Each request contained a single user message with one persona background, the question, and one answer, with no conversation history or other candidate answers. It requested a JSON object containing one integer score. Background lines appeared in the fixed order age, education, technical background, and preferred writing style. Every combination of the four two-level attributes in Table 6 was used, giving 16 personas.

Feature	First level	Second level
Age	22-32	62-72
Education	Highest completed education is high school or less	Highest completed education is a university degree
Technical background	Not technical; no technical/STEM work, training, or study	Very technical; extensive technical/STEM work, training, or study
Preferred writing style	Prefers direct and literal wording	Enjoys expressive and literary wording

Table 6: The four two-level attributes used to form the 16 persona descriptions.

Prompt template. The following text is reproduced from the collection. The placeholders are replaced with the persona background, the question *Why is the sky blue?*, and exactly one answer below.

> Use the following background as context:
> {background}
> People in these groups can disagree. Do not assume any traits or personal history beyond what is listed.
> You just asked, "{question}" One person gave you the answer below, exactly as written, with no further explanation:
> {answer}
> Rate how well the answer works for you.
> Give one overall score from 1 to 10. A score of 1 means it does not work for you at all; 5 means it partly works but has problems; and 10 means it works extremely well for you.
> Return only a JSON object with one key named "score" and an integer from 1 through 10 as its value. Do not provide an explanation.

Fixed answer set. The first four answers vary accessible versus technical content and direct versus expressive style. The last two are negative controls. The headings below identify each answer's role in the study; they were not shown to the evaluator.

Accessible, direct (correct). Sunlight contains many colors, and the air spreads the blue ones around more than the red ones. That is why the sky looks blue.

Accessible, expressive analogy (correct). Imagine white sunlight as a crowd of colors crossing the air. The red colors mostly keep marching forward, while the blue colors are nudged into wandering paths by tiny molecules. Those blue wanderers reach your eyes from all across the heavens, painting the daytime view blue. It is a simple picture of how air scatters light.

Technical, direct (correct). How strongly air redirects light depends sharply on the light's wavelength: the strength varies approximately with the inverse fourth power, so 450-nanometer blue light

is scattered several times more strongly than 650-nanometer red light. This is the Rayleigh regime, which applies because air molecules are much smaller than visible wavelengths. Scattering along every line of sight creates a blue-weighted diffuse radiance. Violet is less prominent because the solar spectrum, atmospheric absorption, and human visual sensitivity reduce its contribution. At sunset, the longer optical path scatters blue out of the direct beam, leaving red and orange.

Technical, expressive (correct). Air does not redirect every color equally. Its effect rises sharply as wavelength shortens, roughly with the inverse fourth power, so 450-nanometer blue light is scattered several times more strongly than 650-nanometer red light. In this Rayleigh regime, blue photons are swept from the Sun's direct beam and distributed across innumerable lines of sight, building the blue radiance overhead; red photons travel onward more faithfully. Violet scatters still more, yet solar output, atmospheric absorption, and the eye's sensitivity mute it. At sunset, the lengthened path spends the blues and leaves the horizon burning red.

Fictional-story control (incorrect). When the world was new, the sky was clear as glass and the sea alone held color. Each morning, the wind carried a little of the sea upward, spreading its blue across the empty air so travelers could tell day from night. Rain returned some of that color to the water, while evening loosened the rest into red and gold before darkness arrived. That is why the sky and ocean share one blue, continually passed between them by wind and weather.

Plain incorrect control (incorrect). The sky is blue because it reflects the color of the oceans below it. Light bounces off the water and spreads across the sky.

G.7.5 Code and saved results

The supplementary archive, `EmpiricalTesting-submission.zip`, contains the code, both input datasets, and saved evaluation results in an `EmpiricalTesting/` directory. Its `README.md` gives installation and reproduction commands, and `data/README.md` describes the inputs. Paths below are relative to that directory:

- Input datasets: `data/coffee/` and `data/persona/`; the latter also contains the synthetic collection materials.
- Comparison results: `results/coffee/comparison/` and `results/persona/comparison/`; each dataset's `beta-source/` directory contains the β runs.
- Learning-curve reductions: `results/coffee/learning-curves/` and `results/persona/learning-curves/`.
- Synthetic-row shuffles: `results/persona/panel-permutations/`.

Saved comparison results include held-out welfare and distortion for each split, seed, method, and group. The ZIP retains these results and their reduction tables, while omitting per-comparison training logs and per-split learning-curve arrays to reduce its size. The figure scripts can draw the paper figures from the bundled tables.

After installing the dependencies, run `python reproduce.py` from the extracted `EmpiricalTesting/` directory. This reruns both datasets' β/protocol grids, GLHF and NLHF comparisons, learner-radius sweeps, learning curves, and five synthetic-row alignments from the input ratings, then generates the paper figures. New results are written to `results/reproduced/`. The README also gives commands for individual experiments, figures from saved results, and a reduced smoke check.

H Additional related work

Social-choice approaches to alignment (Conitzer et al., 2024) and roadmaps for pluralistic alignment that explicitly include user-steerable policies (Sorensen et al., 2024) focus on accommodating diverse preferences. We do this in an unusually strong way, offering a welfare guarantee for every prescribed group relative to the optimal benchmark lottery for that group.

Population welfare and distortion. Distortion quantifies the welfare lost when decisions use ordinal preferences rather than cardinal utilities (Procaccia & Rosenschein, 2006; Boutilier et al., 2015); Anshelevich et al. (2021) survey this literature. Under individual Bradley–Terry comparisons, Gölz et al. (2025) establish the population benchmark used here. Their Theorem 7 bounds NLHF's asymptotic distortion by L/ℓ_β for every reference lottery and KL budget, relative to the best feasible lottery. Their Theorem 3 proves a matching worst-case lower bound with one comparison per user, using unconstrained instances with sufficiently many alternatives. We apply their welfare conversion within each group and separately derive finite-sample comparison guarantees and refinements for preference-aligned groups. The inherited lower bound concerns the asymptotic welfare factor, not the minimax sample complexity of our algorithm. Recent work by Oko et al. (2026) sharpens population-level RLHF distortion bounds and identifies the importance of matching the comparison-sampling distribution to the reference lottery; with this match, clipped RLHF can achieve $O(\beta)$ distortion. These are all guarantees that marginalize over the whole population; our innovation is giving such guarantees conditional on many groupings of the data simultaneously.

Pairwise games and direct comparison feedback. Nash learning from human feedback (Munos et al., 2024) seeks a policy that cannot be beaten in average pairwise comparisons by a competing policy, with a regularized version controlling deviation from a reference. In the unconstrained finite-alternative setting, the corresponding social-choice solution is a maximal lottery (Fishburn, 1984). GHY's finite-sample implementation estimates pairwise margins and solves the resulting game; Munos et al. (2024, Appendix F.2) also give a policy-gradient update using human comparisons without a learned preference model.

Direct learning from comparisons is also studied in contextual dueling bandits, where each round supplies a context and feedback on a queried pair of actions. The SparringEXP4.P algorithm of Dudík et al. (2015) uses two no-regret learners and one comparison per round. It competes with a finite class of contextual policies and, by averaging, yields an approximate *von Neumann winner*: a randomized policy whose expected pairwise disadvantage against every member of that class is small. Its guarantee averages over contexts. Saha & Krishnamurthy (2022) obtain efficient algorithms with a stronger sequential comparison benchmark under a realizable preference-function class and an online regression oracle. GLHF instead controls comparisons conditional on each named group, relative to all fixed feasible lotteries, without a realizability assumption on the context-dependent preferences.

Learning with overlapping-group guarantees. Early work on subgroup fairness emphasized that guarantees for coarse demographic groups can mask large errors on structured subgroups, and developed algorithms enforcing statistical parity or false-positive-rate parity across rich, overlapping group families (Kearns et al., 2018). Multicalibration similarly seeks to produce probabilistic forecasts that are calibrated not just overall, but also within intersecting subgroups (Hébert-Johnson et al., 2018). We too give multigroup guarantees, but for the maximal lottery conditions rather than measures of error or bias.

Rothblum & Yona (2021) study multigroup PAC learning in which a predictor competes against a different groupwise optimal hypothesis on each of multiple intersecting subgroups. Blum & Lykouris (2020) similarly obtain online groupwise regret in an online setting through a sleeping experts construction. Tosh & Hsu (2022) and Globus-Harris et al. (2022) give simple, efficient algorithms for multigroup learning in the batch setting.

Acharya et al. (2024) give a reduction with one inner online learner per group and an outer learner that combines active inner predictions; its efficiency depends on that of the inner learning problem. GLHF instantiates this architecture for feasible lotteries.

Personalized and group-conditioned preference learning. Group Preference Optimization learns a few-shot preference predictor across training groups. Given examples from a target group, including one not seen during training, it predicts that group's preferences and can supply scores to a downstream steering method (Zhao et al., 2024). Its objective is sample-efficient adaptation and does not give the kind of group-wise baseline efficiency guarantees of our method.

Group Robust Preference Optimization conditions a policy on a group label and minimizes the largest group-specific Direct Preference Optimization (DPO) or Identity Preference Optimization (IPO)

loss (Ramesh et al., 2024). For example, its DPO objective has the explicit minimax form

$$\min_{\pi} \max_{g \in [K]} L_g^{\mathrm{DPO}}(\pi),$$

implemented by adaptively increasing the weight on groups with larger losses. Thus its guarantee is group-robust, but it concerns the largest absolute surrogate loss. A group whose loss remains below the maximum loss value need not be close to its own group-specific optimum.

GLHF gives a stronger guarantee: every prespecified, possibly overlapping group competes with its own best feasible fixed lottery. Under the Bradley–Terry assumption, this yields groupwise cardinal-welfare and distortion guarantees.

Park et al. (2024) give sample-complexity guarantees for personalized reward learning using shared representations or clustering, as well as methods for preference aggregation. Zhong et al. (2026) learn party-specific preferences from offline, party-indexed comparisons and optimize collective-welfare objectives under shared reward structure and coverage assumptions. Projection Optimization treats multiple reward objectives with group-specific weights and aggregation functions (Xiong & Singh, 2025). We permit arbitrary preference heterogeneity within and across groups.

Other ways to represent or protect heterogeneous preferences. Distributional preference learning models utility variation induced by hidden context (Siththaranjan et al., 2024). Pairwise-calibrated reward ensembles match population comparison frequencies without requiring annotator identities or predefined groups (Halpern et al., 2025). Population-proportional alignment studies aggregate policies through representation and incentive axioms (Kim et al., 2026). These objectives retain information about diverse preferences, but differ from conditional welfare guarantees for observable groups. Closer to group welfare protection, Latham & Moharrami (2026) study offline alignment with multiple preference oracles, a target-population objective, and protected-group welfare constraints. Their oracle-specific reward estimation and KL-regularized optimization differ from our direct comparison updates, hard feasible set, and benchmark-relative guarantee for every group.