# LINEAR DISTORTION OF UNCLIPPED RLHF AT EVERY KL BUDGET

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement learning from human feedback (RLHF) fits a single reward model to comparisons from users with heterogeneous preferences and then optimizes a policy under a Kullback–Leibler (KL) constraint. Distortion measures the ratio between the optimal average user utility and that attained by the learned policy. Gölz et al. (2025) show that RLHF can suffer distortion exponential in the Bradley–Terry temperature $\beta$, and Oko et al. (2026) trace this growth to distribution mismatch between the preference data and the KL reference policy. Without such mismatch, they prove distortion linear in $\beta$, which is the optimal order. However, their guarantee at finite KL budgets relies on clipping the fitted reward, and the best known bound for the unclipped reward is quadratic in $\beta$. We show that unclipped RLHF attains this optimal order. Without distribution mismatch and with the reward fitted to population comparisons, unclipped RLHF has worst-case distortion at most $C_0\kappa(\beta)$ at every KL budget, where $C_0 < 2.526$ and $\kappa(\beta) = \beta/[2\tanh(\beta/2)]$ is the known algorithm-independent lower bound. The main difficulty is that a KL-constrained policy depends on reward gaps as well as reward rankings. We address it by comparing reward maximization with win-probability maximization inside the same KL ball. In population simulations on MovieLens and KuaiRec, the unclipped reward has mean relative utility loss below $0.15\%$ over each dataset's evaluation grid and lower mean loss than the clipping rule of Oko et al. (2026). These results indicate that clipping is also unnecessary on the evaluated user profiles.

## 1 INTRODUCTION

Post-training aligns large language models with human preferences using pairwise comparisons between candidate outputs. Reinforcement learning from human feedback (RLHF) fits a scalar reward model to these comparisons and then fine-tunes a policy to maximize the fitted reward (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022).

A single reward model aggregates judgments from users with heterogeneous preferences. Users can disagree about which outputs are most helpful, and a reward fitted to their pooled comparisons can rank outputs differently from their average utility (Siththaranjan et al., 2024). Gölz et al. (2025) study the average utility attained by alignment methods through distortion, a measure from social choice theory (Procaccia & Rosenschein, 2006; Boutilier et al., 2015). Distortion is the ratio between the highest achievable average user utility and that attained by the learned policy.

The policy optimization stage also constrains how far the fine-tuned policy can move from a pretrained reference, following relative-entropy and trust-region methods (Peters et al., 2010; Schulman et al., 2015; 2017). Gölz et al. (2025) evaluate each method against the best policy in the same Kullback–Leibler (KL) ball around the reference. They model each user's comparisons with a Bradley–Terry model whose temperature $\beta > 0$ controls how sharply utility differences determine choices. Without a KL constraint, when each user provides a single comparison, every alignment method incurs distortion of order $\beta$ on some instances (Gölz et al., 2025, Theorem 3). Nash learning from human feedback (NLHF) attains this optimal order (Munos et al., 2024; Gölz et al., 2025), whereas RLHF can incur distortion exponential in $\beta$.

Oko et al. (2026) trace this exponential growth to distribution mismatch between the comparison data and the KL reference. When the distributions match, RLHF with a clipped reward attains distortion linear in $\beta$ at every positive KL budget, matching the lower bound up to a constant factor.



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.

Their clipping rule truncates the fitted reward to an interval determined by the sampling distribution. Standard RLHF pipelines optimize the fitted reward without this truncation (Ouyang et al., 2022), motivating a guarantee for the unclipped objective. For that objective, the existing general finite-budget bound is quadratic in $\beta$, and linear bounds require an additional condition on the utility distribution (Oko et al., 2026).

We study whether the unclipped reward attains optimal-order distortion at every KL budget. Without a KL constraint, RLHF concentrates on the highest-reward outputs. These outputs also maximize average win probability against reference-drawn opponents (Siththaranjan et al., 2024; Gölz et al., 2025), and Oko et al. (2026) establish linear distortion in this setting. A binding KL constraint makes reward gaps central to the analysis: the policy spreads probability across outputs according to their reward differences. Because the fitted reward is a nonlinear function of win probabilities, reward maximization and win-probability maximization can produce different allocations despite identical rankings. Gölz et al. (2025) identify controlling reward magnitudes as a major technical challenge. Our analysis addresses this challenge by comparing the two optimization problems within the same KL ball.

### 1.1 Our Contributions

We establish optimal-order distortion for unclipped population RLHF at every KL budget. Without distribution mismatch and with the reward fitted to population comparison probabilities, the worst-case distortion is at most $C_0\kappa(\beta)$ for every $\tau \in [0, \infty]$, where $\kappa(\beta) = \beta/[2\tanh(\beta/2)] \sim \beta/2$ and $C_0 < 2.526$. The bound holds for any finite number of outputs and any full-support reference. The algorithm-independent lower bound $\kappa(\beta)$ at positive finite budgets (Oko et al., 2026) establishes that this dependence on $\beta$ is optimal up to a constant factor. Our contributions are:

- **A comparison between reward and win-score optimization.** We use the aggregate win scores preserved by population reward fitting to compare the two optima within the same KL ball. Combining this comparison with bounds relating win scores to mean utility yields the distortion guarantee in Theorem 1.
- **Reference transfer and lower-bound constructions.** We extend the guarantee to a different full-support reference with a factor $R^2$, where $R$ measures the spread of the reference density ratio (Corollary 5). We construct instances in which unclipped RLHF attains the lower-bound order, including a finite instance at $\beta = 64$ and $\tau = 8$ with distortion exceeding $1.9\kappa(64)$ (Appendix H).
- **Utility comparisons on observed preference profiles.** In population simulations on MovieLens and KuaiRec, unclipped RLHF has mean relative utility loss below $0.15\%$ over each dataset's evaluation grid and lower mean loss than the clipping rule of Oko et al. (2026) (Section 6).

### 1.2 Related Work

**Distortion and utilitarian alignment.** Distortion measures the average utility a voting rule loses from observing only ordinal preferences (Procaccia & Rosenschein, 2006), and under deterministic rankings even randomized rules cannot avoid distortion growing with the number of alternatives (Boutilier et al., 2015). Probabilistic comparison models change this picture, since how often one alternative beats another carries information about the utility gap (Luce, 1959; Plackett, 1975). Under Bradley–Terry comparisons Borda has distortion bounded in $\beta$ (Gölz et al., 2025; Oko et al., 2026), and Alipour & Goyal (2026) obtain a $2\kappa(\beta)$ bound under Plackett–Luce. These select a single winner, i.e. RLHF without a KL constraint. Our closest prior work is Gölz et al. (2025) and Oko et al. (2026): without distribution mismatch, we extend the optimal order of Oko et al. to the unclipped reward at every positive KL budget. Appendix M treats these works and heterogeneous preferences in more detail.

**Reward learning and KL-constrained optimization.** RLHF fits a scalar reward to pooled comparisons and optimizes a policy against it (Christiano et al., 2017; Ouyang et al., 2022). We analyze the population Bradley–Terry optimum and its KL-constrained policy. KL-constrained reward maximization gives an exponential tilt $\pi \propto \mu e^{\lambda r}$ (Peters et al., 2010; Abdolmaleki et al., 2018). Its allocation

depends on reward gaps, not only rankings. Gölz et al. (2025) identify this as the main obstacle to analyzing RLHF. We address it by comparing reward and win-score optimization within the same KL ball. Appendix M reviews training pipelines and reward models.

**Preference optimization without an explicit reward.** Direct preference optimization reparameterizes the KL-regularized reward objective as a loss on preference pairs (Rafailov et al., 2023). Nash and self-play methods instead optimize against competing policies (Munos et al., 2024; Swamy et al., 2024). We study the distortion of the fitted scalar-reward optimum and compare it empirically with NLHF under the same KL budget.

**Proxy objectives and overoptimization.** Reward maximization is a proxy for serving users that degrades under optimization pressure (Amodei et al., 2016; Skalse et al., 2022), with reward-model scores diverging from human judgement as optimization proceeds (Gao et al., 2023) and policy optimization exploiting correlates of reward such as response length (Singhal et al., 2024; Park et al., 2024) while shifting diversity and generalization (Kirk et al., 2024). We quantify the reward–utility gap through the worst-case distortion of the population reward optimum under a KL constraint.

## 2 Population Model and Main Result

Following Gölz et al. (2025) and Oko et al. (2026), we analyze a single prompt with finitely many candidate outputs, called alternatives, $\mathcal{X} = \{1, \dots, m\}$. Each user has a utility vector $u \in [0,1]^m$, and the user population is a distribution $\mathcal{D}$ over these vectors. Alternative $x$ has mean utility $f_x = \mathbb{E}_{u\sim\mathcal{D}} u_x$. A policy $\pi \in \Delta(\mathcal{X})$ assigns probabilities to the alternatives and attains average utility $\mathbb{E}_\pi f = \sum_x \pi_x f_x$.

**Comparison data.** Let $\mu$ be a distribution over alternatives with $\mu_x > 0$ for every $x$. Each comparison draws a user $u \sim \mathcal{D}$ and two alternatives independently from $\mu$, so $(x, y) \sim \mu \otimes \mu$. We also use $\mu$ as the KL reference, giving the setting without distribution mismatch. The user prefers $x$ to $y$ with Bradley–Terry probability $\sigma(\beta(u_x - u_y))$, where $\sigma(v) = 1/(1 + e^{-v})$. The temperature $\beta > 0$ controls how sharply utility differences determine choices. Averaging over users gives

$$p_{xy} = \mathbb{E}_{u\sim\mathcal{D}} \sigma\big(\beta(u_x - u_y)\big). \tag{1}$$

**Reward fitting.** RLHF fits a single Bradley–Terry reward vector to the pooled comparisons (Bradley & Terry, 1952). The fitted model predicts $\sigma(r_x - r_y)$, so reward differences are on the log-odds scale. Following Oko et al. (2026), we analyze the reward $r$ that maximizes the population log-likelihood

$$\mathcal{L}(r) = \sum_{x,y} \mu_x \mu_y \big[p_{xy} \log \sigma(r_x - r_y) + (1 - p_{xy}) \log \sigma(r_y - r_x)\big]. \tag{2}$$

Here the exact comparison probabilities already incorporate $\beta$. Because they lie strictly between zero and one, $r$ is finite and unique up to an additive constant when $m > 1$ (Appendix A).

**Policy optimization.** For a KL budget $\tau \geq 0$, the feasible policies form

$$\mathcal{B}_\mu(\tau) = \{\pi \in \Delta(\mathcal{X}) : \mathrm{KL}(\pi\|\mu) \leq \tau\}, \tag{3}$$

with $\tau = \infty$ denoting the whole simplex. The unclipped RLHF policy and the largest attainable average utility are

$$q \in \arg\max_{\pi\in\mathcal{B}_\mu(\tau)} \mathbb{E}_\pi r, \qquad V = \max_{\pi\in\mathcal{B}_\mu(\tau)} \mathbb{E}_\pi f. \tag{4}$$

We call $q$ the reward policy. A utility-optimal policy attains $V$ and serves as the utility oracle in the experiments. An optimizer of the penalized objective $\mathbb{E}_\pi r - \eta\,\mathrm{KL}(\pi\|\mu)$, with $\eta > 0$, also solves the constrained problem at the KL budget it uses (Gölz et al., 2025). Thus a guarantee at every budget applies to these penalized optima.

**Distortion.** We assume $\mathbb{E}_\mu f > 0$ and define the distortion of an instance as $\mathrm{dist} = V/\mathbb{E}_q f$. When several policies maximize the reward, $q$ may be any of them. The worst-case distortion $D^\star(\beta, \tau)$ is the supremum over all finite alternative sets, full-support references $\mu$, user distributions $\mathcal{D}$, and reward maximizers $q$.

**Theorem 1** (Linear distortion at every budget)**.** *Let the reward maximize the population likelihood* (2) *and let $\mu$ be the KL reference. Define*

$$\kappa(\beta) = \frac{\beta}{2\tanh(\beta/2)}, \qquad e^{z_0} + z_0 = 3, \qquad C_0 = \frac{2}{z_0}, \tag{5}$$

*where $z_0$ is the unique positive solution. Then $C_0 < 2.526$, and for every $\beta > 0$ and $\tau \in [0, \infty]$,*

$$D^\star(\beta, \tau) \le C_0 \kappa(\beta). \tag{6}$$

Since $\kappa(\beta) \sim \beta/2$ as $\beta \to \infty$, the bound is linear in $\beta$ with the same constant at every budget. For every $\beta > 0$ and $0 < \tau \le \infty$, the algorithm-independent lower bounds of Oko et al. (2026, Theorem 5.2) and Gölz et al. (2025, Theorem 3) give (Appendix F)

$$D^\star(\beta, \tau) \ge \kappa(\beta). \tag{7}$$

The theorem therefore determines the worst-case order at every positive budget. At $\tau = 0$, the only feasible policy is $\mu$, so $D^\star(\beta, 0) = 1$. The quantity $\kappa(\beta)$ is the ratio of the sigmoid's slope at zero to its average slope over $[0, \beta]$ (Gölz et al., 2025, Lemma 1). It measures how strongly comparisons compress large utility differences relative to small ones.

**Proof overview.** We connect reward optimization to utility through each alternative's probability of winning against a reference-drawn opponent. Population reward fitting preserves these probabilities. Section 3 compares their best attainable average with their values under the reward policy inside the same KL ball. Section 4 then relates these quantities to mean utility and controls the remaining comparison term on the same utility scale. Combining the bounds gives a distortion ratio with two coefficients. Balancing them yields the constant $C_0$.

## 3 A Comparison over the KL Ball

Identical rankings can lead to different optimal policies under a KL constraint. We control this difference by comparing reward optimization with optimization of the corresponding win probabilities.

### 3.1 Win Scores Preserved by Population Fitting

We center win probabilities at $1/2$ and call the resulting quantities win scores. For the fitted reward $r$, define

$$s_x = \sum_y \mu_y \left[\sigma(r_x - r_y) - \frac{1}{2}\right]. \tag{8}$$

Thus $s_x + 1/2$ is the probability that $x$ beats a reference-drawn opponent under the fitted model, also called its weighted Borda score. Finite rewards give $-1/2 < s_x < 1/2$. Setting the derivative of (2) to zero gives

$$s_x = \mathbb{E}_{u\sim\mathcal{D}, y\sim\mu}\left[\sigma(\beta(u_x - u_y)) - \frac{1}{2}\right]. \tag{9}$$

This identity (Oko et al., 2026, Lemma A.18) shows that fitting preserves the population win scores. Centering also gives $\mathbb{E}_\mu s = 0$, because $\sigma(v) - 1/2$ is odd and both opponents are drawn from $\mu$.

For the fixed reward vector $r$, the score of each alternative satisfies $s_x = \bar{s}(r_x)$, where

$$\bar{s}(v) = \sum_y \mu_y \left[\sigma(v - r_y) - \frac{1}{2}\right]. \tag{10}$$

This function is strictly increasing, so rewards and win scores rank alternatives identically. Siththaranjan et al. (2024, Theorem 3.1) establish this property for equally weighted opponents, and Oko et al. (2026, Lemma A.20) state it for opponents drawn from $\mu$.

### 3.2 Comparing Reward and Win-Score Optimization

Without a KL constraint, rewards and win scores select the same best alternatives. With a binding constraint, the policy depends on reward gaps, which the nonlinear map $\bar{s}$ changes. We introduce an auxiliary convex function to connect win scores to reward log odds and, subsequently, to a lower bound on utility:

$$\chi(v) = \begin{cases} \frac{1}{2}\operatorname{artanh}(2v), & 0 \le v < 1/2, \\ v, & -1/2 < v < 0. \end{cases} \tag{11}$$

Here $\operatorname{artanh}$ is the inverse hyperbolic tangent. On positive scores, $4\chi(v)$ equals the log odds of probability $1/2 + v$. The linear branch on negative scores makes $\chi$ convex, allowing us to average utility bounds in Section 4. On its domain, $\chi(v) \ge v$, $\chi'$ is positive and nondecreasing, and

$$4\chi(v) \ge \log \frac{1/2 + v}{1/2 - v}. \tag{12}$$

We compare the best expected score $S$ with the expected transformed score $Y$ under the reward policy. In Section 4, $S$ enters the upper bound on attainable utility, while $Y$ enters the lower bound on the reward policy's utility. A third quantity, the log moment-generating function $\Lambda$, controls score variation under the reference:

$$S = \max_{\pi \in \mathcal{B}_\mu(\tau)} \mathbb{E}_\pi s, \quad Y = \mathbb{E}_q \chi(s), \quad \Lambda(t) = \log \mathbb{E}_\mu e^{ts}. \tag{13}$$

Jensen's inequality and (12) give $4Y \ge \mathbb{E}_q r - \log \mathbb{E}_\mu e^r$, linking the transformed score to the reward expectation in the KL optimality equations. The parameter $t > 0$ is chosen in the proof. The comparison takes the following form.

**Lemma 2** (Reward-to-score comparison)**.** *For any finite real reward vector $r$ and full-support $\mu$, take any $\tau \in [0, \infty]$ and any $q \in \arg\max_{\pi \in \mathcal{B}_\mu(\tau)} \mathbb{E}_\pi r$ defining $Y$. For every $0 < t \le 4$,*

$$S \le \frac{4Y + \Lambda(t)}{t}. \tag{14}$$

The lemma compares any reward with its induced scores. Population fitting is used in Section 4 through (9) to connect these scores to user utilities. The range $t \le 4$ follows from $\bar{s}'(v) \le 1/4$.

**Proof idea.** For a nonconstant reward and a positive budget too small to concentrate entirely on the highest-reward alternatives, the optimal policy has the exponential form $q_x \propto \mu_x e^{\lambda r_x}$. The multiplier $\lambda > 0$, the inverse KL penalty coefficient, increases with the budget. The variational inequality $S \le [\tau + \Lambda(t)]/t$ relates score optimization to the budget. For $\lambda \le 1$, applying the log-odds bound (12) to the scaled reward gives $\tau \le 4Y$. For $\lambda \ge 1$, we use the variational inequality at scale $t\lambda$ and compare the log normalizers of the reward and score policies. The split at $\lambda = 1$ compares the policy's reward scale with the fitted reward scale. Both arguments yield Lemma 2 (Appendix B).

## 4 From Win Scores to Utility

We now turn the score comparison into a utility ratio. To put its numerator and denominator on the same scale, define $P = \mathbb{E}_{u \sim \mathcal{D}, y \sim \mu}[\sigma(\beta u_y) - 1/2]$. This is the centered probability that a reference-drawn alternative defeats a hypothetical zero-utility alternative. It is positive because $\mathbb{E}_\mu f > 0$. The magnitude of the lowest score, $P_0 = -\min_x s_x$, will connect this baseline to $\Lambda(t)$. Write $a = 4/\beta$ for the conversion factor from scores to utilities and $\kappa = \kappa(\beta)$.

**Lemma 3** (Utility bounds)**.** *For each alternative $x$,*

$$f_x \ge a[P + \chi(s_x)], \qquad f_x \le a[(2\kappa - 1)P + \kappa s_x]. \tag{15}$$

*Moreover,* $0 \le P_0 \le P$.

To see why these bounds involve one factor of $\kappa$, invert the centered comparison function $g = \sigma(\beta v) - 1/2$: $v = \psi(g) = (2/\beta)\operatorname{artanh}(2g)$. For $g \ge 0$, $\psi(g) = a\chi(g)$. The lower bound retains this curvature, using convexity to average over users and opponents. The upper bound uses the chord of $\psi$ over $[0, \sigma(\beta) - 1/2]$, whose slope is $\kappa a$. Thus only the upper bound incurs this factor

(Appendix C). In comparison, the Borda bound of Gölz et al. (2025, Theorem 2) applies linearization twice and has order $\kappa^2$.

For the exponential policy above, $q_x/\mu_x$ and $\chi(s_x)$ increase with the reward. Their nonnegative covariance gives $Y \geq \mathbb{E}_\mu \chi(s) \geq \mathbb{E}_\mu s = 0$. Appendix B establishes $Y \geq 0$ also at boundary budgets and for constant rewards. Averaging Lemma 3 under $q$ and under a utility-optimal policy therefore gives $\mathbb{E}_q f \geq a(P+Y) > 0$ and $V \leq a[(2\kappa-1)P + \kappa S]$. Hence

$$\mathrm{dist} \leq \frac{(2\kappa-1)P + \kappa S}{P+Y}. \tag{16}$$

Lemma 2 controls $S$ up to the term $\Lambda(t)$. Since $P+Y$ can be arbitrarily small, a constant bound on this term would not give a uniform distortion bound. The next lemma instead controls it by $P_0 \leq P$, on the same scale as the denominator.

**Lemma 4** (Log moment-generating function bound)**.** *For every $t > 0$,*

$$\Lambda(t) \leq 2(e^{t/2} - 1 - t/2)P_0. \tag{17}$$

The identities $\mathbb{E}_\mu s = 0$ and $-P_0 \leq s_x \leq 1/2$ give $\mathbb{E}_\mu s^2 \leq P_0/2$. A quadratic upper bound on $e^{tv}$ for $v \leq 1/2$ then yields (17) (Appendix C).

*Proof of Theorem 1.* Set $z = t/2 \in (0, 2]$. Lemmas 2 and 4 imply

$$S \leq \frac{2}{z} Y + \phi(z) P_0, \qquad \phi(z) = \frac{e^z - 1 - z}{z}. \tag{18}$$

We substitute this bound into (16) and use $P_0 \leq P$. The right-hand side becomes a weighted average of $[2 + \phi(z)]\kappa - 1$ and $2\kappa/z$ with nonnegative weights $P$ and $Y$. Consequently,

$$\mathrm{dist} \leq \max\left\{ [2 + \phi(z)]\kappa - 1, \frac{2\kappa}{z} \right\}. \tag{19}$$

The first coefficient increases with $z$, while the second decreases. Dropping $-1$ from the first gives a uniform bound proportional to $\kappa$. The coefficients then balance at $2 + \phi(z) = 2/z$, or $e^z + z = 3$. Its solution $z_0$ lies in $(0, 2]$ and gives $\mathrm{dist} \leq (2/z_0)\kappa = C_0\kappa$. Numerically, $z_0 > 0.792$, so $C_0 < 2.526$. Keeping $-1$ gives the smaller temperature-dependent constant in Appendix D. □

The proof controls reward optimization through bounded win scores and exponential tilts. Lemma 2 supplies the comparison between whole policies that complements the pointwise utility bounds. This yields linear dependence on $\beta$, improving the quadratic unclipped bound of Oko et al. (2026, Theorem 4.4). The constant is also bounded away from one: the instance in Appendix H has distortion above $1.9\kappa(64)$ at $\beta = 64, \tau = 8$, so any uniform constant for RLHF must exceed 1.9.

**Changing the KL reference.** The common distribution $\mu$ connects fitted win scores, policy normalizers, and the mean-zero identity used above. If a fitted reward is optimized around a different reference $\nu$, we transfer the guarantee by controlling how much the ratio $\nu_x/\mu_x$ varies.

**Corollary 5** (Changing the KL reference)**.** *Keep the population reward fitted from $\mu \otimes \mu$, and let the new reference $\nu$ have full support. Define $q_\nu, V_\nu$ by* (4) *with* $\mathrm{KL}(\pi\|\nu) \leq \tau$ *in place of* $\pi \in \mathcal{B}_\mu(\tau)$*. Set* $\rho_x = \nu_x/\mu_x$ *and* $R = \max_x \rho_x / \min_x \rho_x$*. For every $\beta > 0$ and $\tau \in [0, \infty]$, any reward-maximizing $q_\nu$ satisfies*

$$\frac{V_\nu}{\mathbb{E}_{q_\nu} f} \leq C_0 R^2 \kappa(\beta). \tag{20}$$

Changing the reference reweights policies and changes their KL budgets. Theorem 1 applies at the resulting budgets because its guarantee is uniform in $\tau$. The proof accounts for two factors of $R$: one from converting budgets into attainable utility, and one from reweighting utility (Appendix E). At $R = 1$, the original bound is recovered.

## 5 A Lower Bound with Three Alternatives

The lower-bound instances of Oko et al. (2026) make the valuable alternative indistinguishable from others in the comparison data. An algorithm therefore cannot reliably favor it across those instances. We construct instances where comparisons distinguish the alternatives, yet reward optimization favors the one with lower mean utility. The construction uses matching comparison and reference distributions, in contrast to the RLHF lower bound of Gölz et al. (2025, Theorem 6), and explains distortion of order $\kappa(\beta)$ at every fixed positive budget.

There are three alternatives. A background alternative $b$ gives every user zero utility. Alternative $x$ gives utility one to a fraction $p$ of users, while $y$ gives utility $d$ to everyone. For $0 < \varepsilon < 1/2$, the reference and the two user types are

$$(\mu_b, \mu_x, \mu_y) = (1-2\varepsilon, \varepsilon, \varepsilon), \qquad u^{(1)} = (0,1,d), \quad u^{(2)} = (0,0,d), \quad \Pr(u = u^{(1)}) = p, \tag{21}$$

so the mean utilities are $f = (0, p, d)$. For $0 < p < 1$, define

$$d_0(p) = \frac{2}{\beta} \operatorname{artanh}\big(p \tanh(\beta/2)\big). \tag{22}$$

This is the value of $d$ at which $x$ and $y$ have equal limiting win scores as $\varepsilon \downarrow 0$. Strict convexity of $\operatorname{artanh}$ gives $d_0(p) < p$, so we can choose $d_0(p) < d < p$.

**Proposition 6.** *Fix $\beta > 0$ and $0 < \tau < \infty$. For any $0 < p < 1$ and $d_0(p) < d < p$, the instances in (21) satisfy $\lim_{\varepsilon \downarrow 0} \mathrm{dist} = p/d$.*

**Mechanism.** As $\varepsilon \downarrow 0$, a comparison involving $x$ or $y$ is almost always against $b$. Their win scores approach $(p/2)\tanh(\beta/2)$ and $(1/2)\tanh(\beta d/2)$. Since $d > d_0(p)$, the fitted reward ranks $y$ above $x$, although $f_x = p > d = f_y$. At a fixed positive budget, the largest mass assignable to a rare event of reference probability $\delta$ is asymptotic to $\tau/\log(1/\delta)$. Thus $x$ alone and the pair $\{x, y\}$ can receive asymptotically the same mass. A bounded multiplier $\lambda$ would keep the rare mass of order $\varepsilon$ and make the KL divergence vanish. Maintaining a fixed positive budget therefore forces $\lambda \to \infty$. The fitted gap converges to $\beta[d - d_0(p)] > 0$, so $q_x/q_y = e^{-\lambda(r_y - r_x)} \to 0$. Within the rare pair, the reward policy concentrates on $y$ and the utility-optimal policy on $x$, giving distortion $p/d$ (Appendix F).

Letting $d \downarrow d_0(p)$ and then $p \downarrow 0$ recovers $\kappa(\beta)$ at every fixed positive budget. This construction explains the dependence on $\beta$, while the finite instance in Appendix H constrains the uniform constant. For a fixed instance, distortion approaches one as $\tau \downarrow 0$ (Appendix G).

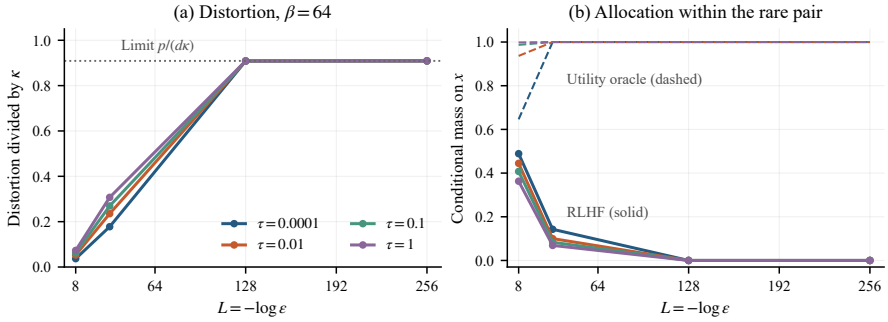


Figure 1: Utility loss as the reference probability $\varepsilon$ of $x$ and $y$ decreases, with $\beta = 64$, $p = 0.01$, and $d = 1.1 d_0(p)$. The horizontal axis is $L = -\log \varepsilon$. Left: distortion divided by $\kappa$ approaches $p/(d\kappa)$ (Proposition 6). Right: conditional on the rare pair, RLHF concentrates on $y$, while the utility-optimal policy concentrates on $x$. Lines connect computed instances.

**Controlled instances.** We evaluate the construction with $p = 0.01$ and $d = 1.1 d_0(p)$ at 48 settings spanning three temperatures, four budgets, and four values of $L = -\log \varepsilon$ (Appendix I). As $L$ grows, RLHF assigns more of its rare-pair mass to the alternative with lower mean utility (Figure 1). At $L = 256$, distortion divided by $\kappa$ is approximately $0.909$ for all 12 temperature–budget pairs: the limit $p/(d\kappa) \approx 1/1.1$ reflects our choice of $d$. From $L = 128$ onward, distortion exceeds $0.9\kappa$ at every tested budget, exhibiting the mechanism of Proposition 6.

# 6 EXPERIMENTS

We measure the utility loss of unclipped RLHF on observed utility profiles, compare reward, win-score and NLHF objectives, and examine clipping and reference changes. We then test for allocation differences on released reward models. In the utility experiments, relative loss is $1 - \mathbb{E}_q f/V = 1 - 1/\operatorname{dist}$, where the utility oracle's value $V$ is the maximum mean utility in the same KL ball.

## 6.1 POPULATION SIMULATIONS ON OBSERVED UTILITY PROFILES

To assess loss beyond the worst-case construction, we simulate population comparisons from MovieLens-1M ratings (Harper & Konstan, 2015) and KuaiRec watch ratios (Gao et al., 2022), mapped to utilities in $[0, 1]$ (Appendix I). Each panel contains items and the users who rated or watched all of them. We fit population Bradley–Terry rewards to the exact probabilities in (1) on 40 MovieLens and 60 KuaiRec panels with 5–20 candidates. The reference is uniform or proportional to rating or exposure counts and also serves as the comparison distribution. Five temperatures from $0.5$ to $128$ and ten budgets from $10^{-4}$ to $\infty$ give 100 settings per panel.

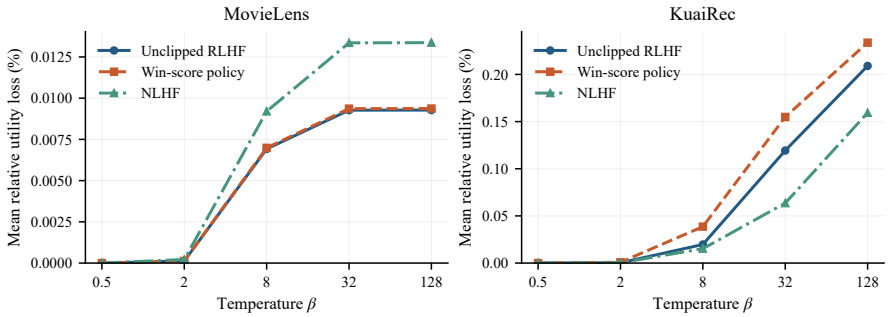


Figure 2: Mean relative utility loss in population simulations using observed utility proxies, relative to the utility oracle. Each point averages the two reference choices and ten budgets within each panel and then averages across panels. Vertical scales differ between panels.

**Loss on observed profiles.** Unclipped RLHF has mean relative loss $0.0051\%$ on MovieLens and $0.070\%$ on KuaiRec (Table 1). Maximum distortions are $1.004$ and $1.062$, attained at saturated budgets where all three policies have the same utility. Figure 2 shows loss increasing with $\beta$, leveling off above 32 on MovieLens and continuing to rise on KuaiRec. This trend is consistent with larger $\beta$ making comparisons approach deterministic utility rankings, while the oracle optimizes utility magnitudes.

**Dependence on the KL budget.** To isolate the effect of the constraint, we vary the budget with reward rankings fixed. At $\beta = 128$, the loss curves averaged across panels and references peak at an intermediate budget, reaching $1.9$ and $2.8$ times their unconstrained levels (Figure 5). Small budgets keep both policies near the reference, while large budgets make their respective maximizing alternatives feasible. The intermediate peaks connect the observed allocation cost to our guarantee for binding KL constraints.

**Win-score and NLHF comparisons.** We compare objectives to examine the allocation distinction in Lemma 2. Reward and win-score objectives rank alternatives identically, yet their constrained optima allocate probability differently. RLHF has lower mean loss than the win-score policy on both datasets (Table 1(a)). NLHF provides a game-based comparison, using the equilibrium closest to $\mu$ (Appendix J). Unclipped RLHF has lower mean loss on MovieLens, while NLHF has lower mean loss on KuaiRec. This dataset dependence is consistent with the distinct scalar and game objectives.

**Utility cost of clipping.** We test the effect of truncating reward gaps with the exact rule of Oko et al. (2026), whose interval has width $2c$. On the 8,000 settings with $\beta \geq 1$, we break ties among clipped optima by maximizing the original reward. At $c = 1/316$, the largest value in their sufficient range, clipping increases mean loss on both datasets, with positive paired-bootstrap intervals for the

Table 1: Mean relative utility loss (%) on MovieLens (ML) and KuaiRec (KR). Blocks (a)/(b) use 100/80 settings per panel. In (b), $\Delta$ is clipped ($c = 1/316$) minus unclipped loss in percentage points, with paired 95% panel-bootstrap intervals.

| Policy | ML | KR |
|---|---|---|
| (a) Full grid: 4,000 / 6,000 settings | | |
| Unclipped RLHF | 0.0051 | 0.0697 |
| Win score | 0.0052 | 0.0856 |
| NLHF | 0.0072 | 0.0477 |
| (b) $\beta \geq 1$: 3,200 / 4,800 settings | | |
| Unclipped RLHF | 0.0064 | 0.0871 |
| Clipped, $c = 1/316$ | 0.2141 | 1.3089 |
| Clipped, $c = 1$ | 0.0290 | 0.3168 |
| $\Delta$ (percentage points) | 0.21 | 1.22 |
| 95% interval | [0.18, 0.24] | [1.07, 1.39] |

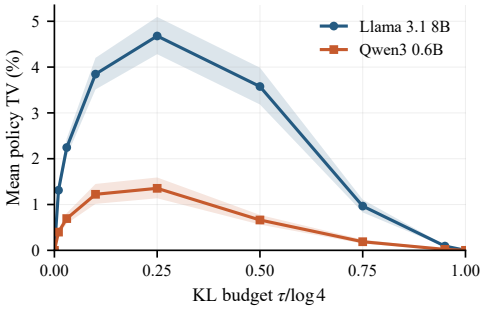


Figure 3: Mean total variation between the reward and win-score policies of released Skywork reward models on 200 HelpSteer prompts, with pointwise 95% prompt-bootstrap intervals. Budgets are divided by $\log 4$, the KL cost of placing all probability on one of four answers under the uniform reference.

increases (Table 1(b)). The loss remains higher on settings with a unique clipped optimum, linking the effect to truncation even when tie selection is absent (Table 2). Across the six widths in Figure 6, larger intervals reduce mean loss, with the unclipped reward retaining the lowest mean. This pattern connects utility loss to the reward spacings that clipping changes and that constrained allocation depends on.

**Changing the reference.** To examine the reference transfer in Corollary 5, we hold comparisons fixed and switch the KL reference between uniform and count-proportional distributions. On the separate, reduced temperature–budget grid in Table 3, dataset means remain below $0.15\%$ and shift by at most $0.013$ percentage points in either direction. The changes reach a density-ratio spread $R = 5.95$ on KuaiRec. This extends the low mean relative losses to changed KL references.

### 6.2 Allocation Differences in Released Reward Models

We next examine whether the same allocation distinction appears with trained reward models. We score four HelpSteer answers for each of 200 prompts (Wang et al., 2024b) using the released Skywork-Reward-V2 Llama 3.1 8B and Qwen3 0.6B models (Liu et al., 2026). We compute implied win scores via (8) with $\mu_i = 1/4$ and compare the reward policy $q_r$ with the win-score policy $q_s$ in the same KL ball. Total variation (TV) measures the probability mass reallocated between the two policies.

Figure 3 shows that mean TV rises and then falls as the budget expands, peaking at $4.68\%$ for Llama and $1.35\%$ for Qwen. Table 4 reports the corresponding uncertainty and prompt-level distributions at each budget. At zero budget both policies equal the reference, and at $\log 4$ they select the same maximizing answer. The differences at interior budgets show how objectives with identical rankings produce different allocations on released rewards.

## 7 Conclusion

Without distribution mismatch, unclipped population RLHF has the optimal distortion order at every positive KL budget. We prove the bound $C_0\kappa(\beta)$ by comparing reward and win-score optimization inside the same KL ball, transfer it to references with a bounded density-ratio spread, and construct a rare-alternative instance in which the fitted reward attains this order at every fixed positive budget. On observed utility profiles, mean relative loss of unclipped RLHF stays below $0.15\%$ when averaged over all panels and parameter settings in each dataset's evaluation grid. The evaluated clipping rule has higher mean loss. A next step is to analyze finite comparison samples and reward models shared across prompts.

### AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure. Generative AI tools helped develop the theoretical models and conceptual frameworks, helped formulate the mathematical claims, provided critical ingredients for proving those claims, and assisted in the writing of the proofs. The mathematical claims, the proofs, and every experimental result were checked and verified by the authors. We additionally used generative AI tools for tasks with recommended disclosure, namely to aid and polish the writing and to edit the paper for readability. We have reviewed all AI-assisted work, including the mathematical derivations and the text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1806.06920. arXiv:1806.06920.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017. URL https://arxiv.org/abs/1705.10528.

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. Nemotron-4 340B technical report, 2024. URL https://arxiv.org/abs/2406.11704. arXiv:2406.11704.

Hamidreza Alipour and Mohak Goyal. Utilitarian distortion under probabilistic voting, 2026. URL https://arxiv.org/abs/2602.11152. arXiv:2602.11152.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, 2016. URL https://arxiv.org/abs/1606.06565. arXiv:1606.06565.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL https://arxiv.org/abs/2204.05862. arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback, 2022b. URL https://arxiv.org/abs/2212.08073. arXiv:2212.08073.

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. URL https://www.sciencedirect.com/science/article/pii/S0004370215000892.

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. URL https://academic.oup.com/biomet/article-abstract/39/3-4/324/326091.

Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5409–5435. PMLR, 2024. URL https://proceedings.mlr.press/v235/calandriello24a.html.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217. arXiv:2307.15217.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Singh Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6116–6135. PMLR, 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4299–4307. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

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. PMLR, 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback, 2023. URL https://arxiv.org/abs/2310.01377. arXiv:2310.01377.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF, 2024. URL https://arxiv.org/abs/2405.07863. arXiv:2405.07863.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization, 2024. URL https://arxiv.org/abs/2402.01306. arXiv:2402.01306.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 2016. URL https://arxiv.org/abs/1512.08562. arXiv:1512.08562.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 540–550, 2022. doi: 10.1145/3511808.3557220. URL https://kuairec.com/.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023. URL https://arxiv.org/abs/2210.10760.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI alignment from human feedback. In *Advances in Neural Information Processing Systems*, volume 37, pp. 80439–80465. Curran Associates, Inc., 2024. doi: 10.52202/079017-2557. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9328208f88ec69420031647e6ff97727-Abstract-Conference.html.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Zhaohan Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems*, volume 38, pp. 29455–29493, 2025. doi: 10.52202/085713-0876. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2526c5e8110bc6bc8b462ba95198161e-Abstract-Conference.html.

GroupLens Research. MovieLens 1M dataset, 2003. URL https://grouplens.org/datasets/movielens/1m/.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online AI feedback, 2024. URL https://arxiv.org/abs/2402.04792. arXiv:2402.04792.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018. URL https://arxiv.org/abs/1801.01290.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.626. URL https://doi.org/10.18653/v1/2024.emnlp-main.626.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2312.14925.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06452. arXiv:2310.06452.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124. arXiv:2411.15124.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050. arXiv:2305.20050.

Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-Reward-V2: Scaling preference data curation via human-AI synergy. In *International Conference on Learning Representations*, pp. 133805–133838, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8836d0294651a867c2ca51012923997-Abstract-Conference.html.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, New York, 1959.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3946. URL https://doi.org/10.52202/079017-3946.

Remi 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. PMLR, 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes, 2017. URL https://arxiv.org/abs/1705.07798. arXiv:1705.07798.

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*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=x4qXiDogRm.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.297. URL https://doi.org/10.18653/v1/2024.findings-acl.297.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, 2007. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

Jan Peters, Katharina Mülling, and Yasemin Altün. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010. doi: 10.1609/aaai.v24i1.7727. URL https://ojs.aaai.org/index.php/AAAI/article/view/7727.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. doi: 10.2307/2346567. URL https://academic.oup.com/jrsssc/article/24/2/193/6953554.

Ariel D. Procaccia and Jeffrey S. Rosenschein. The distortion of cardinal preferences in voting. In Matthias Klusch, Michael Rovatsos, and Terry R. Payne (eds.), *Cooperative Information Agents X*, volume 4149 of *Lecture Notes in Computer Science*, pp. 317–331, Berlin, Heidelberg, 2006. Springer. doi: 10.1007/11839354_23. URL https://link.springer.com/chapter/10.1007/11839354_23.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From $r$ to $q^*$: Your language model is secretly a Q-function, 2024. URL https://arxiv.org/abs/2404.12358. arXiv:2404.12358.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences, 2024. URL https://arxiv.org/abs/2404.03715. arXiv:2404.03715.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*. PMLR, 2015. URL https://arxiv.org/abs/1502.05477.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347. arXiv:1707.06347.

Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018. URL https://jmlr.org/papers/v18/16-206.html.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.03716.

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*, pp. 27448–27471, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7423902b5534e2b267438c85444a54b1-Abstract-Conference.html.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0687. URL https://doi.org/10.52202/068431-0687.

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. PMLR, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020. URL https://arxiv.org/abs/2009.01325. arXiv:2009.01325.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47345–47377. PMLR, 2024. URL https://proceedings.mlr.press/v235/swamy24a.html.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2404.14367.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2402.05749.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://doi.org/10.18653/v1/2024.findings-emnlp.620.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3384. Association for Computational Linguistics,

2024b. doi: 10.18653/v1/2024.naacl-long.185. URL https://aclanthology.org/2024.naacl-long.185/.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2574. URL https://doi.org/10.52202/075280-2574.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2312.11456.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation, 2024. URL https://arxiv.org/abs/2401.08417. arXiv:2401.08417.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024. URL https://arxiv.org/abs/2401.10020. arXiv:2401.10020.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback, 2023. URL https://arxiv.org/abs/2305.10425. arXiv:2305.10425.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2019. URL https://arxiv.org/abs/1909.08593. arXiv:1909.08593.

## A Population Likelihood and Policy Optimization

The symmetry $p_{xy} + p_{yx} = 1$ implies

$$\frac{\partial \mathcal{L}(r)}{\partial r_x} = 2\mu_x \sum_y \mu_y [p_{xy} - \sigma(r_x - r_y)].$$

For any direction $v$,

$$v^\top \nabla^2 \mathcal{L}(r) v = -\sum_{x,y} \mu_x \mu_y \sigma'(r_x - r_y)(v_x - v_y)^2.$$

All weights are positive. Therefore the objective is strictly concave on a gauge subspace such as $r_1 = 0$. If a reward difference diverges, the strictly positive probability of the comparison opposing that difference makes the corresponding expected log-likelihood tend to minus infinity. There are finitely many pairs and the remaining log-likelihood terms are nonpositive. Thus the objective is coercive on the gauge subspace and has a unique finite maximizer there. Its first-order equations imply (9).

For a nonconstant objective vector $v$, define $J_v(\lambda) = \log \mathbb{E}_\mu e^{\lambda v}$ and $\pi_\lambda(x) = \mu_x e^{\lambda v_x - J_v(\lambda)}$. The KL divergence is $K_v(\lambda) = \lambda J_v'(\lambda) - J_v(\lambda)$, with $K_v'(\lambda) = \lambda \operatorname{Var}_{\pi_\lambda}(v) > 0$ for $\lambda > 0$. It increases from zero to $-\log \mu(\arg\max v)$. For its unique interior root $K_v(\lambda) = \tau$ and any feasible $\pi$,

$$\lambda(\mathbb{E}_{\pi_\lambda} v - \mathbb{E}_\pi v) = \tau - \mathrm{KL}(\pi \| \mu) + \mathrm{KL}(\pi \| \pi_\lambda) \geq 0.$$

This proves optimality and uniqueness of the interior tilt. At a saturated budget, the conditional reference distribution on $\arg\max v$ is an optimal choice. These formulas are used for the reward, win-score, and utility objectives in the experiments. A scalar root solve determines the multiplier for each interior budget.

## B Proof of Lemma 2

We say that a quantity $w_x$ indexed by alternatives is nondecreasing in the reward if $w_x = g(r_x)$ for a nondecreasing function $g$. Two such quantities are ordered in the same way over alternatives, so their covariance under any distribution on $\mathcal{X}$ is nonnegative. By (10), win scores are nondecreasing in the reward. We first treat a nonconstant reward and an interior budget $0 < \tau < -\log \mu(\mathcal{X}^\star)$, where $\mathcal{X}^\star = \arg\max_x r_x$. The last two paragraphs treat the remaining budgets and ties and show that $Y \geq 0$.

**Exponential tilt.** Write $J(\gamma) = \log \mathbb{E}_\mu e^{\gamma r}$ for the log normalizer at scale $\gamma \geq 0$. For a nonconstant reward and an interior budget, the reward policy is the unique exponential tilt

$$q_x = \mu_x e^{\lambda r_x - J(\lambda)}, \quad \tau = \lambda \mathbb{E}_q r - J(\lambda), \quad \lambda > 0. \tag{23}$$

This is the usual form of a KL-regularized optimum, with $1/\lambda$ in the role of the KL penalty coefficient. The KL divergence of the tilt increases continuously from zero to $-\log \mu(\mathcal{X}^\star)$ as $\lambda$ grows, because its derivative is $\lambda \operatorname{Var}_q(r) > 0$ (Appendix A). Every interior budget therefore has a unique multiplier $\lambda$.

**Proof strategy.** We treat $\lambda \leq 1$ and $\lambda \geq 1$ separately. Both cases bound $S$ through the variational inequality (24) and differ in how they control its right-hand side. Both convert rewards into log odds through $4Y \geq \mathbb{E}_q r - J(1)$. The cutoff is 1 because $Y$ and the bound (26) below use the reward at its fitted scale. The multiplier $\lambda$ depends jointly on the reward, the reference, and the budget, so the two cases do not correspond to a fixed range of budgets. In the first case, $Y$ controls the budget $\tau$. In the second, the log normalizer $J$ controls the log moment-generating function $\Lambda$.

**Two inequalities.** Two inequalities connect the tilt with win scores. The first is a variational bound on the score optimum. For every $\theta > 0$,

$$S \leq \frac{\tau + \Lambda(\theta)}{\theta}, \tag{24}$$

which follows from the nonnegativity of $\mathrm{KL}(\pi \| \mu e^{\theta s - \Lambda(\theta)})$ for each $\pi \in \mathcal{B}_\mu(\tau)$. The second converts rewards into log odds. For fixed $x$, the map $v \mapsto e^{r_x}/(e^{r_x} + v)$ is convex on $v > 0$, and Jensen's inequality gives

$$\frac{1}{2} + s_x = \mathbb{E}_{y \sim \mu} \frac{e^{r_x}}{e^{r_x} + e^{r_y}} \geq \frac{e^{r_x}}{e^{r_x} + \mathbb{E}_\mu e^r}. \tag{25}$$

Taking log odds and applying (12), we obtain

$$4\chi(s_x) \geq r_x - J(1), \qquad 4Y \geq \mathbb{E}_q r - J(1). \tag{26}$$

**Case $0 < \lambda \leq 1$: the budget is controlled by $Y$.** Keep the policy $q$ fixed and scale the reward inside the win score. For $\gamma \geq 0$, define

$$\begin{aligned} s_x(\gamma) &= \sum_y \mu_y [\sigma(\gamma(r_x - r_y)) - 1/2], \\ F(\gamma) &= \mathbb{E}_q \chi(s(\gamma)). \end{aligned} \tag{27}$$

We show below that $F$ is nondecreasing, because $q_x/\mu_x$ and $\chi'(s_x(\gamma))$ are both nondecreasing in the reward. Applying (26) to the scaled reward $\lambda r$ gives $4\chi(s_x(\lambda)) \geq \lambda r_x - J(\lambda) = \log(q_x/\mu_x)$. Averaging under $q$ and using $\lambda \leq 1$, we obtain

$$Y = F(1) \geq F(\lambda) \geq \frac{1}{4} \mathrm{KL}(q \| \mu) = \frac{\tau}{4}. \tag{28}$$

Substituting $\tau \leq 4Y$ into (24) with $\theta = t$ proves the lemma in this case.

**Monotonicity of $F$.** Write $w_x = q_x/\mu_x$ and symmetrize the derivative of $F$:

$$\begin{aligned} F'(\gamma) = \frac{1}{2} \sum_{x,y} \mu_x \mu_y \left[ w_x \chi'(s_x(\gamma)) - w_y \chi'(s_y(\gamma)) \right] \\ \cdot (r_x - r_y) \sigma'(\gamma(r_x - r_y)) \geq 0. \end{aligned} \tag{29}$$

Both $w_x$ and $\chi'(s_x(\gamma))$ are positive and nondecreasing in $r_x$. Their product has the same ordering, so every term in the sum is nonnegative. The argument also covers ties and $\gamma = 0$.

**Case $\lambda \geq 1$: log moment-generating functions are compared.** For $\lambda > 1$, monotonicity gives $F(1) \leq F(\lambda)$, so $Y$ no longer controls the budget. We instead take $\theta = t\lambda$ in (24) and compare normalizers. We show below that

$$J(\lambda) - \lambda J(1) \geq \Lambda(t\lambda) - \lambda\Lambda(t). \tag{30}$$

Using $\tau = \lambda \mathbb{E}_q r - J(\lambda)$ from (23), we obtain

$$\begin{aligned} S &\leq \frac{\tau + \Lambda(t\lambda)}{t\lambda} \\ &\leq \frac{\mathbb{E}_q r - J(1) + \Lambda(t)}{t} \leq \frac{4Y + \Lambda(t)}{t}, \end{aligned} \tag{31}$$

where the last step uses (26).

Both sides of (30) have the same form. For a vector $v$ on $\mathcal{X}$, the gap $\log \mathbb{E}_\mu e^{\lambda v} - \lambda \log \mathbb{E}_\mu e^v$ is nonnegative for $\lambda \geq 1$ by convexity. It is zero when $v$ is constant and measures how unevenly $v$ is spread under $\mu$. The left side is this gap for the reward $r$, and the right side is the gap for the scaled win scores $ts$. The inequality rests on the contraction from rewards to win scores. The function $\bar{s}$ in (10) has derivative in $(0, 1/4]$. For $0 < t \leq 4$, both $ts$ and the residual $\ell = r - ts$ are therefore nondecreasing in the reward. The reward is thus the scaled win score plus a residual ordered in the same way. To show that adding this residual cannot decrease the gap, we interpolate between the two exponents through $r^{(\xi)} = (1-\xi)ts + \xi r$ for $\xi \in [0, 1]$ and set

$$G(\xi) = \log \mathbb{E}_\mu e^{\lambda r^{(\xi)}} - \lambda \log \mathbb{E}_\mu e^{r^{(\xi)}}. \tag{32}$$

The endpoint values $G(0)$ and $G(1)$ are the right and left sides of (30). The derivative $G'(\xi)$ is a covariance between $\ell$ and $e^{(\lambda-1)r^{(\xi)}}$ under a tilted distribution. For $\lambda \geq 1$, both are nondecreasing in the reward, so the covariance is nonnegative and $G(1) \geq G(0)$. The computation follows.

**Computation for $\lambda \geq 1$.** The score function $\bar{s}$ in (10) satisfies

$$0 < \bar{s}'(v) = \sum_y \mu_y \sigma'(v - r_y) \leq \frac{1}{4}. \tag{33}$$

Since $s_x = \bar{s}(r_x)$, for $0 < t \leq 4$ both $ts_x$ and $\ell_x = r_x - ts_x$ are nondecreasing in $r_x$. So is $r_x^{(\xi)} = (1-\xi)ts_x + \xi r_x$ for every $\xi \in [0, 1]$. With $G$ from (32), put

$$\pi_x^{(\xi)} = \frac{\mu_x e^{r_x^{(\xi)}}}{\mathbb{E}_\mu e^{r^{(\xi)}}}, \qquad W_x(\xi) = e^{(\lambda-1)r_x^{(\xi)}}.$$

Since $\partial r^{(\xi)}/\partial \xi = \ell$, differentiation gives

$$G'(\xi) = \lambda \, \frac{\mathrm{Cov}_{\pi^{(\xi)}}(\ell, W(\xi))}{\mathbb{E}_{\pi^{(\xi)}} W(\xi)}. \tag{34}$$

Because $\lambda \geq 1$, both $\ell$ and $W(\xi)$ are nondecreasing in the reward. For any distribution $\pi$ on $\mathcal{X}$, this common ordering implies

$$\mathrm{Cov}_\pi(\ell, W) = \frac{1}{2} \sum_{x,y} \pi_x \pi_y (\ell_x - \ell_y)(W_x - W_y) \geq 0.$$

Thus $G(1) \geq G(0)$, which is (30).

**Boundary budgets and ties.** If $r$ is constant, then $s = 0$ and the result is immediate. If $\tau = 0$, then $q = \mu$ and $S = 0$, while $Y \geq \mathbb{E}_\mu s = 0$ and $\Lambda(t) \geq 0$. If $\tau \geq -\log \mu(\mathcal{X}^\star)$, a policy supported on $\mathcal{X}^\star$ is feasible and every reward optimum is supported there. Rewards and win scores have the same maximizers, so $S = \max_x s_x \geq 0$ and $Y = \chi(\max_x s_x) \geq S$. Now $4/t \geq 1$ and $\Lambda(t) \geq 0$ prove the bound. This includes the unrestricted budget and every feasible distribution within a tied maximum set.

**Nonnegativity of $Y$.** Section 4 uses $Y \geq 0$. For an interior tilt, $q_x/\mu_x$ and $\chi(s_x)$ are both nondecreasing in the reward. Their covariance under $\mu$ is therefore nonnegative, and $Y = \mathbb{E}_\mu[(q/\mu)\chi(s)] \geq \mathbb{E}_\mu \chi(s) \geq \mathbb{E}_\mu s = 0$. The boundary cases above give $Y \geq 0$ for every reward maximizer $q$. For a constant reward, every feasible policy maximizes the reward, and $Y = 0$ for each of them.

## C Proofs of Lemmas 3 and 4

*Proof of Lemma 3.* Put $h(v) = \sigma(\beta v) - 1/2$ and $\psi(g) = (2/\beta)\,\mathrm{artanh}(2g)$, so $\psi(h(v)) = v$. For a fixed user and opponent $y$, define $Q = h(u_y) \geq 0$ and $g = h(u_x - u_y)$, so $u_x - u_y = \psi(g)$ and $u_y = \psi(Q)$. Averaging $Q$ over users and opponents gives $P$. The lower bound keeps $\psi$ exact for positive $g$, which produces $\chi$. Nonnegativity of $u_x$ gives $g \geq -Q$. Averaging this inequality proves $s_x \geq -P$, and hence $P_0 \leq P$. Since $\mathbb{E}_\mu s = 0$, the smallest win score is nonpositive, so $P_0 \geq 0$.

For the lower utility bound, observe that $u_x = \psi(Q) + \psi(g)$ and $\psi(Q) \geq aQ$. When $g \geq 0$, $\psi(g) = a\chi(g)$. When $g < 0$, we have $|g| \leq Q$. The function $v \mapsto \psi(v) - av$ is nondecreasing for $v \geq 0$, so

$$\begin{aligned}\psi(Q) + \psi(g) - a(Q+g) = [\psi(Q) - aQ] \\ - [\psi(|g|) - a|g|] \geq 0.\end{aligned}$$

Thus $u_x \geq a[Q + \chi(g)]$ in both cases. Averaging and applying the convexity of $\chi$ gives the lower half of (15).

For the upper bound, convexity of $\psi$ on $[0, h(1)]$ gives the chord inequality $\psi(g) \leq \kappa a g$ for $0 \leq g \leq h(1)$. For $g < 0$, oddness of $\psi$ gives

$$\psi(g) \leq ag \leq \kappa a g + (\kappa - 1)aQ,$$

where the second inequality follows from $g + Q \geq 0$. After averaging $u_x - u_y = \psi(g)$, we obtain $f_x \leq \mathbb{E}_\mu f + \kappa a s_x + (\kappa - 1)aP$. The chord bound applied to $Q$ also gives $\mathbb{E}_\mu f \leq \kappa a P$. Combining these two inequalities proves the result. □

*Proof of Lemma 4.* Since $-P_0 \leq s_x \leq 1/2$ and $\mathbb{E}_\mu s = 0$, averaging $(s_x + P_0)(1/2 - s_x) \geq 0$ gives $\mathbb{E}_\mu s^2 \leq P_0/2$. Taylor's integral remainder gives

$$\frac{e^{tv} - 1 - tv}{v^2} = t^2 \int_0^1 (1 - \vartheta) e^{t\vartheta v}\, d\vartheta,$$

with its continuous value at zero. This expression increases with $v$. For $v \leq 1/2$, we therefore have

$$e^{tv} \leq 1 + tv + 4(e^{t/2} - 1 - t/2)v^2.$$

Take expectations, use the second-moment bound, and apply $\log(1 + w) \leq w$ to obtain (17). □

**Every reward maximizer and constant rewards.** Lemma 3 holds for each alternative. Averaging its lower bound over any distribution $q$ gives $\mathbb{E}_q f \geq a[P + \mathbb{E}_q \chi(s)]$, and averaging its upper bound over a utility optimum gives $V \leq a[(2\kappa - 1)P + \kappa S]$. Appendix B shows that $Y \geq 0$ and Lemma 2 hold for every reward maximizer. Hence (16) and the proof of Theorem 1 apply to every choice of $q$ allowed in the definition of $D^\star(\beta, \tau)$. This includes a constant reward, for which every feasible policy is a reward maximizer. In that case $s = 0$, so $S = Y = 0$, and (16) gives $\mathrm{dist} \leq 2\kappa - 1 \leq C_0 \kappa$. The bound $\mathbb{E}_q f \geq aP > 0$ also holds for every such policy.

**Why no additional factor of $\beta$ appears.** The unclipped bound of Oko et al. (2026, Theorem 4.4) is quadratic in $\beta$. Their analysis bounds the fitted reward itself and uses the fact that its maximum is at most $\beta$ under their normalization (Oko et al., 2026, Section 4.2). Our proof never bounds reward magnitudes. It uses the reward only through the exponential tilt (23) and the log-odds bound $4Y \geq \mathbb{E}_q r - J(1)$. All other estimates concern win scores, which lie in $[-1/2, 1/2]$ for every $\beta$.

**Where the absence of distribution mismatch is used.** The proof uses $\mu$ in two roles. Win scores average over opponents drawn from the comparison distribution, whereas $\mathcal{B}_\mu(\tau)$ and $\Lambda(t)$ are defined through the KL reference. Because both distributions equal $\mu$, the win scores have mean zero under the reference. This fact gives $Y \geq 0$ and the second-moment bound behind Lemma 4. For the same reason, the baseline $P$ and the expectation defining $\Lambda(t)$ use the same distribution, which lets Lemma 4 relate them. Corollary 5 shows that changing the reference costs at most a factor $R^2$, where $R$ measures how much the density ratio between the two references varies across alternatives.

## D A Refined Constant for a Fixed Temperature

**Corollary 7.** *Let $z_\beta$ be the unique positive solution of*

$$e^{z_\beta} + (1 - 1/\kappa(\beta))z_\beta = 3. \tag{35}$$

*Then $z_0 \leq z_\beta \leq \log 3$ and*

$$D^\star(\beta, \tau) \leq \frac{2\kappa(\beta)}{z_\beta}. \tag{36}$$

*Proof.* The two coefficients in (19) coincide at $z = z_\beta$ from (35). Since $\kappa \geq 1$, the unique positive solution lies in $[z_0, \log 3] \subset (0, 2]$. Substituting it into (19) proves the bound. □

## E Transferring a Guarantee to a Different Reference

We prove Corollary 5 through a general property of KL-constrained reward optimization. The reward and utility vectors remain fixed throughout this argument. Only the reference in the constraint changes. The corollary keeps the standing assumption $\mathbb{E}_\mu f > 0$ of Section 2. Because $f \geq 0$ and both references have full support, this is equivalent to $\mathbb{E}_\nu f > 0$, so the ratio in (20) is well defined.

**Proof idea.** Around either reference, the reward policy is an exponential tilt of the same reward. At a common multiplier $\lambda$, the tilt around $\nu$ is the tilt around $\mu$ reweighted by $\rho$. The tilt around $\mu$ is the reward policy for $\mu$ at the budget it uses, and Theorem 1 applies there because it holds at every budget. Reweighting a utility optimum for $\nu$ by $1/\rho$ gives a feasible policy for $\mu$ at another budget. The factor $R^2$ collects two losses, up to normalizing constants that cancel. Reweighting changes KL divergences by at most a factor $R$, so the two budgets differ by at most this factor. Concavity of the attainable utility in the budget turns this budget ratio into a utility ratio. Moving mean utility between the two references costs the second factor.

**Lemma 8** (Common reweighting)**.** *Let $\pi, \zeta$ be probability distributions on a finite set, with $\zeta_i > 0$, and let $0 < \underline{w} \leq w_i \leq \overline{w}$. Write*

$$c_\pi = \mathbb{E}_\pi w, \qquad c_\zeta = \mathbb{E}_\zeta w,$$
$$\pi_i^w = \frac{w_i \pi_i}{c_\pi}, \qquad \zeta_i^w = \frac{w_i \zeta_i}{c_\zeta}.$$

*Then*

$$\frac{\underline{w}}{c_\pi} \mathrm{KL}(\pi \| \zeta) \leq \mathrm{KL}(\pi^w \| \zeta^w) \leq \frac{\overline{w}}{c_\pi} \mathrm{KL}(\pi \| \zeta). \tag{37}$$

*Proof.* For $v \geq 0$ and $z > 0$, define $h(v, z) = v \log(v/z) - v + z \geq 0$, using $h(0, z) = z$. For every $\eta > 0$,

$$\begin{aligned} \underline{w} \sum_i h(\pi_i, \eta\zeta_i) &\leq \sum_i w_i h(\pi_i, \eta\zeta_i) \\ &\leq \overline{w} \sum_i h(\pi_i, \eta\zeta_i). \end{aligned}$$

The unweighted sum is minimized at $\eta = 1$, with value $\mathrm{KL}(\pi \| \zeta)$. The weighted sum is minimized at $\eta = c_\pi / c_\zeta$, with value $c_\pi \, \mathrm{KL}(\pi^w \| \zeta^w)$. Taking infima over $\eta$ proves (37). □

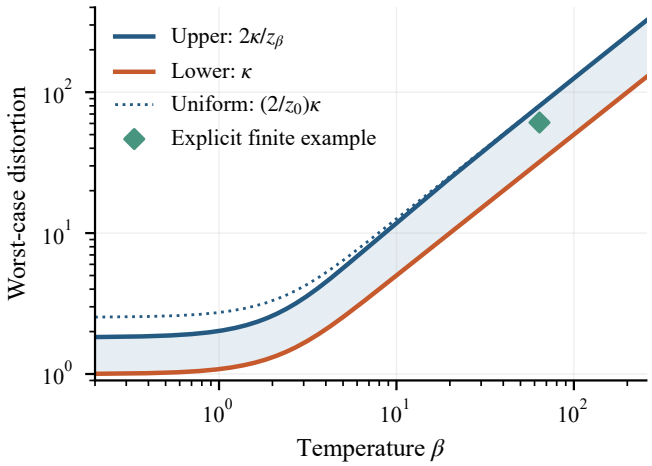


Figure 4: Our upper bounds and the known lower bound (Oko et al., 2026). The solid upper curve uses Corollary 7. The dotted curve uses the uniform constant $C_0$. The diamond is the explicit instance in Appendix H, with distortion 60.94 at $(\beta, \tau) = (64, 8)$.

For a full-support reference $\zeta$, use

$$\mathcal{B}_\zeta(t) = \{\pi : \mathrm{KL}(\pi\|\zeta) \le t\}, \qquad V_\zeta(t) = \max_{\pi \in \mathcal{B}_\zeta(t)} \mathbb{E}_\pi f.$$

The following proposition applies to any finite reward vector and nonnegative utility vector with a guarantee at all budgets.

**Proposition 9** (Reference transfer)**.** *Fix $r \in \mathbb{R}^m$, $f \in \mathbb{R}^m_+$, full-support references $\mu, \nu$, and $A \ge 1$. Suppose that for every finite $t \ge 0$ and every $\bar{q} \in \arg\max_{\pi \in \mathcal{B}_\mu(t)} \mathbb{E}_\pi r$,*

$$V_\mu(t) \le A\mathbb{E}_{\bar{q}} f. \tag{38}$$

*With $R$ defined in Corollary 5, every $\tau \in [0, \infty]$ and $q \in \arg\max_{\pi \in \mathcal{B}_\nu(\tau)} \mathbb{E}_\pi r$ satisfy*

$$V_\nu(\tau) \le AR^2 \mathbb{E}_q f. \tag{39}$$

*Proof.* Set $\rho_i = \nu_i/\mu_i$, $\underline{\rho} = \min_i \rho_i$, and $\overline{\rho} = \max_i \rho_i$, so $R = \overline{\rho}/\underline{\rho}$. First suppose $r$ is nonconstant and $0 < \tau < -\log \nu(\mathcal{X}^\star)$, where $\mathcal{X}^\star = \arg\max_i r_i$. By the exponential-tilt characterization, for some $\lambda > 0$,

$$q_i = \frac{\nu_i e^{\lambda r_i}}{\mathbb{E}_\nu e^{\lambda r}}, \qquad \mathrm{KL}(q\|\nu) = \tau.$$

At the same multiplier, define

$$\bar{q}_i = \frac{\mu_i e^{\lambda r_i}}{\mathbb{E}_\mu e^{\lambda r}}, \qquad \tau_1 = \mathrm{KL}(\bar{q}\|\mu) > 0, \quad c = \mathbb{E}_{\bar{q}}\rho.$$

Then $\bar{q}$ is reward-optimal at budget $\tau_1$ under $\mu$, and $q_i = \rho_i \bar{q}_i / c$. Applying the upper bound in (37) gives

$$\tau \le \frac{\overline{\rho}}{c}\tau_1, \qquad \tau_1 \ge \frac{c}{\overline{\rho}}\tau. \tag{40}$$

Choose any utility-optimal policy $o \in \mathcal{B}_\nu(\tau)$ and reverse the reweighting:

$$\omega_i = \frac{o_i/\rho_i}{\sum_j o_j/\rho_j}, \qquad d = \mathbb{E}_\omega \rho, \qquad \tau_2 = \mathrm{KL}(\omega\|\mu).$$

Since $o_i = \rho_i \omega_i / d$, the lower bound in (37) gives

$$\tau_2 \le \frac{d}{\underline{\rho}} \mathrm{KL}(o\|\nu) \le \frac{d}{\underline{\rho}}\tau. \tag{41}$$

The policy $o$ can have zero coordinates and need not exhaust its budget. Thus this step includes saturated utility optima.

Let $\alpha = \overline{\rho} d/(\underline{\rho} c)$. Because $d \geq \underline{\rho}$ and $c \leq \overline{\rho}$, we have $\alpha \geq 1$, and (40)–(41) imply $\tau_2 \leq \alpha \tau_1$. Mixing feasible policies and using convexity of KL shows that $V_\mu$ is nondecreasing and concave. Since $V_\mu(0) \geq 0$,

$$V_\mu(\tau_2) \leq V_\mu(\alpha \tau_1) \leq \alpha V_\mu(\tau_1). \tag{42}$$

Nonnegative utility and (38) now yield

$$\begin{aligned} V_\nu(\tau) = \frac{\mathbb{E}_\omega(\rho f)}{d} &\leq \frac{\overline{\rho}}{d} V_\mu(\tau_2) \\ &\leq \frac{\overline{\rho}^{\,2}}{\underline{\rho}\, c} A \mathbb{E}_{\bar{q}} f, \\ \mathbb{E}_q f = \frac{\mathbb{E}_{\bar{q}}(\rho f)}{c} &\geq \frac{\underline{\rho}}{\overline{c}} \mathbb{E}_{\bar{q}} f. \end{aligned}$$

Combining these inequalities gives (39). The normalization factors $c$ and $d$ cancel.

For the remaining budgets, take the finite value $t_* = \max_i \log(1/\mu_i)$. Every point mass is feasible under $\mu$ at budget $t_*$. In particular, each $\delta_j$ with $j \in \mathcal{X}^\star$ is reward-optimal, so the premise implies

$$\max_i f_i = V_\mu(t_*) \leq A f_j \qquad (j \in \mathcal{X}^\star).$$

If $\tau \geq -\log \nu(\mathcal{X}^\star)$, every reward-optimal $q$ is supported on $\mathcal{X}^\star$. Consequently, $V_\nu(\tau) \leq \max_i f_i \leq A \mathbb{E}_q f$. This covers saturated budgets, constant rewards, and $\tau = \infty$. At $\tau = 0$, the only feasible policy is $\nu$, and the conclusion follows from $AR^2 \geq 1$. □

Theorem 1 supplies (38) with $A = C_0 \kappa(\beta)$ for the population BT reward. Proposition 9 therefore proves Corollary 5. The guarantee also holds with the refined factor from Corollary 7 in place of $C_0 \kappa(\beta)$.

## F Details of the Rare-Alternative Limit

Our instance adapts the idea behind Gölz et al. (2025, Theorem 3), in which a minority strongly values one alternative and comparisons compress this value. We first explain the limit in Proposition 6 in four steps and then give the formal arguments.

**The fitted reward prefers $y$.** As $\varepsilon \downarrow 0$, almost every comparison involving $x$ or $y$ is against $b$. Against $b$, the minority prefers $x$ with probability $\sigma(\beta)$ and the other users are indifferent. The win score of $x$ therefore approaches $p[\sigma(\beta) - 1/2] = (p/2)\tanh(\beta/2)$. Every user prefers $y$ to $b$ with probability $\sigma(\beta d)$, so the win score of $y$ approaches $(1/2)\tanh(\beta d/2)$. The threshold $d_0(p)$ is the utility at which these two limits coincide. Since $d > d_0(p)$ and rewards follow win scores, the fitted reward ranks $y$ above $x$. Normalize $r_b = 0$. Almost all opponents are $b$, so $\sigma(r_x) - 1/2$ and $\sigma(r_y) - 1/2$ approach the limiting win scores above. The paragraph on convergence of the fitted reward below proves this limit. Hence

$$r_x \longrightarrow \beta d_0(p), \qquad r_y \longrightarrow \beta d. \tag{43}$$

**Mean utility prefers $x$.** Strict convexity of $\operatorname{artanh}$ gives $d_0(p) < p$, so the interval $(d_0(p), p)$ is nonempty and $f_x = p > d = f_y$. The minority's large utility for $x$ is compressed by the sigmoid, whereas the small common utility of $y$ is reflected almost linearly in comparisons.

**The KL budget reaches both rare alternatives at the same cost.** Let $\bar{w}_\delta$ be the largest mass that a policy in the KL ball can place on an event of reference probability $\delta$. For small $\delta$, $\bar{w}_\delta \log(1/\delta) \to \tau$. Hence $\bar{w}_\varepsilon / \bar{w}_{2\varepsilon} \to 1$, and placing mass on $x$ alone costs essentially as much as placing it on the pair $\{x, y\}$. Since $b$ has zero utility, both $V$ and $\mathbb{E}_q f$ are determined by how this rare mass is split. Rarity makes the reward policy split it almost as an argmax would. If its multiplier $\lambda$ stayed bounded, both rare masses would be $O(\varepsilon)$ and the KL divergence would vanish. Hence $\lambda$ diverges as $\varepsilon \downarrow 0$. Because $\mu_x = \mu_y$, the tilt (23) gives $q_x/q_y = e^{-\lambda(r_y - r_x)} \to 0$, and the reward policy puts almost all rare mass on $y$. The utility optimum puts it on $x$, and the ratio of their utilities approaches $p/d$.

**The limit equals $\kappa(\beta)$.** The ratio $p/d$ approaches $p/d_0(p)$ as $d \downarrow d_0(p)$. For small $p$, $d_0(p) \approx (2/\beta)p\tanh(\beta/2)$, so $p/d_0(p) \to \kappa(\beta)$. In terms of win scores, each unit of the mean utility of $x$ adds $\tanh(\beta/2)/2$ to its win score. For small $d$, each unit of the mean utility of $y$ adds about $\beta/4$ to its win score. The ratio of these two rates is $\kappa(\beta)$. It is the ratio between the chord slope and the slope at zero of $\psi$ in Lemma 3.

**Convergence of the fitted reward.** In the instance (21), the population comparison probabilities in (1) are

$$p_{xb} = p\sigma(\beta) + (1-p)/2, \qquad p_{yb} = \sigma(\beta d),$$
$$p_{xy} = p\sigma(\beta(1-d)) + (1-p)\sigma(-\beta d).$$

After division by row weights, the score equations for $x$ and $y$ are

$$0 = (1-2\varepsilon)[\sigma(r_x) - p_{xb}] + \varepsilon[\sigma(r_x - r_y) - p_{xy}],$$
$$0 = (1-2\varepsilon)[\sigma(r_y) - p_{yb}] + \varepsilon[\sigma(r_y - r_x) - p_{yx}].$$

At $\varepsilon = 0$, these equations have the solution in (43), with Jacobian $\mathrm{diag}(\sigma'(\beta d_0(p)), \sigma'(\beta d))$. Both diagonal entries are positive. The implicit function theorem and uniqueness of the population fit in the gauge $r_b = 0$ (Appendix A) give a continuous solution for small positive $\varepsilon$. The remaining row equation follows from the weighted sum of the score residuals being zero. In particular, $r_y - r_x$ stays strictly positive, while $f_x = p > d = f_y$.

**The KL ball permits comparable rare-event mass.** Let $\bar{w}_\delta$ be the largest probability $v$ satisfying $\mathrm{kl}(v\|\delta) \le \tau$, where kl is binary relative entropy. By data processing, it is the largest mass that a feasible policy can place on an event of reference probability $\delta$. For small $\delta$, this constraint is an equality at $\bar{w}_\delta < 1$. Writing $H(v) = -v\log v - (1-v)\log(1-v)$, we have

$$\mathrm{kl}(v\|\delta) = v\log(1/\delta) - H(v) - (1-v)\log(1-\delta). \tag{44}$$

It follows first that $\bar{w}_\delta \to 0$ and then that

$$\bar{w}_\delta \log(1/\delta) \longrightarrow \tau, \qquad \frac{\bar{w}_\varepsilon}{\bar{w}_{2\varepsilon}} \longrightarrow 1. \tag{45}$$

The leading KL cost of reaching one rare alternative therefore matches that of reaching either member of the rare pair.

**Reward optimization allocates the rare mass to $y$.** For small $\varepsilon$, the budget is interior because concentrating on $y$ costs $\log(1/\varepsilon) > \tau$. Its tilt parameter $\lambda_\varepsilon$ must tend to infinity. Otherwise, along a bounded subsequence of multipliers, both rare policy masses would be $O(\varepsilon)$ and the KL divergence would tend to zero. Equation (43) therefore gives

$$\frac{q_x}{q_y} = e^{-\lambda_\varepsilon(r_y - r_x)} \longrightarrow 0. \tag{46}$$

Let $w = q_x + q_y$. Relative-entropy data processing gives $w \le \bar{w}_{2\varepsilon}$, and hence

$$\mathbb{E}_q f \le [d + o(1)]\bar{w}_{2\varepsilon}. \tag{47}$$

The utility optimum can assign mass $\bar{w}_\varepsilon$ to $x$ and retain the reference proportions elsewhere. This is feasible and gives $V \ge p\,\bar{w}_\varepsilon$. Combining with (45),

$$\liminf_{\varepsilon \downarrow 0} \mathrm{dist} \ge \frac{p}{d}.$$

*Completion of the lower bound.* For every fixed $p$ and $d \in (d_0(p), p)$, the preceding family gives $D^\star(\beta,\tau) \ge p/d$. First let $d \downarrow d_0(p)$ and then let $p \downarrow 0$. We obtain

$$D^\star(\beta,\tau) \ge \lim_{p\downarrow 0} \frac{\beta p}{2\,\mathrm{artanh}(p\tanh(\beta/2))} = \kappa(\beta).$$

For $\tau = \infty$, small positive $\varepsilon$ makes $y$ the unique reward maximizer and $x$ the unique mean-utility maximizer. Their distortion is exactly $p/d$, giving the same lower bound. At $\tau = 0$, both policies are necessarily $\mu$ and the distortion is one. □

**Relation to the algorithm-independent lower bounds.** For every $\tau > 0$ and $\beta > 0$, Oko et al. (2026, Theorem 5.2) construct finitely many instances without distribution mismatch such that, on an instance drawn uniformly at random, the expected distortion of any algorithm is at least $\kappa(\beta) - \epsilon$ for an arbitrary $\epsilon > 0$. The reward policy is one such algorithm, and some instance attains at least the expected value. Hence $D^\star(\beta, \tau) \geq \kappa(\beta) - \epsilon$ for every $\epsilon > 0$, which gives (7). For $\tau = \infty$, Gölz et al. (2025, Theorem 3) show that no voting rule can guarantee distortion below $\kappa(\beta)$ for large $m$. Both results assume that each user provides a single comparison, as in our model, where every comparison draws a new user. With several comparisons per user, Gölz et al. (2025, Theorem 3) prove the bound only for rules that put at most $1/m$ mass on a Condorcet loser. The explicit construction above establishes $D^\star(\beta, \tau) \geq \kappa(\beta)$ for the reward policy at every $0 < \tau \leq \infty$.

For $0 < \tau < \infty$, to strengthen the lower-limit result to equality, write $w = q_x + q_y$ and apply the KL chain rule:

$$\begin{aligned}\tau = \; & \mathrm{kl}(w \| 2\varepsilon) \\ & + w \, \mathrm{KL}\left((q_x/w, q_y/w) \,\|\, (1/2, 1/2)\right).\end{aligned}$$

The second term is at most $w \log 2$ and tends to zero because $w \leq \bar{w}_{2\varepsilon} \to 0$. Equation (44) then gives $w \log(1/(2\varepsilon)) \to \tau$, so $w/\bar{w}_{2\varepsilon} \to 1$. Combining this with (46), we obtain $\mathbb{E}_q f/(d\,\bar{w}_{2\varepsilon}) \to 1$. Every feasible policy puts at most $\bar{w}_{2\varepsilon}$ mass on the rare pair. Therefore

$$p\,\bar{w}_\varepsilon \leq V \leq p\,\bar{w}_{2\varepsilon}.$$

Using (45) proves $\mathrm{dist} \to p/d$ for fixed $p, d$.

## G Small Budgets for a Fixed Reference

The construction in Section 5 varies $\mu$ along a sequence of instances. For a fixed instance, both $V$ and $\mathbb{E}_q f$ approach $\mathbb{E}_\mu f > 0$ as $\tau \downarrow 0$, so the distortion approaches one. In the construction, the alternatives with positive utility become arbitrarily rare under $\mu$, and this approach to one is not uniform across instances. This appendix quantifies the dependence on reference mass.

Let $\bar{f} = \mathbb{E}_\mu f > 0$ and $\Delta = \max f - \min f$. For every feasible $\pi$,

$$|\mathbb{E}_\pi f - \bar{f}| \leq \Delta \, \mathrm{TV}(\pi, \mu) \leq \Delta\sqrt{\tau/2}.$$

Applying this inequality to the utility-optimal and reward policies gives, whenever $\Delta\sqrt{\tau/2} < \bar{f}$,

$$\mathrm{dist} \leq \frac{\bar{f} + \Delta\sqrt{\tau/2}}{\bar{f} - \Delta\sqrt{\tau/2}}. \tag{48}$$

If $\mu_{\min} = \min_x \mu_x$, then $\bar{f} \geq \mu_{\min} \max_x f_x \geq \mu_{\min}\Delta$. Hence for $\sqrt{\tau/2} < \mu_{\min}$,

$$\mathrm{dist} \leq \frac{1 + \sqrt{\tau/2}/\mu_{\min}}{1 - \sqrt{\tau/2}/\mu_{\min}}.$$

This approach to one is uniform over nonnegative utility distributions for a fixed full-support reference. The reference probabilities in (21) vary with $\varepsilon$, which is why the uniform worst-case lower bound can hold at every positive budget.

## H A Finite Instance at $\beta = 64$ and $\tau = 8$

Proposition 6 describes a limit. The following instance has a background alternative, two rare alternatives, and three user types, with rational utility vectors and user weights. Each rare alternative has reference mass $e^{-8}$, so the budget $\tau = 8$ is exactly the cost of concentrating on one of them. The reward policy concentrates on the rare alternative with the higher win score, while the utility optimum concentrates on the other. NLHF selects the alternative that beats both others in pairwise comparisons, which here also has the highest mean utility. The instance shows that a uniform constant in Theorem 1 must exceed 1.9. It also separates the fitted scalar reward from the complete comparison game of NLHF at a fixed temperature and budget.

**Proposition 10.** *At $\beta = 64$ and $\tau = 8$, let $\mu = (1 - 2e^{-8}, e^{-8}, e^{-8})$. Consider the distribution*

| *Utility vector* | *Probability* |
|---|---|
| $(1, 0, 37/40)$ | $27/2000$ |
| $(0, 0, 1)$ | $133/10000$ |
| $(0, 1/2300, 0)$ | $2433/2500.$ |

*The unclipped RLHF policy uniquely selects alternative 2 and has distortion $1186225/19464 > 60.9 > 1.9\kappa(64)$. The NLHF policy uniquely selects alternative 3 and has distortion one.*

*Proof.* Write $w_1 = 27/2000$, $w_2 = 133/10000$, and $w_3 = 2433/2500$ for the probabilities of the three user types. The antisymmetric margin matrix $M_{ij} = 2p_{ij} - 1$ has entries

$$\begin{aligned} M_{21} &= -w_1 \tanh 32 + w_3 \tanh(8/575), \\ M_{31} &= -w_1 \tanh(12/5) + w_2 \tanh 32, \\ M_{32} &= w_1 \tanh(148/5) + w_2 \tanh 32 \\ &\quad - w_3 \tanh(8/575). \end{aligned}$$

These expressions satisfy the strict bounds

$$\begin{aligned} 39 \cdot 10^{-6} < M_{21} < 40 \cdot 10^{-6}, \\ 20 \cdot 10^{-6} < M_{31} < 21 \cdot 10^{-6}, \\ 0.0132 < M_{32} < 0.0134, \\ 0.000335 < e^{-8} < 0.000336. \end{aligned}$$

For completeness, the bounds can be obtained entirely with rational arithmetic. For $x > 0$ and $n + 2 > x$, put $S_n(x) = \sum_{j=0}^{n} x^j/j!$. Positivity of the series terms and a geometric bound on its tail give

$$S_n(x) < e^x < S_n(x) + \frac{x^{n+1}}{(n+1)!} \frac{1}{1 - x/(n+2)}.$$

Apply this inequality with $n = 200$ and $x \in \{8, 64, 24/5, 296/5, 16/575\}$, and use $\tanh v = (e^{2v} - 1)/(e^{2v} + 1)$.

Let $\varepsilon = e^{-8}$. Since $s = M\mu/2$,

$$\begin{aligned} s_2 - s_3 &= \frac{(1 - 2\varepsilon)(M_{21} - M_{31}) - 2\varepsilon M_{32}}{2} \\ &> 4 \cdot 10^{-6} > 0. \end{aligned}$$

Furthermore, $s_3 > 0 > s_1$. The population likelihood equations preserve these scores, and the fitted reward has the same strict ordering: $r_2 > r_3 > r_1$. Concentrating on alternative 2 costs $\mathrm{KL}(\delta_2 \| \mu) = 8$, so it is the unique reward optimum.

The mean utilities are

$$f = \left( \frac{27}{2000}, \frac{2433}{5750000}, \frac{2063}{80000} \right).$$

Thus alternative 3 is the unique utility maximizer and is also feasible at budget 8. Its utility divided by that of alternative 2 is exactly $1186225/19464$. The displayed exponential bounds also give $1186225/19464 > 1.9\kappa(64)$.

Finally, $M_{31} > 0$ and $M_{32} > 0$, so alternative 3 defeats both other alternatives. For any feasible $q \neq \delta_3$, $\delta_3^\top M q = M_{31} q_1 + M_{32} q_2 > 0$. The feasible response $\delta_3$ therefore excludes $q$ from the equilibria of the zero-sum game. Consequently, $\delta_3$ is the unique NLHF policy and attains the utility optimum. □

# I Additional Experimental Specification

**Controlled rare-alternative instances.** The grid contains $\beta \in \{4, 16, 64\}$, $\tau \in \{10^{-4}, 10^{-2}, 10^{-1}, 1\}$, and $L = -\log \varepsilon \in \{8, 32, 128, 256\}$, giving 48 instances. The reward

is refitted from the full comparison matrix at each point. We use arbitrary-precision arithmetic to represent the small reference probabilities and optimize the corresponding policies. Figure 1 shows the distortion and conditional rare-pair allocation at $\beta = 64$. For each budget, the reward policy allocates an increasing share of its rare mass to the alternative with lower mean utility as $L$ grows, and the utility optimum moves in the opposite direction. The value $0.909$ at $L = 256$ reflects the choice of $d$, which exceeds $d_0(p)$ by ten percent, while $p/d_0(p)$ is close to $\kappa$ at $p = 0.01$. Across the tested temperatures and budgets, the normalized distortion exceeds $0.9$ at $L \geq 128$, where the two valuable alternatives have very small reference mass. At the smallest tested depth, $L = 8$ gives $\varepsilon \approx 3.4 \times 10^{-4}$. For $\beta = 64$, distortion ranges from 1.2 to 2.3 across the four budgets.

**Utility proxies and panels.** A movie rating is mapped to utility by $(\text{rating} - 1)/4$. For KuaiRec, we use $\min\{\max\{\text{watch_ratio}, 0\}, 2\}/2$ as a utility proxy. We sample 40 MovieLens panels with 5 or 10 candidates and 60 KuaiRec panels with 5, 10, or 20 candidates. Each panel assigns equal weight to its users. Panel selection uses observation coverage, with 20 panels per candidate count and dataset. The natural-profile panel generator uses seed 2026091301. Candidate pools are ranked by full-data counts, with stable index ordering for ties. Candidates are sampled uniformly without replacement within a panel. For the MovieLens 1M release (GroupLens Research, 2003), the pool contains the 50 most-rated movies. Five- and ten-candidate panels require at least 100 and 50 complete users, respectively. For KuaiRec, the pool contains the 200 most-exposed videos, and every panel requires at least 500 complete users. Exposure counts sum `show_cnt` across the daily item records. We then retain users whose entire selected profile is observed. MovieLens five-candidate panels contain 201–767 users, and ten-candidate panels contain 91–241 users. KuaiRec panels contain 1,356–1,409 users for five candidates, 1,298–1,403 for ten, and 1,218–1,398 for twenty.

**Population fitting.** For each panel, the empirical comparison matrix is

$$p_{xy} = \frac{1}{n} \sum_{i=1}^{n} \sigma\big(\beta(u_{ix} - u_{iy})\big).$$

The win-score vector is $(p - 1/2)\mu$. We solve its population likelihood equations after fixing the first reward to zero. The Jacobian is obtained by differentiating the sigmoid, and the policy problems use the scalar tilt described in Appendix A. The simulations use these exact probabilities instead of sampling comparisons. The uniform and count-proportional references are also the comparison distributions. The empirical user weights are uniform within each selected complete panel.

**Parameter grid and aggregation.** Each panel uses a uniform reference and a reference proportional to the candidate's full-data rating or exposure count. In both cases, the reference is also the comparison distribution. We evaluate $\beta \in \{0.5, 2, 8, 32, 128\}$ and $\tau \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1, 2, 4, 8, \infty\}$, which gives 100 settings per panel and 10,000 settings in total. Dataset means give equal weight to every setting of every panel, so they describe the fixed evaluation grid. Within each dataset, the maximum distortion occurs at saturated budgets on a single panel, where the three policies attain the same value. Figure 2 averages references and budgets within each panel before averaging panels at each $\beta$. Figure 5 shows the dependence on the KL budget.

## J Clipping and Nash Comparison Policies

**Comparison objectives.** Table 1(a) compares reward, win-score and NLHF policies over the same full grid. The first two optimize scalar objectives with identical rankings and different score spacings. NLHF solves a KL-constrained zero-sum game on the complete population margin matrix. We select the equilibrium closest to $\mu$ in KL divergence, as specified below. Each objective is evaluated against the same utility oracle in the same KL ball.

**The exact clipping rule.** Oko et al. (2026) introduce clipping as a technical device that keeps the policy from spending its KL budget on outputs whose rewards lie outside the informative range. The clipping interval is determined by the fitted reward and the sampling distribution. For $0 < c \leq 1/316$,

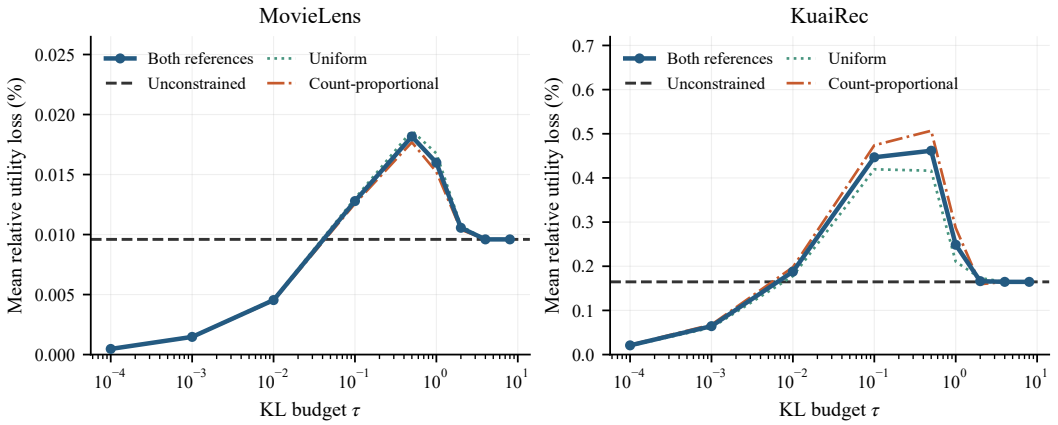


Figure 5: Mean relative utility loss of the unclipped reward policy at $\beta = 128$ in population simulations using observed utility proxies, on 40 MovieLens and 60 KuaiRec panels. The bold curve averages the two references within each panel and then averages panels, matching the main-text comparison. Thin curves separate the references. The horizontal dashed line gives the pooled mean at $\tau = \infty$. Vertical scales differ between panels.

Equation (7) of Oko et al. (2026) chooses the lower clipping level $r_{\min}$ to solve

$$\sum_y \mu_y \sigma(r_{\min} - r_y) = \frac{1}{2} - \frac{c^3}{16}, \tag{49}$$

$$\widetilde{r}_x = \max\{r_{\min}, \min\{r_x, r_{\min} + 2c\}\}.$$

The left-hand side is continuous and strictly increasing from zero to one, so $r_{\min}$ is unique. Translating all rewards by a constant translates $r_{\min}$ by that constant and leaves the policy unchanged. We solve the equivalent centered equation $\frac{1}{2}\sum_y \mu_y \tanh((r_{\min} - r_y)/2) + c^3/16 = 0$.

**Selection among clipped optima.** Clipping gives several alternatives the same maximal clipped reward, so the clipped objective can have many maximizers. Let $T = \arg\max_x \widetilde{r}_x$ and $b_T = -\log \mu(T)$. If $\tau < b_T$, the unique solution is the exponential tilt of $\widetilde{r}$ given in Appendix A. If $\tau \geq b_T$, every optimum is supported on $T$, and

$$\mathrm{KL}(q\|\mu) = \mathrm{KL}(q\|\mu(\cdot \mid T)) + b_T.$$

The conditional-reference rule returns $\mu(\cdot \mid T)$. Our main clipping comparison uses the lexicographic rule, which we call reward-based tie selection below. Within $T$, it maximizes $r$ at remaining budget $\tau - b_T$ relative to $\mu(\cdot \mid T)$. A separate optimistic bound replaces this secondary objective with $f$, giving the greatest utility among all clipped-reward optima. This bound uses utility information and is reported separately from the comparison policies.

**Clipping sensitivity.** Table 2 separates selection among clipped optima from the effect of clipping itself. On the settings where the clipped optimum is unique, clipping retains higher mean loss than the unclipped reward. On all 8,000 clipping settings, maximizing utility among clipped optima also leaves higher mean loss. Thus the loss increase persists both without ties and under the most favorable utility selection. The smaller values of $c$ in the theorem's sufficient range produce similar mean losses to $c = 1/316$.

To examine the wider parameter range in Figure 6, we additionally evaluate $c \in \{0.01, 0.03, 0.1, 0.3, 1\}$ using reward-based tie selection. These parameters lie beyond the cited theorem's sufficient range and are fixed before the comparison. Mean loss decreases over these evaluated widths, bringing the clipped policy closer to the unclipped reward's mean utility.

**Nash equilibrium computation.** Let $M_{xy} = p_{xy} - p_{yx}$ and $\mathcal{B} = \{\pi : \mathrm{KL}(\pi\|\mu) \leq \tau\}$. The population game of Gölz et al. (2025) is $\max_{q\in\mathcal{B}} \min_{\rho\in\mathcal{B}} q^\top M \rho$. Antisymmetry makes its value zero. Its equilibrium set is convex, and we select its unique minimizer of $\mathrm{KL}(q\|\mu)$.

Table 2: Clipping sensitivity on the $\beta \geq 1$ grid. Losses are mean relative utility losses in percent. The upper block uses 3,200 MovieLens and 4,800 KuaiRec settings (80 per panel). The lower block retains the settings where the $c = 1/316$ clipped optimum is unique. The utility tie rule maximizes utility among clipped optima. Reward-based selection is used unless another rule is specified. Paired intervals for the primary comparison appear in Table 1(b).

| Policy / clipping selection | MovieLens | KuaiRec |
|---|---|---|
| Full clipping grid: 3,200 / 4,800 settings | | |
| Unclipped RLHF | 0.0064 | 0.0871 |
| $c = 1/316$, reward ties | 0.2141 | 1.3089 |
| $c = 1/316$, conditional reference | 2.1451 | 11.8555 |
| $c = 1/316$, utility ties | 0.2103 | 1.2802 |
| $c = 1/632$, reward ties | 0.2151 | 1.3089 |
| $c = 1/1264$, reward ties | 0.2148 | 1.3091 |
| $c = 1$, reward ties | 0.0290 | 0.3168 |
| Unique clipped optima: 1,546 / 2,352 settings | | |
| Unclipped RLHF | 0.0045 | 0.1010 |
| $c = 1/316$ | 0.3697 | 2.3201 |

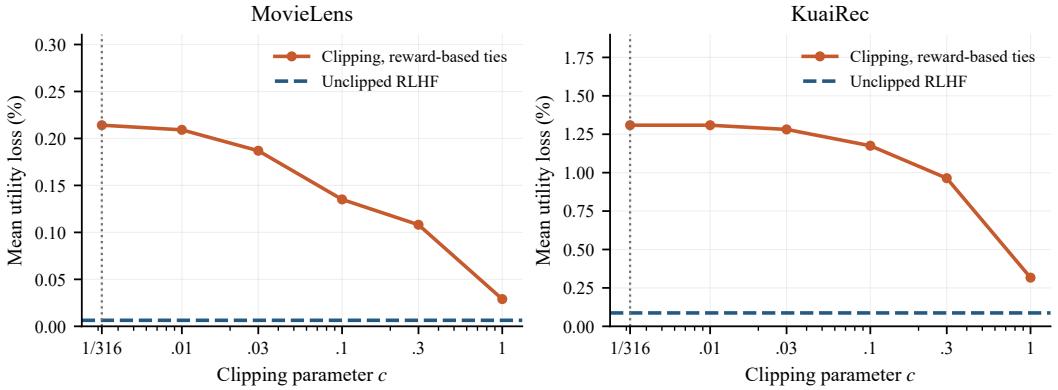


Figure 6: Clipping parameter sensitivity on the same 8,000 settings with $\beta \geq 1$. The solid curves use the exact threshold equation and reward-based tie selection. The dashed lines show the unclipped policy. The vertical dotted line marks $c = 1/316$, the largest value in the sufficient range of Oko et al. (2026, Theorem 4.1). Larger values are evaluated as parameter sensitivity beyond that range. Each point averages the two references, four temperatures and ten budgets within each panel, followed by averaging across panels.

At $\tau = 0$, the only feasible policy is $\mu$. For positive budgets, when the full-simplex equilibrium of least KL divergence is feasible, it is also the selected equilibrium for the constrained game. Otherwise, the KL constraint is active and the equilibrium has positive coordinates satisfying, for some $\eta > 0$,

$$q_i \propto \mu_i \exp((Mq)_i/\eta), \qquad \mathrm{KL}(q\|\mu) = \tau.$$

We solve these equations in log-probability coordinates using the analytic Jacobian. For each returned policy, the independent best-response value

$$g(q) = \max_{\rho \in \mathcal{B}} \rho^\top M q$$

measures its distance from equilibrium in margin units. The best-response problem is a scalar KL tilt of the objective $Mq$. At equilibrium $g(q) = 0$. This computation uses the observed population margin matrix directly, without fitting a scalar reward or multiplying the matrix entries by reference sampling probabilities.

**Panel-level uncertainty.** The paired bootstrap first computes each panel's mean loss difference over the two references, four temperatures with $\beta \geq 1$, and ten budgets. It then resamples 20 panels with

replacement within each candidate-count stratum. We use 5,000 replicates with seed 2026091317 and report percentile 95% intervals. Each replicate retains the complete grid within a selected panel. The intervals describe variation under this panel-sampling protocol. Table 1(b) reports the paired differences and intervals for reward-based tie selection at $c = 1/316$.

## K Population Experiments with a Changed Reference

Table 3: Reference changes on the observed utility profiles. All loss columns are mean relative losses in percent. The first unclipped column uses $\alpha = 0$. The second unclipped column and the three other policies use $\alpha = 1$, the same reference, budget, and utility oracle. Each direction and endpoint uses 12 settings per panel on 40 MovieLens and 60 KuaiRec panels, a separate grid from Table 1. Max. $R$ is the largest value of $R$ in Corollary 5 across panels.

| Dataset | Sampling → reference | Max. $R$ | Unclipped, $\alpha = 0$ | Unclipped, $\alpha = 1$ | NLHF | Clip $c = 1/316$ | Clip $c = 1$ |
|---|---|---|---|---|---|---|---|
| MovieLens | Uniform → counts | 2.02 | 0.0069 | 0.0064 | 0.0103 | 0.1951 | 0.0326 |
| MovieLens | Counts → uniform | 2.02 | 0.0065 | 0.0070 | 0.0108 | 0.2297 | 0.0404 |
| KuaiRec | Uniform → counts | 5.95 | 0.1133 | 0.1248 | 0.0817 | 1.4273 | 0.4288 |
| KuaiRec | Counts → uniform | 5.95 | 0.1316 | 0.1186 | 0.0740 | 1.2034 | 0.4873 |

To examine reference changes covered by Corollary 5, we retain the same 100 complete-user panels and observed utility profiles. Let $\mu$ be either the uniform or the count-proportional distribution, and let $\mu_{\mathrm{alt}}$ be the other one. We fix the comparison distribution at $\mu \otimes \mu$, fit its population reward, and optimize with reference

$$\nu_\alpha = (1 - \alpha)\mu + \alpha\mu_{\mathrm{alt}}, \qquad \alpha \in \{0, 1/2, 1\}.$$

The utility oracle and both players of the Nash game use the same $\mathcal{B}_{\nu_\alpha}(\tau)$. For clipped policies, the threshold in (49) is still computed against $\mu$, and the policy uses $\nu_\alpha$ in its constraint. We use reward-based tie selection for $c = 1/316$ and $c = 1$. The latter is a fixed wider clipping range.

The grid uses $\beta \in \{2, 32, 128\}$ and $\tau \in \{10^{-3}, 10^{-1}, 1, \infty\}$. The two sampling directions, three interpolation values, and 100 panels give 7,200 settings. We average with equal weights across panels and the complete grid within each direction. Table 3 compares the unchanged reference $\alpha = 0$ with the changed reference $\alpha = 1$.

Table 3 shows small changes in unclipped-RLHF loss after the reference switch. The largest shift in either direction is $0.013$ percentage points, while the maximum density-ratio spread reaches $R = 5.95$. The losses remain small even as the reference-transfer bound allows an $R^2$ factor. At the interpolated midpoint, median $R$ ranges from $1.28$ to $1.37$ across datasets and directions. The largest unclipped distortions remain $1.0039$ on MovieLens and $1.0618$ on KuaiRec throughout the grid.

The remaining columns of Table 3 compare objectives after the reference switch. In both directions, unclipped RLHF has lower mean loss than NLHF on MovieLens, while NLHF has lower mean loss on KuaiRec. Both clipping choices have higher mean loss than unclipped RLHF in every dataset–direction pair. Increasing $c$ from $1/316$ to $1$ reduces that difference. Thus the baseline comparisons retain the same dataset-level ordering after changing the reference.

## L Released Reward Models on Four-Answer Candidate Sets

**Data and selection.** We use the validation split of HelpSteer (Wang et al., 2024b). Grouping rows by exact prompt yields 320 groups with four distinct, nonempty responses. We retain groups whose four complete conversations fit within 1,024 tokens under both released reward-model tokenizers, leaving 248 eligible prompts. We sort these prompts lexicographically and sample 200 without replacement using seed 2026091322. We preserve the four original responses and their order. This selection uses prompt grouping and token lengths before reward scoring.

**Reward evaluation.** The two checkpoints are `Skywork-Reward-V2-Llama-3.1-8B` and `Skywork-Reward-V2-Qwen3-0.6B` (Liu et al., 2026). Each conversation contains one user prompt and its assistant response. We apply the released chat template without a system message

Table 4: Reward versus implied-win-score allocation on 200 HelpSteer prompts. All TV values are percentages. Mean 95% intervals use 5,000 whole-prompt bootstrap samples. Medians and 90th percentiles describe the distribution across prompts at each fixed budget.

| | Llama 3.1 8B | | | Qwen3 0.6B | | |
|---|---|---|---|---|---|---|
| $\tau/\log 4$ | Mean [95% CI] | Median | 90th pct. | Mean [95% CI] | Median | 90th pct. |
| 0 | 0.00 [0.00, 0.00] | 0.00 | 0.00 | 0.00 [0.00, 0.00] | 0.00 | 0.00 |
| 0.01 | 1.31 [1.21, 1.43] | 1.23 | 2.37 | 0.40 [0.34, 0.46] | 0.21 | 1.00 |
| 0.03 | 2.24 [2.06, 2.44] | 2.10 | 4.21 | 0.69 [0.58, 0.81] | 0.36 | 1.79 |
| 0.1 | 3.85 [3.51, 4.20] | 3.59 | 7.27 | 1.22 [1.01, 1.45] | 0.59 | 3.04 |
| 0.25 | 4.68 [4.28, 5.09] | 4.21 | 8.71 | 1.35 [1.14, 1.59] | 0.68 | 3.63 |
| 0.5 | 3.58 [3.18, 3.99] | 2.64 | 7.54 | 0.66 [0.56, 0.77] | 0.44 | 1.32 |
| 0.75 | 0.97 [0.83, 1.11] | 0.63 | 2.37 | 0.19 [0.15, 0.23] | 0.08 | 0.45 |
| 0.95 | 0.09 [0.08, 0.11] | 0.03 | 0.30 | 0.02 [0.01, 0.03] | 0.00 | 0.05 |
| 1 | 0.00 [0.00, 0.00] | 0.00 | 0.00 | 0.00 [0.00, 0.00] | 0.00 | 0.00 |

or generation prompt and retain the complete sequence. The longest selected inputs have 1,013 and 1,024 tokens, respectively. We evaluate the original weights in float32 on a CPU, with batches of eight for Llama and four for Qwen. The scalar output logits define $r$ at their released scale.

**Policy paths and uncertainty.** We use $\mu_i = 1/4$ and nine budgets from zero to $\log 4$, the budget at which a policy can concentrate on a single answer:

$$\frac{\tau}{\log 4} \in \{0, 0.01, 0.03, 0.10, 0.25, 0.50, 0.75, 0.95, 1\}.$$

For each prompt, we construct the implied scores in (8) and solve both KL-constrained objectives using the exponential tilts in Appendix A. We use arbitrary-precision arithmetic for the win scores and policy paths. At a saturated budget, we choose the conditional uniform distribution on the objective's maximizers. The primary summary averages TV over the seven interior budgets within each prompt and then weights the 200 prompts equally. We obtain 95% percentile intervals from 5,000 bootstrap samples of complete prompt paths, using seed 2026091323. The same resampled prompt indices are used for both models and all budgets. Table 5 separates this interior-budget average from threshold and maximum statistics over each complete prompt path.

**Distribution at a fixed budget.** Table 4 reports the mean, median and 90th percentile across prompts at every budget. A TV of $5\%$ means that $5\%$ of the answer probability mass must be reallocated to transform one policy into the other. At $\tau/\log 4 = 0.25$, Qwen's median TV is $0.68\%$ and its 90th percentile is $3.63\%$: the allocation differences are concentrated in a subset of candidate sets even at the peak of the mean curve.

Table 5: Summaries over the same 200 prompts as Table 4. Interior TV first averages the seven interior budgets per prompt. Brackets give 95% whole-prompt bootstrap intervals. The threshold row counts prompts whose maximum TV over all nine tested budgets exceeds $5\%$. The maximum row ranges over all prompt–budget pairs. Reward span is $\max_i r_i - \min_i r_i$ on the model's released scale.

| Statistic | Llama 3.1 8B | Qwen3 0.6B |
|---|---|---|
| Mean interior TV (%), 95% interval | 2.39 [2.20, 2.59] | 0.65 [0.55, 0.75] |
| Prompts with path maximum TV $> 5\%$ | 43% (86/200) | 6% (12/200) |
| Maximum prompt–budget TV (%) | 17.10 | 8.98 |
| Mean within-prompt reward span | 14.65 | 4.00 |

**Variation along the budget path.** Table 5 shows larger allocation differences for Llama by both the interior-budget average and the fraction of prompts crossing the TV threshold. At zero budget, both policies equal the reference. At $\tau = \log 4$, both select the same reward maximizer. The rise and fall in Figure 3 therefore locates the allocation effect at interior budgets, where score spacings affect the policy before concentration on a single answer.

## M Additional Related Work

**Distortion in social choice.** Distortion was introduced to measure how much average utility a voting rule loses when it observes only ordinal preferences instead of cardinal utilities (Procaccia & Rosenschein, 2006; Boutilier et al., 2015). With deterministic rankings, this loss can be severe. The Borda rule has unbounded distortion (Procaccia & Rosenschein, 2006), and even randomized rules cannot avoid distortion that grows with the number of alternatives (Boutilier et al., 2015). They also study expected-welfare objectives under probabilistic preference models. Probabilistic comparison models change this picture, because how often one alternative beats another carries information about the size of their utility gap (Luce, 1959; Plackett, 1975). Under Bradley–Terry comparisons, the distortion of Borda is bounded in $\beta$ (Gölz et al., 2025; Oko et al., 2026). Alipour & Goyal (2026) obtain a $2\kappa(\beta)$ bound for Borda and linear distortion for Copeland under the Plackett–Luce model when $\beta \geq 1$. These results concern the selection of a single winner, which corresponds to RLHF without a KL constraint.

The link to reward learning is that maximum-likelihood Bradley–Terry fitting ranks alternatives by their Borda scores when comparisons pool different hidden contexts (Siththaranjan et al., 2024). Gölz et al. (2025) use this equivalence to identify unconstrained RLHF with the Borda rule, and without distribution mismatch the relevant Borda score is the average win probability against a reference-drawn opponent. We start from this ranking property and supply the additional comparison that is needed once the KL constraint makes reward gaps matter.

**Utilitarian analysis of alignment methods.** Our closest prior work is that of Gölz et al. (2025) and Oko et al. (2026). Gölz et al. brought distortion to alignment, incorporated the KL constraint, and separated NLHF from RLHF. Oko et al. traced the exponential distortion of RLHF to distribution mismatch and proved optimal-order guarantees for a clipped reward without such mismatch. We show that the same order holds for the unclipped reward at every KL budget. Single-reward alignment has been criticized for pooling disagreeing users (Chakraborty et al., 2024; Conitzer et al., 2024), and recent work models diverse users explicitly (Sorensen et al., 2024). Other work studies different aspects of heterogeneous preferences. Chakraborty et al. (2024) learn a mixture of rewards to maximize the utility of the worst-off group. Ge et al. (2024) show that RLHF with linear reward models can violate axioms such as Pareto optimality and pairwise majority consistency. We keep average utility as the objective throughout, which permits a direct comparison with the distortion bounds of Gölz et al. (2025) and Oko et al. (2026) developed in the main text. Sorensen et al. (2024) distinguish representing a range of reasonable answers, adapting to a chosen perspective, and matching a population's distribution of preferences.

**Learning rewards and optimizing policies.** Christiano et al. (2017) learn a reward function from human comparisons of trajectory segments and train an RL policy with the learned reward. Ziegler et al. (2019) apply preference-based reward learning to language continuation and summarization, preceding the instruction-following pipeline of Ouyang et al. (2022). These methods separate the estimation of human preferences from optimization of the resulting scalar reward. Related pipelines include summarization and helpfulness training (Stiennon et al., 2020; Bai et al., 2022a;b), with deployed reward models trained from richer feedback (Cui et al., 2023; Wang et al., 2024a; Adler et al., 2024; Lambert et al., 2024). Annotators may weight attributes differently (Wu et al., 2023; Lightman et al., 2023), and Casper et al. (2023); Kaufmann et al. (2025) survey these pipelines. Relative-entropy policy search constrains information loss during policy updates and obtains exponential reweighting solutions (Peters & Schaal, 2007; Peters et al., 2010). Related approaches include entropy-regularized formulations (Neu et al., 2017; Fox et al., 2016), trust-region methods (Schulman et al., 2015; 2017), maximum-entropy methods (Haarnoja et al., 2018), and constrained variants that impose a hard divergence budget (Achiam et al., 2017). The resulting exponential tilt depends on reward gaps (Abdolmaleki et al., 2018), so two objectives with the same ranking can allocate probability differently within a KL ball. Many methods that never train an explicit reward model rest on the same reward-based formulation. In language-model alignment, direct preference optimization (DPO) reparameterizes a KL-regularized reward objective as a classification loss on preference pairs (Rafailov et al., 2023). Our analysis concerns the average utility attained by exact reward optimization within a given KL budget.

**Extensions of direct preference optimization.** Variants change the link function or drop the reference model (Gheshlaghi Azar et al., 2024; Zhao et al., 2023; Ethayarajh et al., 2024; Meng et al., 2024; Hong et al., 2024; Xu et al., 2024; Rafailov et al., 2024). Other methods use online or self-play updates (Xiong et al., 2024; Guo et al., 2024; Dong et al., 2024; Tajwar et al., 2024; Yuan et al., 2024). Nash-style methods include Munos et al. (2024); Swamy et al. (2024); Calandriello et al. (2024); Rosset et al. (2024), while Tang et al. (2024) provide a unified view.

**Preference objectives and games.** Gheshlaghi Azar et al. (2024) introduce a family of transformed preference objectives that includes identity preference optimization (IPO). Munos et al. (2024) instead learn pairwise preferences and optimize a policy against competing policies through a regularized Nash game. Swamy et al. (2024) develop self-play updates based on comparison win proportions. Calandriello et al. (2024) connect online IPO to self-play in a regularized preference game. These results explain relationships between preference objectives and optimization procedures. The utilitarian comparison of Gölz et al. (2025) evaluates scalar reward and Nash policies using the same welfare criterion. Our main theorem determines the worst-case order of the unclipped reward policy without distribution mismatch.

**Probabilistic choice and statistical ranking.** The Bradley–Terry model assigns each individual a latent-score description of pairwise choice (Bradley & Terry, 1952). The choice framework of Luce (1959) and the ranking model of Plackett (1975) supply related probabilistic descriptions of choices among several alternatives. Aggregating different users' comparison probabilities need not yield a single model of this form. In statistical ranking, Shah & Wainwright (2018) define an alternative's population Borda score as its probability of defeating a uniformly sampled opponent. Our score in (8) is a centered analogue weighted by the reference distribution. Siththaranjan et al. (2024) connect Bradley–Terry fitting with Borda aggregation over hidden contexts. Lemma 2 builds on the moment identity (9) to address the subsequent constrained allocation.