THEORETICAL TENSIONS IN RLHF: RECONCILING EMPIRICAL SUCCESS WITH INCONSISTENCIES IN SOCIAL CHOICE THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning from Human Feedback (RLHF) can violate fundamental social-choice axioms, including Pareto optimality, majority consistency, pairwise-majority consistency, and Condorcet consistency. Prior work relates these failures to the aggregation rule implicitly implemented by reward learning: unrestricted Bradley–Terry training recovers the Borda count, which satisfies Pareto optimality but not the other axioms. Restricting the reward model does not generally resolve the problem; linear and any fixed-degree polynomial reward models can fail even Pareto optimality. This appears at odds with the empirical success of RLHF and motivates us to ask when its axiomatic guarantees can be recovered. We first show that, with unrestricted rewards, replacing comparison frequencies by majority labels changes the induced rule from Borda to Copeland, thereby recovering all four axioms. For linear rewards, we characterize the regularized estimator as a projection of its unrestricted counterpart and identify how feature geometry and regularization determine whether these guarantees survive. For polynomial rewards, we establish a sharp distinction in the order of quantifiers: although every fixed degree admits an adverse candidate set, every fixed finite candidate set has an interpolation degree beyond which unrestricted aggregation is recovered. Above this threshold, standard RLHF recovers Borda and hence Pareto optimality, whereas majority-based RLHF recovers Copeland and satisfies all four axioms. Together, these results show how aggregation design and reward expressivity jointly determine the axiomatic behavior of RLHF. Experiments on UltraFeedback and Chatbot Arena confirm this division of labor: the gap between the two label choices opens only as a reward model's capacity approaches the per-prompt unrestricted fit, and practical reward models remain far from that regime.

1 INTRODUCTION

Large language models (LLMs), such as GPT-6 Astra (OpenAI, 2026) and Claude Fable 5.1 (Anthropic, 2026), perform strongly across tasks from code synthesis to mathematical reasoning (Bubeck et al., 2023; Chowdhery et al., 2023; Touvron et al., 2023; Team et al., 2023) and are increasingly deployed in domains once thought to require human judgment (Eloundou et al., 2024). The dominant method for aligning them with human values is Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Casper et al., 2023; Dong et al., 2024), which fits a reward model to human preference comparisons, typically under the Bradley–Terry (BT) model (Bradley & Terry, 1952), and then fine-tunes the LLM to generate responses with high reward.

A reward model aggregates the preferences of many labelers into one score, so it can be analyzed as a social choice function (Brandt et al., 2016). Four classical axioms (Section 2.1) require that a response preferred by every labeler be ranked higher (Pareto optimality, PO), that a majority's top choice be ranked first (majority consistency, MC), that the ranking formed by pairwise majorities be returned when it exists (pairwise majority consistency, PMC), and that a response beating every other by pairwise majority be ranked first (Condorcet consistency, CC).

RLHF fares poorly against them (Noothigattu et al., 2020; Siththaranjan et al., 2024; Ge et al., 2024a; Maura-Rivero et al., 2025; Liu et al., 2025). With unrestricted rewards, BT reward modeling

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.

Table 1: Axioms of the regularized BT reward estimator (✓: guaranteed; ✗: violated on some profile). Complete profiles; standard labels P_{ij} induce Borda, majority labels $\mathbb{1}[P_{ij} > 1/2]$ induce Copeland (no pairwise majority ties). Score-complete: $\mathrm{rank}[\mathbf{1}\ \Phi] = n$. Appendix C.4 extends the $\forall k\ \exists \mathcal{Y}$ row to every $\lambda > 0$. Shaded: our main technical result, where fixing the candidate set before the degree reverses the polynomial impossibility.

Reward model	Labels	Condition	PO	MC	PMC	CC	Source
Unrestricted	Standard	–	✓	✗	✗	✗	Siththaranjan et al. (2024)
	Majority	–	✓	✓	✓	✓	**Ours**
Linear	Standard	any features	✗	✗	✗	✗	Ge et al. (2024a)
	Standard	score-complete	✓	✗	✗	✗	**Ours**
	Majority	score-complete	✓	✓	✓	✓	**Ours**
Polynomial	Standard	$\forall k\ \exists \mathcal{Y}$	✗	✗	✗	✗	Hollender & Kraiczy (2026)
	Standard	$\forall \mathcal{Y}\ \exists k_*, k \geq k_*$	✓	✗	✗	✗	**Ours**
	Majority	$\forall \mathcal{Y}\ \exists k_*, k \geq k_*$	✓	✓	✓	✓	**Ours**

implements the Borda count and satisfies only PO (Siththaranjan et al., 2024), so a response that a majority prefers to every alternative can still be ranked below another. With linear rewards it can violate even PO (Ge et al., 2024a), reversing a comparison on which all labelers agree, and polynomial rewards do not help, since for every degree some candidate set makes BT violate PO (Hollender & Kraiczy, 2026). Nor can the policy step repair these failures: for any reference policy there is a profile on which no choice of the Kullback–Leibler (KL) coefficient restores MC, PMC, or CC (Appendix D.2). These are failures of the aggregation rule itself, not of finite data or optimization. This raises two questions:

Why does RLHF work so well in practice, and how can it be modified to satisfy these axioms?

Both works propose remedies that are hard to deploy. Ge et al. (2024a) replace the loss by leximax Copeland subject to Pareto optimality (LCPO), a combinatorial rule that selects a ranking for one fixed candidate set and therefore cannot serve as a training loss for a reward model. Hollender & Kraiczy (2026) recover PO for linear rewards in an idealized limit where comparisons cover every direction of the embedding space uniformly, which requires reweighting or collecting additional data.

We revisit these questions for the regularized BT estimator used in practice, focusing on reward modeling (the policy step is treated in Appendix D); Table 1 summarizes the results. For *unrestricted rewards*, replacing each pair's preference fraction by its majority label turns Borda into Copeland and recovers all four axioms (Section 3). Standard practice already makes this change incidentally: when each comparison has a single labeler (Ouyang et al., 2022), preference fractions are 0 or 1, standard BT coincides with the Copeland loss, and the reward is PMC and CC even if comparisons from different labelers form a cycle. For *linear rewards*, the estimator is the Bregman projection of the unrestricted one onto the feasible score space (Section 4); with score-complete features, which holds generically once the feature dimension is at least $n - 1$, the majority-label estimator satisfies all four axioms, and otherwise every comparison with a large enough unrestricted margin is preserved.

Main technical result: the order of quantifiers. For *polynomial rewards* the existing picture is pessimistic: for every degree k there is a candidate set on which BT violates even PO, and Hollender & Kraiczy (2026) conjecture that fixed-width, fixed-depth networks fail for the same reason. This suggests that no amount of capacity can secure even the weakest axiom. We show that the conclusion hinges on the order of quantifiers (Section 5). The negative result reads $\forall k\ \exists \mathcal{Y}$: the candidates are chosen adversarially after the degree. In RLHF, however, the candidate set of a prompt is given before the reward model is fitted, and the quantifiers reverse: for every candidate set with distinct features there is an interpolation degree $k_* \leq n - 1$, equal to $\min\{k : \binom{d+k}{k} \geq n\}$ for generic features, such that for all $k \geq k_*$ the polynomial estimator coincides with the unrestricted one. Beyond k_*, standard BT recovers Borda and hence PO, and the majority-label loss recovers Copeland and all four axioms. Capacity is therefore not futile: it decides whether a reward model reaches the rule its loss targets.

The loss chooses the target; capacity decides whether it is reached. The labels determine which rule a reward model converges to, and its expressivity how close it gets; a model inherits the axioms of its target only once it nearly realizes its unrestricted solution. We make this concrete with a *capacity*

ladder that interpolates between a single reward head shared by all prompts and a separate head per prompt (Section 6.2). On UltraFeedback, the gap between standard and majority labels on PMC grows from 0.019 to 0.981 with capacity, yet a model that must generalize to unseen prompts keeps it below 0.04. Practical reward models are therefore far enough from either unrestricted solution that the Borda–Copeland distinction is not what limits them, and changing the labels is necessary but not sufficient to secure the axioms.

Experiments confirm these predictions (Section 6). In simulation every reward-stage statement holds exactly. On UltraFeedback, where Borda and Copeland disagree on 18.3% of prompts, the unrestricted Copeland estimator always returns the pairwise majority ranking there and Borda never does; learned reward models, however, are far from either target, and the choice of labels changes no axiom by more than 0.01, even with a 7B encoder or end-to-end LoRA fine-tuning. On Chatbot Arena, Borda and Copeland differ only on near-ties, one reason why the axiomatic failures of standard RLHF may cost little in practice.

1.1 Related Work

A growing body of work studies RLHF through the lens of social choice theory (Conitzer et al., 2024; Dai & Fleisig, 2024; Mishra, 2023; Zhong et al., 2024; Park et al., 2024; Chakraborty et al., 2024; Ge et al., 2024b; Swamy et al., 2024); see Halpern et al. (2026) for a survey. Three position papers argue that axiomatic frameworks should guide RLHF design (Conitzer et al., 2024; Dai & Fleisig, 2024; Mishra, 2023), and recent work analyzes further properties of RLHF as an aggregation rule: robustness to cloned responses (Procaccia et al., 2025), distortion (Gölz et al., 2025), proportional representation (Kim et al., 2026), and calibration to annotator disagreement (Halpern et al., 2025).

Noothigattu et al. (2020) first showed that the BT maximum-likelihood estimator (MLE) in Eq. (2.1) can violate pairwise majority consistency; their counterexample varies the number of comparisons across pairs, exactly the imbalance excluded by Theorem 3.1. Other works replace reward modeling altogether: alignment via the maximal lottery is Condorcet consistent (Maura-Rivero et al., 2025), and so is Nash learning from human feedback (Liu et al., 2025). Closest to our unrestricted analysis, An et al. (2026) construct losses whose population optima implement the Borda, Copeland, or Kemeny rule. We instead analyze the regularized BT estimator on a finite candidate set, including restricted reward classes, and characterize when a reward model is expressive enough to realize the rule its loss targets. Further related work is discussed in Appendix F.

2 Preliminaries

Reward modeling. RLHF (Ouyang et al., 2022) proceeds in three steps: supervised fine-tuning, reward modeling, and policy optimization. We study the second; the third is treated in Appendix D. Given a prompt x, labelers compare pairs of responses, and a reward $r(x, \cdot)$ is fitted to the resulting dataset D_{comp} of winner–loser pairs $(y_{\text{w}}, y_{\text{l}})$ by maximum likelihood under the BT model (Bradley & Terry, 1952):

$$\min_{r} -\mathbb{E}_{(x,y_{\text{w}},y_{\text{l}})\sim D_{\text{comp}}}\big[\log \sigma\big(r(x, y_{\text{w}}) - r(x, y_{\text{l}})\big)\big], \tag{2.1}$$

where $\sigma(u) := 1/(1 + e^{-u})$. We omit the dependence on x when it causes no confusion.

Let $\mathcal{Y} = \{y_1, \ldots, y_n\}$ be a set of n distinct responses to a prompt, also referred to as *candidates* in social choice theory, and let $\mathcal{U} = \{1, \ldots, m\}$ be a set of m labelers, known as *voters*.

Definition 2.1 (Preference Profile)**.** Given any two distinct responses $y, y' \in \mathcal{Y}$, a human labeler $i \in \mathcal{U}$ may express a preference for either $y \succ y'$ or $y' \succ y$. The preference of labeler i is represented as a set of binary comparisons over the responses, denoted by τ_i[1]. The collective preference profile of the group is defined as $\boldsymbol{\tau} = (\tau_i)_{i\in\mathcal{U}}$.

In the classical setting, referred to as the *complete preference profile*, each voter provides a complete and transitive ranking; that is, τ_i corresponds to a permutation over $\mathcal{Y}$. We denote by $\mathcal{P}(y \succ y')$ the proportion of comparisons between y and y' in which y is ranked above y', and write $P_{ij} := \mathcal{P}(y_i \succ y_j)$. If $P_{ij} > 1/2$, we say that y_i defeats y_j under the pairwise majority rule. A profile need not

[1]For simplicity, we assume strict individual preferences. A tie, if present, can be handled by splitting the vote equally between the two responses, as in our experiments.

be representable by any BT model: when the pairwise majorities form a Condorcet cycle, such as $y_1 \succ y_2 \succ y_3 \succ y_1$, no reward vector respects all of them (Gehrlein, 2006), and such cycles arise with high probability (Liu et al., 2025).

2.1 RLHF AND SOCIAL CHOICE

Social choice theory studies how to aggregate individual preferences into a collective decision (Brandt et al., 2016).

Definition 2.2 (Social Choice Function)**.** A social choice function is a function $f : \mathcal{T}^m \to \mathbb{S}(\mathcal{Y})$, where $\mathcal{T}^m$ denotes the set of all possible preference profiles over m voters, and $\mathbb{S}(\mathcal{Y})$ denotes the set of rankings over the response set $\mathcal{Y}$.

A desirable social choice function may satisfy one or more of the following properties.

Definition 2.3. Let $\boldsymbol{\tau} = (\tau_i)_{i\in\mathcal{U}}$ be a preference profile over a set of responses $\mathcal{Y}$, and let $f(\boldsymbol{\tau})$ denote the aggregate ranking. A social choice function f may satisfy the following (desirable) properties:

1. Pareto Optimality (PO): If all comparisons between y_i and y_j that exist agree on $y_i \succ y_j$, i.e., $\mathcal{P}(y_i \succ y_j) = 1$, then the aggregate ranking must also place y_i above y_j.
2. Majority Consistency (MC): If there exists a response y^{mw}, called the majority winner, that is ranked in the top position by more than 50% of voters, then y^{mw} must be ranked first in the aggregate ranking.
3. Pairwise Majority Consistency (PMC): If there exists a ranking τ^*, called the pairwise majority consistent ranking, such that for all $y_i, y_j \in \mathcal{Y}$, $y_i \succ_{\tau^*} y_j$ if and only if a majority of voters prefer y_i over y_j, i.e., $\mathcal{P}(y_i \succ y_j) > 1/2$, then the social choice function must return this ranking, i.e., $f(\boldsymbol{\tau}) = \tau^*$.
4. Condorcet Consistency (CC): If there exists a response y^{cw}, called the Condorcet winner, that defeats every other response in pairwise majority comparisons, i.e., $\mathcal{P}(y^{\mathrm{cw}} \succ y_i) > 1/2$ for all $y_i \in \mathcal{Y} \setminus \{y^{\mathrm{cw}}\}$, then y^{cw} must be ranked first in the aggregate ranking.

Pareto optimality is a minimal requirement, ensuring that universally preferred responses are respected, and is satisfied by virtually all standard social choice functions. A desirable social choice function should also satisfy at least one of the properties (2)–(4) in Definition 2.3.

Reward modeling is a social choice function: the reward-modeling step is the map $s_2 : \mathcal{T}^m \to \mathbb{R}^n$,

$$s_2(\boldsymbol{\tau}) = \operatorname*{arg\,min}_{\boldsymbol{r}\in\mathbb{R}^n} - \sum_{(y_i \succ y_j)\in\boldsymbol{\tau}} \log \sigma(r_i - r_j) + \frac{\lambda}{2}\|\boldsymbol{r}\|_2^2, \qquad \lambda > 0, \tag{2.2}$$

and ranks $\mathcal{Y}$ by the resulting rewards, following the regularized formulation of Siththaranjan et al. (2024). The penalty is needed for s_2 to be well defined: the unregularized BT likelihood has a finite maximizer only if the directed win graph is strongly connected (Yan, 2016), which fails, for instance, whenever each pair is compared once and an empirical Condorcet winner exists. With $\lambda > 0$ the objective is strongly convex and coercive, so the minimizer is unique and finite; when a finite MLE exists, the ordering results below also hold for $\lambda = 0$.

Pairwise weights and the reward objective. When every pair is compared the same number N of times ($N = m$ for complete profiles, $N = 1$ with one labeler per comparison), dividing Eq. (2.2) by N gives, with λ replaced by λ/N,

$$F_{q,\lambda}(\boldsymbol{r}) = \sum_{i<j} \Big[q_{ij} \log\big(1 + e^{r_j - r_i}\big) + q_{ji} \log\big(1 + e^{r_i - r_j}\big) \Big] + \frac{\lambda}{2}\|\boldsymbol{r}\|_2^2, \tag{2.3}$$

with $q_{ij} = P_{ij}$. More generally, any $q_{ij} \in [0, 1]$ with $q_{ij} + q_{ji} = 1$ is a *balanced weight*; we consider

$$\textit{standard BT: } q_{ij} = P_{ij}, \qquad \textit{Copeland loss: } q_{ij} = \mathbb{1}[P_{ij} > 1/2] + \tfrac{1}{2}\,\mathbb{1}[P_{ij} = 1/2].$$

The *score* $b_i = \sum_{j\neq i} q_{ij}$ is then the Borda count BC_i or the Copeland score C_i (pairwise majority victories, half a point per tie), respectively. We also write $\widehat{b}_i(\boldsymbol{r}) = \sum_{j\neq i} \sigma(r_i - r_j)$ and $A_\lambda(\boldsymbol{r}) =$

$\sum_{i<j} \log(e^{r_i} + e^{r_j}) + \frac{\lambda}{2}\|\boldsymbol{r}\|_2^2$, so that $F_{q,\lambda}(\boldsymbol{r}) = A_\lambda(\boldsymbol{r}) - \langle b, \boldsymbol{r}\rangle$; D_{A_λ} denotes the Bregman divergence of A_λ.

Feasible score spaces. The likelihood is shift invariant, so the minimizer of $F_{q,\lambda}$ over $\mathbb{R}^n$ lies in $\mathcal{H} = \mathbf{1}^\perp$; let $H = I_n - \mathbf{1}\mathbf{1}^\top/n$. A reward model restricts the centered scores to a subspace $\mathcal{S} \subseteq \mathcal{H}$:

$$r^\star = \arg\min_{\boldsymbol{r}\in\mathcal{H}} F_{q,\lambda}(\boldsymbol{r}), \qquad r^{\mathcal{S}} = \arg\min_{\boldsymbol{r}\in\mathcal{S}} F_{q,\lambda}(\boldsymbol{r}). \tag{2.4}$$

Under balanced comparisons, $r^\star$ is $s_2(\boldsymbol{\tau})$ with penalty λ/N (Section 3); linear and polynomial models are particular choices of $\mathcal{S}$ (Sections 4–5). Throughout, λ penalizes candidate scores, not model coefficients.

3 Unrestricted Reward Models

Among the axioms of Definition 2.3, prior work shows that the reward-modeling step of RLHF satisfies only Pareto optimality. A fundamental result was established by Siththaranjan et al. (2024), who showed that the ℓ_2-regularized BT estimator implicitly implements the Borda count rule; the same ordering holds for the unregularized MLE whenever it exists. Specifically, the Borda count of a response y_i is $\mathrm{BC}_i = \sum_{k\neq i} P_{ik}$, and the reward estimator satisfies $r_i > r_j \Leftrightarrow \mathrm{BC}_i > \mathrm{BC}_j$. Since the Borda count satisfies Pareto optimality but violates properties (2)–(4), so does $s_2(\boldsymbol{\tau})$.

3.1 Conditions for RLHF to Satisfy Social Choice Axioms

We now provide the conditions under which the regularized reward estimator satisfies the desirable properties outlined earlier. Since the standard BT loss fails, we replace the preference fractions P_{ij} in Eq. (2.3) by general balanced weights q_{ij} and ask which weights produce a well-behaved ranking.

Theorem 3.1. *Let $q_{ij} \in [0,1]$ satisfy $q_{ij} + q_{ji} = 1$ for every pair, and let $b_i = \sum_{j\neq i} q_{ij}$. For every $\lambda > 0$, the unrestricted estimator $r^\star$ in Eq.* (2.4) *exists uniquely and satisfies*

$$r_i^\star > r_j^\star \quad \Longleftrightarrow \quad b_i > b_j.$$

The balance condition matters: with unequal numbers of comparisons across pairs the fitted order need not follow any score of this form (Noothigattu et al., 2020). Theorem 3.1 reduces the analysis of the reward ranking to the design of q_{ij}, and $\lambda > 0$ makes the estimator finite and unique without changing its order. With $q_{ij} = P_{ij}$ it recovers the Borda result above; if the score vector b satisfies an axiom among (2)–(4), so does the reward.

Copeland Score. A natural choice is the majority indicator of the Copeland loss in Section 2, $q_{ij} = \mathbb{1}[P_{ij} > 1/2]$ when there are no pairwise majority ties. Then $b_i = C_i$, the number of pairwise majority victories of y_i, which is the classical Copeland score. In Appendix A.2, we provide a formal proof demonstrating that the Copeland score satisfies the required social choice axioms.

We refer to Eq. (2.3) with these weights as the Copeland reward loss. Some datasets already aggregate annotator judgments into such labels (Wang et al., 2024; Köpf et al., 2023); for them, Theorem 3.1 provides a theoretical justification.

3.2 Why Does RLHF Achieve Remarkable Success?

To reconcile the Borda failures with the empirical success of RLHF, we revisit the structure of the preference profiles that arise in practice. One possible explanation is that only two responses are compared per prompt, which rules out cycles. This holds for some public pairwise datasets, but it does not describe the data collection behind InstructGPT and its successors. As Ouyang et al. (2022) report, "We present labelers with anywhere between $K = 4$ and $K = 9$ responses to rank. This produces $\binom{K}{2}$ comparisons for each prompt shown to a labeler." A single labeler's ranking is transitive, but once the $\binom{K}{2}$ comparisons of a prompt are judged pair by pair, or by different labelers, they can form a cycle.

The key observation is that each comparison is typically labeled by a single labeler, since assigning two labelers to every comparison would double the annotation cost. Every pair then carries exactly

one vote, so the standard BT weight $q_{ij} = P_{ij} \in \{0, 1\}$ coincides with the majority indicator $\mathbb{1}[P_{ij} > 1/2]$. In other words, the standard BT loss *is* the Copeland reward loss, and the Borda and Copeland scores are identical. By Theorem 3.1 and Appendix A.2, the regularized reward is therefore pairwise majority consistent and Condorcet consistent in this regime, including when comparisons judged by different labelers form a cycle. Pareto optimality and majority consistency are not implied: a single labeler's cyclic comparisons $y_1 \succ y_2 \succ y_3 \succ y_1$ give $r_1 = r_2 = r_3$ although every available comparison favors y_1 over y_2, and a majority winner need not be defined when individual preferences are cyclic.

We stress that this agreement is largely incidental. It comes from the cost constraint that limits each comparison to one labeler, not from any property of the BT objective. Once several labelers judge the same comparison, P_{ij} becomes a vote share, the Borda and Copeland scores separate, and the Borda failures described above can reappear. Developers who collect multiple judgments per comparison should therefore aggregate them deliberately, for example with the Copeland reward loss, rather than feeding raw vote counts to the BT likelihood.

4 Linear Reward Models

The unrestricted analysis treats candidate rewards as independent. A shared linear model instead couples them through fixed features $\phi_i = \phi(x, y_i) \in \mathbb{R}^d$: $r_\theta(y_i) = \langle \theta, \phi_i \rangle$. Let $\Phi = (\phi_i^\top)_{i=1}^n$ and $\mathcal{S}_\Phi = \mathrm{col}(H\Phi) \subseteq \mathcal{H}$. We work with centered scores $r = H\Phi\theta$, which preserve all linear reward differences. A strict ranking is *linearly feasible* if some θ induces all its comparisons strictly. Except where specified, the results below allow arbitrary complete strict preference profiles. With balanced weights q and the objective in Eq. (2.3), the linear estimator is the restricted estimator of Eq. (2.4) with $\mathcal{S} = \mathcal{S}_\Phi$:

$$r^{\mathrm{lin}} = \operatorname*{arg\,min}_{r \in \mathcal{S}_\Phi} F_{q,\lambda}(r), \qquad \lambda > 0, \tag{4.1}$$

to be compared with the unrestricted $r^\star$. Since the penalty acts on centered scores, this is equivalent to fitting an intercept and penalizing the resulting scores.

Theorem 4.1 (Projected aggregation)**.** *The estimator r^{lin} is finite and unique, and $r^\star$ orders the candidates by b (Theorem 3.1). The linear estimator satisfies*

$$\Phi^\top \left(\widehat{b}(r^{\mathrm{lin}}) - b + \lambda r^{\mathrm{lin}} \right) = 0, \qquad r^{\mathrm{lin}} = \operatorname*{arg\,min}_{r \in \mathcal{S}_\Phi} D_{A_\lambda}(r, r^\star), \tag{4.2}$$

with A_λ and D_{A_λ} as in Section 2. If $\mathrm{rank}[\mathbf{1}\ \Phi] = n$, then $\mathcal{S}_\Phi = \mathcal{H}$ and $r^{\mathrm{lin}} = r^\star$.

Thus linear BT projects the unrestricted fitted rewards, rather than Euclidean-projecting the voting scores. With a score-complete representation, standard BT recovers Borda: PO holds, but MC, PMC, and CC can fail. The Copeland loss instead recovers the Copeland weak order and satisfies all four axioms on complete profiles without majority ties. Restricted features need not preserve either order.

Proposition 4.1 (A global recovery bound)**.** *Let $\delta_\Phi = \mathrm{dist}(r^\star, \mathcal{S}_\Phi)$ and $\varepsilon_\Phi = \sqrt{1 + n/(4\lambda)}\, \delta_\Phi$. Then*

$$\|r^{\mathrm{lin}} - r^\star\|_2 \le \varepsilon_\Phi. \tag{4.3}$$

Every comparison whose unrestricted reward gap exceeds $\sqrt{2}\, \varepsilon_\Phi$ is preserved. In particular, a unique Borda or Copeland winner is recovered if its smallest reward gap exceeds this threshold.

The bound needs no local Hessian assumption, and it applies to deterministic comparisons, including empirical Condorcet winners, because the positive penalty guarantees existence.

Feasibility and the scope of the guarantees. The geometry of the features matters independently of the loss. A response can be uniquely top-ranked by a linear reward exactly when its feature is an exposed vertex of the feature convex hull and is not shared by another candidate. Even when the majority ranking is feasible, a fixed positive penalty does not automatically make majority-label fitting PMC; a sufficiently small, explicitly bounded penalty does. For standard BT, the four-candidate example in Appendix B rejects a majority and Condorcet winner despite agreement between the Borda and majority rankings. Conversely, if the centered score space is one-dimensional and every voter ranking is linearly feasible, regularized BT satisfies all four axioms for every $\lambda > 0$, including unanimous profiles. These are statements about specified feature and preference domains, not a monotonicity claim about model expressivity. All proofs, counterexamples, the unregularized limit, and the feasibility-aware LCPO alternative of Ge et al. (2024a) are given in Appendix B.

5 POLYNOMIAL REWARD MODELS

Does increasing the expressivity of the reward model resolve its axiomatic failures? Hollender & Kraiczy (2026) show that bounded-degree polynomial rewards can still violate Pareto optimality. We complement this worst-case perspective by fixing the candidate set and studying how the induced aggregation rule changes with the polynomial degree.

Fix a prompt and distinct response features $\phi_i = \phi(x, y_i) \in \mathbb{R}^d$, $i \in [n]$, with $n \geq 2$. For an integer $k \geq 1$, let $\psi_k(z) = (z^\alpha)_{|\alpha| \leq k}$ contain all monomials of total degree at most k, including the constant. The reward is $r_\theta(y_i) = \langle \theta, \psi_k(\phi_i) \rangle$. Write $V_k = (\psi_k(\phi_i)^\top)_{i=1}^n$. The feasible centered scores form $\mathcal{S}_k = \mathrm{col}(HV_k) \subseteq \mathcal{H}$, and with the objective of Eq. (2.3) the degree-k estimator is

$$r^{(k)} = \operatorname*{arg\,min}_{r \in \mathcal{S}_k} F_{q,\lambda}(r), \qquad \lambda > 0, \tag{5.1}$$

the restricted estimator of Eq. (2.4) with $\mathcal{S} = \mathcal{S}_k$. As before, the penalty acts on candidate rewards and is not coefficient weight decay.

Theorem 5.1 (Projection and interpolation)**.** *The estimator $r^{(k)}$ exists uniquely, and $r^\star$ orders the candidates by b (Theorem 3.1). Moreover,*

$$V_k^\top \left(\widehat{b}(r^{(k)}) - b + \lambda r^{(k)} \right) = 0, \qquad r^{(k)} = \operatorname*{arg\,min}_{r \in \mathcal{S}_k} D_{A_\lambda}(r, r^\star). \tag{5.2}$$

Define the interpolation degree $k_* := \min\{k : \mathrm{rank}(V_k) = n\}$. Then

$$k_* \leq n - 1, \qquad k_* = \min \left\{ k : \binom{d+k}{k} \geq n \right\} \text{ for generic features.} \tag{5.3}$$

Here generic means outside a proper algebraic exceptional set. For every $k \geq k_*$, $r^{(k)} = r^\star$.

Thus polynomial reward fitting interpolates between a constrained aggregation rule and unrestricted Borda or Copeland. The threshold depends on feature geometry, not only on the number of coefficients; collinear features attain the worst-case value $k_* = n - 1$. Below this threshold, exact recovery can still occur for individual profiles. The following bound gives a sufficient condition without local curvature assumptions.

Proposition 5.1 (Recovery before interpolation)**.** *Let $\delta_k = \mathrm{dist}(r^\star, \mathcal{S}_k)$ and $\varepsilon_k = \sqrt{1 + n/(4\lambda)}\, \delta_k$. Then*

$$\|r^{(k)} - r^\star\|_2 \leq \varepsilon_k, \qquad D_{A_\lambda}(r^{(k+1)}, r^\star) \leq D_{A_\lambda}(r^{(k)}, r^\star). \tag{5.4}$$

Every comparison with $r_i^\star - r_j^\star > \sqrt{2}\, \varepsilon_k$ is preserved by $r^{(k)}$. In particular, a unique winner is preserved when its smallest reward gap exceeds this bound.

Axiomatic implications. For complete strict preference profiles and $k \geq k_*$, standard polynomial BT inherits Borda's PO guarantee, but can still violate MC, PMC, and CC. In contrast, the polynomial Copeland loss satisfies all four axioms when pairwise majority ties are absent. Below k_* these guarantees require additional conditions, such as Proposition 5.1. Increasing the degree improves the objective, but does not by itself establish an axiom. Worst-case PO failure persists at every fixed degree and every fixed $\lambda > 0$ when the candidate set may vary. Quadratic rewards can already place any distinct candidate uniquely first, while a three-candidate example violates MC, PMC, and CC for every $k \geq 1$ and $\lambda > 0$.

These statements concern the reward stage on the fixed candidate set; transfer to the KL-regularized policy is discussed in Appendix D. Appendix C provides all proofs, the counterexamples, and the distinction between finite-set interpolation and uniform-coverage guarantees on unseen comparisons.

6 EXPERIMENTS

We test the reward-stage theory in three settings: a controlled simulation where every quantity is known, reward models trained on real LLM representations, and the human votes of Chatbot Arena. Full setups and additional results are in Appendix E.

Table 2: Reward models on UltraFeedback (5,554 held-out prompts; the four annotation aspects act as voters). Axiom columns are rates among applicable prompts; "PMC (con.)" restricts to the 52 applicable prompts on which Borda and Copeland disagree; Kendall is the distance to the target order (Borda for standard labels, Copeland for majority labels). Full results are in Tables 6 and 7.

Reward model	Labels	Pair acc.	PO	MC	PMC	CC	PMC (con.)	Kendall
Unrestricted	Standard	0.998	1.000	0.995	0.978	0.995	0.000	0.000
	Majority	1.000	1.000	1.000	1.000	1.000	1.000	0.000
Frozen 1.5B, linear ($d = 1536$)	Standard	0.760	0.718	0.612	0.255	0.584	0.115	0.271
	Majority	0.759	0.711	0.615	0.254	0.586	0.135	0.272
Frozen 7B, linear ($d = 3584$)	Standard	0.824	0.864	0.707	0.363	0.659	0.173	0.215
	Majority	0.822	0.854	0.710	0.365	0.662	0.154	0.218
LoRA 1.5B	Standard	0.824	0.849	0.707	0.366	0.665	0.192	0.216
	Majority	0.822	0.844	0.710	0.364	0.668	0.250	0.217

6.1 Controlled simulation

We draw $n = 10$ candidates with features in $\mathbb{R}^2$ and $m = 5$ linearly feasible voters, and fit both losses at $\lambda = 0.1$ under generic features ($k_* = 3$) and collinear features ($k_* = 9$). Over $13{,}000$ constrained fits, every reward-stage statement holds on every instance: the unrestricted estimator orders candidates by b, the recovery bound of Propositions 4.1 and 5.1 is never violated, and $r^{(k)} = r^\star$ for all $k \geq k_*$ (Table 4 in Appendix E.1). Labels and expressivity play the separate roles the theory assigns them (Figure 3 in Appendix E.1). Under majority labels, every axiom improves with the degree and reaches 1 at k_*. Under standard labels, a higher degree makes matters *worse*: MC falls from 0.854 to 0.714 and CC from 0.832 to 0.719 as the fit converges to Borda. With collinear features at $k = 1$, even PO holds on only 0.134 and 0.114 of the applicable profiles under the two losses, the finite-sample counterpart of the impossibility result of Ge et al. (2024a).

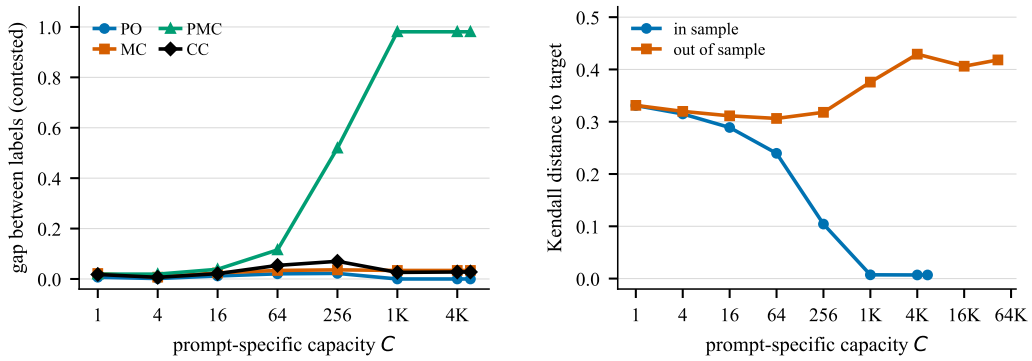
6.2 Reward models on LLM representations

Setup. We use UltraFeedback (Cui et al., 2024), in which each of $55{,}527$ retained instructions has four completions rated on four annotation aspects; the aspects act as four voters, an aspect that rates two completions equally splits its vote, and every pair of a profile faces the same electorate. Borda and Copeland disagree on 18.3% of prompts, which we call *contested*. Rewards are fitted on the objective in Eq. (2.3) summed over prompts, with $\lambda = 0.1$ penalizing candidate scores rather than coefficients, on an $80/10/10$ split by prompt; a head shared across prompts is thus an empirical test of the per-prompt theory. We compare linear heads on frozen last-token representations of `Qwen2.5-1.5B-Instruct` and `Qwen2.5-7B-Instruct` (Qwen Team, 2024), projected to d dimensions by nested random projections, and a `Qwen2.5-1.5B-Instruct` reward model fine-tuned end to end with LoRA (Hu et al., 2022).

The aggregation rule is decisive in the unrestricted limit. On the contested prompts, the unrestricted Borda estimator returns the pairwise majority ranking in *none* of the 52 cases where one exists, while the unrestricted Copeland estimator returns it in all 52 (Table 2). Pooled over all prompts the gap nearly vanishes, since the two rules coincide on the remaining 82%; reporting only pooled numbers would make the choice of loss look immaterial.

Learned reward models are far from either rule. A stronger model helps every axiom: moving from the frozen 1.5B encoder to the frozen 7B encoder or to end-to-end LoRA raises pairwise accuracy from 0.760 to 0.824 and CC from 0.584 to about 0.66, and the two stronger models land at the same Kendall distance of about 0.216. Yet the labels remain indistinguishable. Across all models and all projection dimensions, standard and majority labels differ by at most 0.011 on any axiom over all prompts; the largest contested gap, 0.058, is three of 52 prompts, and its sign differs between the 7B and LoRA models. As in the low-expressivity regime of Section 6.1, representation error, not the aggregation objective, is what binds.

When does the aggregation rule become visible? To vary how close the fitted reward can get to its unrestricted target, we partition prompts into C clusters and fit one head per cluster, interpolating



(a) Label gap on contested prompts, in sample.

(b) Distance to the target order.

Figure 1: Prompt-specific capacity on UltraFeedback, with one reward head per cluster of prompts. **(a)** The gap between standard and majority labels opens only as the fitted reward approaches its unrestricted target; on pairwise majority consistency it goes from 0.019 at $C = 1$ to 0.981 at $C = 1024$. **(b)** The same sweep in and out of sample. Approximation improves monotonically with capacity, generalization turns around near $C = 64$, and the capacity that makes the aggregation rule visible is exactly the capacity a model that must generalize cannot use.

between a shared head ($C = 1$) and one head per prompt, which recovers the per-prompt unrestricted solution. In sample, the contested PMC gap between the labels rises from 0.019 at $C = 1$ to 0.115 at $C = 64$, 0.519 at $C = 256$, and 0.981 at $C = 1024$, in lockstep with the Kendall distance falling from 0.331 to 0.007 (Figure 1a). The loss therefore decides the outcome only once the model can realize it. Out of sample, the distance bottoms out at 0.306 near $C = 64$ and then deteriorates, and the gap never exceeds 0.038 (Figure 1b): reward models that must generalize across prompts sit in the regime where the aggregation rule is invisible.

6.3 Borda and Copeland on Chatbot Arena

On Chatbot Arena (Chiang et al., 2024), candidates are language models and voters are anonymous human users. Among the twenty most-battled models ($42{,}539$ votes), subsampling every pair to the same 63 votes balances the comparisons, and the regularized BT order then coincides with the Borda order in all twenty subsamples, as Theorem 3.1 predicts; on the full unbalanced data used by the live leaderboard, the two orders differ. Borda and Copeland disagree on 3.7% of pairs, and five of the top eight positions change, but every disagreement sits on a near-tie: the largest head-to-head margin that the BT order reverses is $|P_{ij} - \frac{1}{2}| = 0.059$, and none of the 88 pairs with a statistically clear majority is reversed (Appendix E.5). On real preference data, the axiom violations of Borda concentrate where the electorate is close to indifferent, which is where they cost least.

7 Conclusion

We asked when the axiomatic guarantees of RLHF can be recovered and found that two choices decide: the labels a reward model is fitted to, which select Borda or Copeland, and its expressivity relative to the candidate set, which decides whether that rule is reached. Majority labels recover all four axioms for unrestricted rewards; linear rewards inherit them for score-complete features, and polynomial rewards beyond an interpolation degree once the candidate set is fixed, so capacity is decisive rather than futile.

Our guarantees concern a fixed candidate set without pairwise majority ties, and majority labels need several judgments per comparison. Since reward models trained on LLM representations remain far from their unrestricted targets, extending the analysis to reward models that generalize across prompts is the main open direction.

AI USE STATEMENT

We used generative AI tools as follows. **Writing.** AI tools were used to polish the text for readability, drafted parts of the paper from detailed outlines written by the authors, searched for related literature; the authors checked the content. **Theory.** The research question, the formulation, and the main results were developed by the authors. AI tools were used to check the correctness of the proofs, to polish their presentation, and to repair some arguments; the authors verified the results. **Experiments.** AI tools were used to write and run the code; the authors checked the codes the results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theoretical results state their assumptions in the main text, and complete proofs are given in the appendix. For the experiments, All datasets and models are publicly available.

REFERENCES

Parand A. Alamdari, Soroush Ebadian, and Ariel D. Procaccia. Policy aggregation. In *Advances in Neural Information Processing Systems*, 2024.

Zhiyu An, Duaa Nakshbandi, and Wan Du. Differential voting: Loss functions for axiomatically diverse aggregation of heterogeneous preferences. *arXiv preprint arXiv:2601.18824*, 2026.

Anthropic. Claude Fable 5.1 and Claude Mythos 5.1 System Card. Technical report, Anthropic, September 2026. URL https://www.anthropic.com/claude-fable-5-1-mythos-5-1-system-card.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 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, 2024.

Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, 2022.

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.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb.

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, 2024.

Jonathan D Chang, Wenhao Shan, Owen Oertell, Kianté Brantley, Dipendra Misra, Jason D Lee, and Wen Sun. Dataset reset policy optimization for RLHF. *arXiv preprint arXiv:2404.08495*, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Jessica Dai and Eve Fleisig. Mapping social choice theory to RLHF. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

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. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a13aYUU9eU.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024.

Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. In *Proceedings of the 25th ACM Conference on Economics and Computation*, 2024.

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*, 2024a. URL https://openreview.net/forum?id=cmBjkpRuvw.

Luise Ge, Brendan Juba, and Yevgeniy Vorobeychik. Learning Linear Utility Functions From Pairwise Comparison Queries. *arXiv preprint arXiv:2405.02612*, 2024b.

William V Gehrlein. *Condorcet's paradox*. Springer, 2006.

Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems*, 2025.

Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Daniel Halpern, Evi Micha, Ariel D. Procaccia, and Itai Shapira. Pairwise calibrated rewards for pluralistic alignment. In *Advances in Neural Information Processing Systems*, 2025.

Daniel Halpern, Evi Micha, Ariel D. Procaccia, Benjamin Schiffer, Itai Shapira, and Shirley Zhang. AI alignment from social choice perspectives. *ACM SIGecom Exchanges*, 2026. To appear; arXiv:2606.21550.

Alexandros Hollender and Sonja Kraiczy. Enforcing axioms for AI alignment under loss-based rules. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MpYSoTK65s.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022.

Ruochen Jin, Bojian Hou, Jiancong Xiao, Weijie J Su, and Li Shen. Fine-tuning attention modules only: Enhancing weight disentanglement in task arithmetic. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dj0TktJcVI.

Kihyun Kim, Jiawei Zhang, Asuman Ozdaglar, and Pablo A. Parrilo. Beyond RLHF and NLHF: Population-proportional alignment under an axiomatic framework. In *The Fourteenth International Conference on Learning Representations*, 2026.

Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, and Marta Kwiatkowska. Strategyproof reinforcement learning from human feedback. In *Advances in Neural Information Processing Systems*, 2025.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. OpenAssistant conversations – democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47669–47681, 2023.

Ziniu Li, Tian Xu, and Yang Yu. Policy optimization in RLHF: The impact of out-of-preference data. *arXiv preprint arXiv:2312.10584*, 2023.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 29128–29163, 2024.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NQEe7B7bSw.

Kaizhao Liu, Qi Long, Zhekun Shi, Weijie J Su, and Jiancong Xiao. Statistical impossibility and possibility of aligning LLMs with human preferences: From Condorcet paradox to Nash equilibrium. *arXiv preprint arXiv:2503.10990*, 2025.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

Roberto-Rafael Maura-Rivero, Marc Lanctot, Francesco Visin, and Kate Larson. Jackpot! alignment as a maximal lottery. *arXiv preprint arXiv:2501.19266*, 2025.

Abhilash Mishra. AI alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*, 2023.

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 *Forty-first International Conference on Machine Learning*, 2024.

Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 1587–1594, 2018.

Ritesh Noothigattu, Dominik Peters, and Ariel D Procaccia. Axioms for learning from pairwise comparisons. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17745–17754, 2020.

OpenAI. GPT-6 Astra System Card. Technical report, OpenAI, September 2026. URL `https://deploymentsafety.openai.com/gpt-6-astra`.

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, 2022.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled RLHF from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Ariel D. Procaccia, Benjamin Schiffer, and Shirley Zhang. Clone-robust AI alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Tianyi Qiu. Representative social choice: From learning theory to AI alignment. In *NeurIPS 2024 Workshop on Pluralistic Alignment*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, 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.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to Q^*: Your language model is secretly a Q-function. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=kEVcNxtqXk`.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhekun Shi, Kaizhao Liu, Qi Long, Weijie J Su, and Jiancong Xiao. Fundamental limits of game-theoretic LLM alignment: Smith consistency and preference matching. *arXiv preprint arXiv:2505.20627*, 2025.

Ali Shirali, Arash Nasr-Esfahany, Abdullah Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel D. Procaccia. Direct alignment with heterogeneous preferences. In *Advances in Neural Information Processing Systems*, 2025.

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

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

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, 2024.

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*, pp. 47441–47474, 2024.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mingzhi Wang, Chengdong Ma, Qizhi Chen, Linjian Meng, Yang Han, Jiancong Xiao, Zhaowei Zhang, Jing Huo, Weijie J Su, and Yaodong Yang. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PDnEDS244P.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, et al. 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, 2024.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J. Su, and Li Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025a.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552):2154–2164, 2025b.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO superior to PPO for LLM alignment? A comprehensive study. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54983–54998, 2024.

Ting Yan. Ranking in the generalized Bradley–Terry models when the strong connection condition fails. *Communications in Statistics—Theory and Methods*, 45(2):344–358, 2016.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 2027–2032. IEEE, 2024.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 81773–81807, 2024.

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

A PROOFS FOR UNRESTRICTED REWARD MODELS

This appendix supports Section 3: Appendix A.1 proves Theorem 3.1, Appendix A.2 proves that the Copeland score satisfies the axioms invoked in Section 3, and Appendix A.3 shows that the listwise Plackett–Luce loss, a natural alternative when labelers rank several responses, fails the same axioms.

A.1 PROOF OF THEOREM 3.1

We work with the objective $F_{q,\lambda}$ of Eq. (2.3), balanced weights $q_{ij} + q_{ji} = 1$, and scores $b_i = \sum_{j \neq i} q_{ij}$.

Lemma A.1. *Let $q_{ij} \in [0,1]$ with $q_{ij} + q_{ji} = 1$ for every pair, and fix $\lambda > 0$. Then $F_{q,\lambda}$ has a unique finite minimizer $r^\star$ over $\mathbb{R}^n$, it lies in $\mathcal{H}$, and for all $i, j \in [n]$,*

$$r_i^\star > r_j^\star \iff b_i > b_j.$$

If $\lambda = 0$ and a finite MLE exists under an identifying normalization, the same ordering equivalence holds.

Proof. The likelihood part of $F_{q,\lambda}$ is convex. The quadratic term makes the full objective λ-strongly convex and coercive, so a unique finite minimizer $r^\star$ exists. Since $\partial_{r_k}\left[q_{kj}\log(1 + e^{r_j - r_k}) + q_{jk}\log(1 + e^{r_k - r_j})\right] = \sigma(r_k - r_j) - q_{kj}$, its first-order conditions are

$$\lambda r_k^\star + \sum_{j \neq k} \sigma(r_k^\star - r_j^\star) = b_k, \qquad k \in [n].$$

Summing over k and using $\sum_k \widehat{b}_k(r) = \sum_k b_k = n(n-1)/2$ gives $\sum_k r_k^\star = 0$, so $r^\star \in \mathcal{H}$. Subtracting the equation for j from that for i gives

$$b_i - b_j = \lambda(r_i^\star - r_j^\star) + \left(e^{r_i^\star} - e^{r_j^\star}\right) A_{ij},$$

where

$$A_{ij} = \sum_{k \neq i,j} \frac{e^{r_k^\star}}{(e^{r_i^\star} + e^{r_k^\star})(e^{r_j^\star} + e^{r_k^\star})} + \frac{1}{e^{r_i^\star} + e^{r_j^\star}} > 0.$$

Both terms on the right have the same sign as $r_i^\star - r_j^\star$. Hence $b_i - b_j$ and $r_i^\star - r_j^\star$ have the same sign, proving the equivalence. When $\lambda = 0$ and a finite normalized MLE exists, the same calculation applies with the first term omitted. □

Lemma A.1 is Theorem 3.1. For the count form in Eq. (2.2) with N comparisons per pair, apply it with $q_{ij} = P_{ij}$ and penalty λ/N (Section 2). The ranking is therefore determined by the design of q_{ij}.

Copeland Rule. For the Copeland loss without pairwise majority ties,

$$q_{ij} = \mathbb{1}[P_{ij} > 1/2], \qquad b_i = \sum_{k \neq i} \mathbb{1}[P_{ik} > 1/2] = C_i,$$

which is exactly the Copeland score.

Borda Count Rule. For standard BT, $q_{ij} = P_{ij}$ is the fraction of voters who prefer y_i over y_j, so

$$b_i = \sum_{k \neq i} P_{ik} = \mathrm{BC}_i,$$

the Borda count. Therefore, Lemma A.1 reproduces the finding that ℓ_2-regularized BT reward modeling implements the Borda count rule; the same conclusion holds for the unregularized MLE whenever it exists (Siththaranjan et al., 2024; Maura-Rivero et al., 2025).

A.2 THE COPELAND SCORE SATISFIES THE SOCIAL CHOICE AXIOMS

Throughout, C_k denotes the Copeland score of y_k, the number of candidates that y_k defeats by a strict pairwise majority, and the Copeland rule ranks candidates by C_k. Pareto optimality and majority consistency concern complete profiles, in which every voter submits a strict ranking of $\mathcal{Y}$. Pairwise majority consistency and Condorcet consistency depend only on the pairwise majority relation, so Theorems A.3 and A.4 hold for any profile; in particular they cover the single-labeler profiles of Section 3, whose comparisons may form cycles.

Theorem A.1 (Pareto optimality)**.** *Consider a complete profile without pairwise majority ties. If every voter strictly prefers y to y', then $C_y > C_{y'}$, so the Copeland rule ranks y strictly above y'.*

Proof. Fix $z \in \mathcal{Y} \setminus \{y, y'\}$. On every ballot, $y' \succ z$ implies $y \succ z$, because y is above y'. Hence the voters preferring y' to z are a subset of those preferring y to z, and if y' defeats z then so does y. Summing over z,

$$\#\{z \neq y, y' : \ y' \text{ defeats } z\} \leq \#\{z \neq y, y' : \ y \text{ defeats } z\}.$$

In addition, y defeats y' unanimously while y' does not defeat y. Therefore $C_y \geq C_{y'} + 1$. □

Theorem A.2 (Majority consistency)**.** *Consider a complete profile. If a strict majority of voters rank y^{mw} first, then y^{mw} is the unique Copeland winner.*

Proof. The voters who rank y^{mw} first, more than half of all voters, prefer it to every other candidate, so y^{mw} defeats every rival and $C_{y^{\mathrm{mw}}} = n - 1$. Every other candidate loses to y^{mw}, so its score is at most $n - 2$. □

Theorem A.3 (Pairwise majority consistency)**.** *Let $\succ_M$ denote the strict pairwise majority relation of any profile. If $\succ_M$ is complete (there are no pairwise majority ties) and transitive, then the Copeland rule returns the ranking $\succ_M$.*

Proof. Relabel the candidates so that $y_{(1)} \succ_M y_{(2)} \succ_M \cdots \succ_M y_{(n)}$. Since $\succ_M$ is a complete linear order, $y_{(k)}$ defeats exactly the $n - k$ candidates below it, so $C_{y_{(k)}} = n - k$. These scores are strictly decreasing in k, so sorting by Copeland score returns $\succ_M$. □

Theorem A.4 (Condorcet consistency)**.** *For any profile, if y^{cw} is a Condorcet winner, then it is the unique Copeland winner.*

Proof. By definition y^{cw} defeats every rival, so $C_{y^{\mathrm{cw}}} = n - 1$. Every other candidate loses to y^{cw}, so its score is at most $n - 2$; this bound also holds if a pairwise tie is counted as half a point. □

A.3 LISTWISE (PLACKETT–LUCE) LOSSES

Section 3 notes that InstructGPT-style labelers rank several responses. Instead of splitting each ranking into pairwise comparisons, one can fit the listwise Plackett–Luce (PL) likelihood. Theorem 3.1 concerns the pairwise loss and does not extend to this objective, and the next result shows that PL minimization fails the same three axioms as Borda. We use the following first-order characterization.

Lemma A.2 (First-order characterization of the listwise PL objective)**.** *Fix $n \geq 2$. Let $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$. Given m observed full rankings $\{\tau_t\}_{t=1}^m$ (each τ_t is a permutation of $[n]$), consider the empirical PL loss*

$$\mathcal{L}_{\mathrm{PL}}^{(m)}(r) := -\frac{1}{m} \sum_{t=1}^{m} \sum_{k=1}^{n} \log \frac{\exp\big(r_{\tau_t(k)}\big)}{\sum_{j=k}^{n} \exp\big(r_{\tau_t(j)}\big)}. \tag{A.1}$$

For each (t, k) define the remaining set

$$R_{t,k} := \{\tau_t(k), \tau_t(k+1), \ldots, \tau_t(n)\},$$

and for $i \in R_{t,k}$ define the stage-k softmax probability

$$p_{t,k}(i; r) := \frac{e^{r_i}}{\sum_{j \in R_{t,k}} e^{r_j}}.$$

Then, for every $i \in [n]$,

$$\frac{\partial \mathcal{L}_{\mathrm{PL}}^{(m)}(r)}{\partial r_i} = -\frac{1}{m}\sum_{t=1}^{m}\sum_{k=1}^{n}\Big(\mathbb{1}[\tau_t(k)=i] - \mathbb{1}[i \in R_{t,k}]\, p_{t,k}(i;r)\Big). \tag{A.2}$$

Consequently, any minimizer $\hat{r}$ of Eq. (A.1) *(under an identifiability constraint, e.g. $\sum_{i=1}^{n} r_i = 0$ or $r_1 = 0$) satisfies the first-order condition*

$$\sum_{t=1}^{m}\sum_{k=1}^{n}\mathbb{1}[\tau_t(k)=i] = \sum_{t=1}^{m}\sum_{k=1}^{n}\mathbb{1}[i \in R_{t,k}]\, p_{t,k}(i;\hat{r}), \qquad \forall i \in [n]. \tag{A.3}$$

Proof. Rewrite the summand in Eq. (A.1) as

$$-\log \frac{e^{r_{\tau_t(k)}}}{\sum_{j\in R_{t,k}} e^{r_j}} = -r_{\tau_t(k)} + \log \sum_{j\in R_{t,k}} e^{r_j}.$$

Differentiating w.r.t. r_i gives

$$\frac{\partial}{\partial r_i}\big(-r_{\tau_t(k)}\big) = -\mathbb{1}[\tau_t(k)=i],$$
$$\frac{\partial}{\partial r_i}\log\sum_{j\in R_{t,k}} e^{r_j} = \mathbb{1}[i\in R_{t,k}]\,\frac{e^{r_i}}{\sum_{j\in R_{t,k}} e^{r_j}} = \mathbb{1}[i\in R_{t,k}]\, p_{t,k}(i;r).$$

Summing over $k = 1,\dots,n$ and $t = 1,\dots,m$ and multiplying by $1/m$ yields Eq. (A.2). Setting Eq. (A.2) to zero gives Eq. (A.3). □

Theorem A.5 (PL minimization violates MC, PMC, and CC)**.** *There exist three full rankings of $n = 3$ candidates such that the pairwise majority relation is transitive, one candidate is both the majority winner and the Condorcet winner, and the unique minimizer $\hat{r}$ of Eq.* (A.1) *with $\sum_i \hat{r}_i = 0$ ranks a different candidate first. The PL minimizer therefore violates majority, pairwise majority, and Condorcet consistency on this profile.*

Proof. Let one voter submit $1 \succ 2 \succ 3$ and two voters submit $3 \succ 1 \succ 2$. Candidate 3 is ranked first by two of the three voters, so it is the majority winner. Pairwise, $\#(3 \succ 1) = 2$, $\#(3 \succ 2) = 2$, and $\#(1 \succ 2) = 3$, so 3 is the Condorcet winner and the pairwise majority relation is the transitive order $3 \succ 1 \succ 2$.

Draw an edge from the item chosen at each stage to every item still remaining at that stage. The edges $1 \to 2$, $2 \to 3$, and $3 \to 1$ already make this graph strongly connected. Along any nonzero direction in $\{r : \sum_i r_i = 0\}$, some edge $i \to j$ has $r_j - r_i$ increasing, and the corresponding stage term grows linearly, so the loss is coercive on this subspace. Its Hessian is a sum of softmax covariance matrices whose null space intersects the subspace only at zero, again by connectivity, so the loss is strictly convex there and the minimizer $\hat{r}$ exists and is unique. Solving Eq. (A.3) numerically by Newton's method gives

$$\hat{r} = (0.498,\ -0.859,\ 0.361),$$

at which the first-order conditions of Lemma A.2 hold to machine precision. Hence $\hat{r}_1 > \hat{r}_3 > \hat{r}_2$: candidate 1 is ranked first, so the majority winner and the Condorcet winner 3 is not, and the returned order $1 \succ 3 \succ 2$ differs from the pairwise majority order $3 \succ 1 \succ 2$. □

B Proofs and Further Results for Linear Reward Models

This appendix supports Section 4. Throughout, $n \geq 2$ candidates and a complete profile are fixed, so every pair is compared equally often and the objective takes the normalized form of Eq. (2.3) (Section 2). We write $L_q(r) = F_{q,\lambda}(r) - \frac{\lambda}{2}\|r\|_2^2$ for the unpenalized likelihood term. The penalty acts on the centered scores $H\Phi\theta$, which is equivalent to fitting $\Phi\theta + a\mathbf{1}$ with a free intercept and penalizing its squared norm, since the likelihood ignores a and the optimal intercept centers the scores. It is not coefficient weight decay, which induces a different score-space penalty (Appendix C.6). Appendices B.1 and B.2 prove Theorem 4.1 and Proposition 4.1; the remaining subsections prove the statements in the paragraph "Feasibility and the scope of the guarantees."

B.1 Proof of Theorem 4.1

Direct differentiation gives $\nabla F_{q,\lambda}(r) = \widehat{b}(r) - b + \lambda r$, and the Hessian is

$$\nabla^2 F_{q,\lambda}(r) = \lambda I_n + \sum_{i<j} \sigma(r_i - r_j)\sigma(r_j - r_i)(e_i - e_j)(e_i - e_j)^\top.$$

Hence $F_{q,\lambda}$ is coercive and λ-strongly convex, and its restriction to the closed subspace $\mathcal{S}_\Phi$ has a unique finite minimizer, also with deterministic labels. The ordering of $r^\star$ by b is Theorem 3.1.

At r^{lin}, first-order optimality gives $(H\Phi)^\top \nabla F_{q,\lambda}(r^{\mathrm{lin}}) = 0$. Both $\widehat{b}(r)$ and b sum to $n(n-1)/2$, so the gradient is centered whenever r is, and H may be dropped, giving the moment equation in Eq. (4.2). Since $F_{q,\lambda} = A_\lambda - b^\top r$ and $\nabla A_\lambda(r^\star) = b$ by the first-order condition of $r^\star$,

$$D_{A_\lambda}(r, r^\star) = F_{q,\lambda}(r) - F_{q,\lambda}(r^\star) \qquad (r \in \mathcal{H}), \tag{B.1}$$

and minimizing over $\mathcal{S}_\Phi$ proves the projection characterization. Finally, $\mathrm{rank}[\mathbf{1}\ \Phi] = 1 + \mathrm{rank}(H\Phi)$, so $\mathrm{rank}[\mathbf{1}\ \Phi] = n$ gives $\mathcal{S}_\Phi = \mathcal{H}$ and $r^{\mathrm{lin}} = r^\star$. With standard labels the fitted order is then the Borda weak order; with majority labels and no pairwise majority ties it is the Copeland order, which satisfies PO, MC, PMC, and CC on complete profiles (Appendix A.2). If majority ties occur, the half-point weights of Section 2 still sum to one within each pair, and the same argument applies to Copeland scores that award half a point for a tie.

B.2 Proof of Proposition 4.1

Because $\sigma(t)\sigma(-t) \le 1/4$ and $\sum_{i<j}(e_i - e_j)(e_i - e_j)^\top = nH$, the Hessian on $\mathcal{H}$ obeys the global bounds

$$\lambda I_{\mathcal{H}} \preceq \nabla^2 A_\lambda(r)|_{\mathcal{H}} \preceq (\lambda + n/4) I_{\mathcal{H}}.$$

Integrating the Hessian along line segments gives

$$\frac{\lambda}{2}\|u - v\|_2^2 \le D_{A_\lambda}(u, v) \le \frac{\lambda + n/4}{2}\|u - v\|_2^2 \qquad (u, v \in \mathcal{H}). \tag{B.2}$$

Let r^Π be the Euclidean projection of $r^\star$ onto $\mathcal{S}_\Phi$. By Eq. (B.1),

$$\frac{\lambda}{2}\|r^{\mathrm{lin}} - r^\star\|_2^2 \le D_{A_\lambda}(r^{\mathrm{lin}}, r^\star) \le D_{A_\lambda}(r^\Pi, r^\star) \le \frac{\lambda + n/4}{2}\delta_\Phi^2.$$

This proves Eq. (4.3). Furthermore,

$$|(r_i^{\mathrm{lin}} - r_j^{\mathrm{lin}}) - (r_i^\star - r_j^\star)| \le \|e_i - e_j\|_2 \|r^{\mathrm{lin}} - r^\star\|_2 \le \sqrt{2}\,\varepsilon_\Phi. \tag{B.3}$$

Thus the stated gap condition preserves the required comparison. Unlike a condition on the minimum gap over all candidates, it does not require unrelated candidates to have different voting scores.

A useful version uses voting-score differences directly. For any $i \ne j$,

$$\widehat{b}_i(r) - \widehat{b}_j(r) = \sigma(r_i - r_j) - \sigma(r_j - r_i) + \sum_{\ell \notin \{i,j\}} [\sigma(r_i - r_\ell) - \sigma(r_j - r_\ell)], \tag{B.4}$$

which has the sign of $r_i - r_j$. Subtracting the ith and jth first-order conditions $\widehat{b}(r^\star) + \lambda r^\star = b$ and using $\sigma' \le 1/4$ in Eq. (B.4) gives, whenever $b_i > b_j$,

$$\frac{b_i - b_j}{\lambda + n/4} \le r_i^\star - r_j^\star < \frac{b_i - b_j}{\lambda}. \tag{B.5}$$

In particular, under majority labels a Condorcet winner has a voting advantage of at least one. It remains the unique linear-reward winner whenever $\sqrt{2}\,\varepsilon_\Phi < 1/(\lambda + n/4)$.

B.3 Unregularized estimators and continuity at zero penalty

Theorem 4.1 remains valid for $\lambda = 0$ whenever the unregularized minimizer is finite. For the unrestricted estimator, strong connectivity of the directed graph with an edge $i \to j$ whenever $q_{ij} > 0$ suffices. For a restricted score space it is sufficient but not necessary; the exact condition is as follows.

Lemma B.1 (Existence on the feasible score space)**.** *The unregularized loss L_q has a finite minimizer on $\mathcal{S}_\Phi$ if and only if there is no nonzero $h \in \mathcal{S}_\Phi$ such that $h_i \geq h_j$ for every directed comparison with $q_{ij} > 0$. When it exists, its centered score vector is unique.*

Proof. If such h exists, $L_q(r + th)$ is nonincreasing in t and strictly decreases for at least one term. Indeed, h is centered and nonzero, so some pair has unequal coordinates; complete comparisons force at least one observed edge on that pair, and its margin must be positive. Every finite point can therefore be improved along h.

Conversely, consider $r^\ell \in \mathcal{S}_\Phi$ with $\|r^\ell\|_2 \to \infty$. Along a subsequence, $r^\ell/\|r^\ell\|_2 \to h \in \mathcal{S}_\Phi$, $\|h\|_2 = 1$. The assumed absence of a separating direction supplies an observed edge $i \to j$ with $h_i < h_j$. Then $q_{ij} \log(1 + e^{r_j^\ell - r_i^\ell}) \to \infty$. Thus a bounded sublevel set cannot contain an unbounded sequence, and L_q is coercive. Its Hessian is positive definite on $\mathcal{H}$ at every finite vector, so the restricted minimizer is unique. □

Lemma B.2 (Persistence of a strict counterexample)**.** *Let $\mathcal{S} \subseteq \mathcal{H}$ be a closed linear subspace, such as $\mathcal{S}_\Phi$ or the polynomial space $\mathcal{S}_k$ of Section 5, and suppose L_q has a finite minimizer r^0 on $\mathcal{S}$. If r^λ minimizes $L_q(r) + \lambda\|r\|_2^2/2$ over $\mathcal{S}$, then $r^\lambda \to r^0$ as $\lambda \downarrow 0$, and any strict reversed comparison at r^0 persists for all sufficiently small $\lambda > 0$.*

Proof. Optimality and $L_q(r^\lambda) \geq L_q(r^0)$ imply

$$\|r^\lambda\|_2 \leq \|r^0\|_2, \qquad 0 \leq L_q(r^\lambda) - L_q(r^0) \leq \frac{\lambda}{2}\|r^0\|_2^2.$$

Every accumulation point therefore minimizes L_q on $\mathcal{S}$. For complete balanced comparisons the Hessian of L_q is positive definite on $\mathcal{H}$ at every finite vector, so this minimizer is unique and equals r^0, proving convergence. Strict inequalities persist by continuity. □

B.4 A FOUR-CANDIDATE COUNTEREXAMPLE

Proposition B.3 (Linear BT can reject a majority and Condorcet winner)**.** *There are four candidates and four linearly feasible voter rankings whose Borda and majority orders are both $D \succ B \succ A \succ C$, with D ranked first by three quarters of voters, but whose finite unregularized linear BT estimator ranks B above D. The same failure occurs for all sufficiently small $\lambda > 0$ in Eq.* (4.1)*.*

Proof. Take

$$\phi_A = (0, 0), \quad \phi_B = (1, 0), \quad \phi_C = (0, 1), \quad \phi_D = (2, 2).$$

One voter has parameter $(1, -3/2)$ and submits $B \succ A \succ D \succ C$. Three have parameter $(1, -1/4)$ and submit $D \succ B \succ A \succ C$. The Borda scores multiplied by four are $(5, 9, 0, 10)$ in the order (A, B, C, D). The majority and Condorcet claims follow directly.

Write $\theta = (u, v)$, let $q = r_D - r_B = u + 2v$, and put $\ell(t) = \log(1 + e^{-t})$. Then the raw scores are $(0, u, (q - u)/2, u + q)$. Four times the normalized unregularized loss is

$$\begin{aligned} L(u, q) = {} & 4\ell(u) + 4\ell((3u - q)/2) + 4\ell((3u + q)/2) \\ & + 3\ell(q) + \ell(-q) + 3\ell(u + q) + \ell(-u - q) + 4\ell((u - q)/2). \end{aligned} \tag{B.6}$$

On the boundary $q = 0$, the unique minimizer u_0 solves

$$F(u_0) := 8\sigma(u_0) + 12\sigma(3u_0/2) + 2\sigma(u_0/2) - 21 = 0. \tag{B.7}$$

The function F is strictly increasing and

$$F(2\log(1 + \sqrt{2})) = \frac{51\sqrt{2} - 83}{7} < 0.$$

Hence $u_0 > 2\log(1 + \sqrt{2})$. At this point,

$$\partial_q L(u_0, 0) = 2\left[\frac{1}{1 + e^{u_0/2}} - \frac{2}{1 + e^{u_0}}\right] > 0. \tag{B.8}$$

Convexity and $\partial_u L(u_0, 0) = 0$ show that $(u_0, 0)$ minimizes L over $q \geq 0$. Moving toward $q < 0$ decreases the loss, so the unconstrained minimizer must have $q < 0$.

Existence can be checked without assuming a strongly connected unrestricted win graph. The mixed-label terms for D versus B and D versus A are $3\ell(q) + \ell(-q)$ and $3\ell(u+q) + \ell(-u-q)$. Each is bounded below by the absolute value of its argument. Thus $L(u, q) \geq |q| + |u+q|$, which is coercive in (u, q). The pairwise feature differences span $\mathbb{R}^2$, so the Hessian is positive definite and the optimum is unique. Centering changes none of these comparisons, and Lemma B.2 extends the strict reversal to sufficiently small $\lambda > 0$. □

B.5 Feasible majority rankings: an explicit positive bound and its limit

For the unregularized majority-label loss, Ge et al. (2024a, Theorem 3.6) use a rule defined through the infimum attained within a ranking region. This is distinct from minimizing a loss with a fixed positive penalty. We can give a finite-estimator guarantee with an explicit penalty threshold.

Proposition B.4 (Recovery of a feasible majority ranking)**.** *Suppose the strict majority relation is a total order ρ and some $h \in \mathcal{S}_\Phi$ realizes it with $\gamma = \min_{i \succ_\rho j}(h_i - h_j) > 0$. Let $E = \binom{n}{2}$ be the number of pairs, $R = \|h\|_2$, and*

$$\lambda_\rho = \frac{\gamma^2 \log 2}{R^2 \{\log(2E/\log 2)\}^2}. \tag{B.9}$$

Under majority labels, every $0 < \lambda < \lambda_\rho$ makes r^{lin} induce ρ.

Proof. For $t > 0$, all margins of th are at least $t\gamma$, and $\log(1 + e^{-z}) \leq e^{-z}$ gives

$$F_{q,\lambda}(th) \leq Ee^{-t\gamma} + \lambda t^2 R^2/2.$$

Choose $t = \gamma^{-1} \log(2E/\log 2)$. Then the first term is $\log 2/2$, and the assumed bound on λ makes the second strictly smaller than $\log 2/2$. The optimum therefore has total objective less than $\log 2$. Any score vector that reverses or ties a majority comparison has a single logistic term at least $\log 2$, a contradiction. □

The threshold cannot simply be omitted: the next construction gives an explicit PO failure under strong score regularization.

Proposition B.5 (Strong regularization can reverse a unanimous comparison)**.** *Take four features $(0,0), (1,-3), (2,-2), (10,-3)$, named A, B, C, D. If all voters submit the linearly feasible order $D \succ C \succ B \succ A$, then for every $\lambda \geq 100$ the estimator in Eq. (4.1) ranks B above C. Thus both standard and majority-weighted linear BT violate PO and PMC on this instance.*

Proof. The common ranking is induced by $\theta = (1, 0)$, so standard and majority weights coincide. Let P_Φ be Euclidean projection onto $\mathcal{S}_\Phi$ and set

$$z = P_\Phi(b - \tfrac{3}{2}\mathbf{1}) = \frac{1}{146}(-183, -17, -35, 235)^\top, \qquad b = (0, 1, 2, 3)^\top.$$

This projection follows by solving the two-dimensional least-squares system for $H\Phi$. In particular, $z_B - z_C = 9/73$ and $\|z\|_2 < 3$.

For general n, the projected optimality equation reads $\lambda r^{\mathrm{lin}} = z - P_\Phi(\widehat{b}(r^{\mathrm{lin}}) - \widehat{b}(0))$. Convexity of L_q and $\nabla L_q(0) = \frac{n-1}{2}\mathbf{1} - b$ imply $\lambda\|r^{\mathrm{lin}}\|_2 \leq \|z\|_2$. The unregularized Hessian has operator norm at most $n/4$ on $\mathcal{H}$. Therefore

$$\|\lambda r^{\mathrm{lin}} - z\|_2 \leq \frac{n}{4}\|r^{\mathrm{lin}}\|_2 \leq \frac{n\|z\|_2}{4\lambda}. \tag{B.10}$$

For this $n = 4$ example and $\lambda \geq 100$,

$$\lambda(r_B^{\mathrm{lin}} - r_C^{\mathrm{lin}}) \geq \frac{9}{73} - \frac{\sqrt{2}\|z\|_2}{\lambda} > \frac{9}{73} - \frac{3\sqrt{2}}{100} > 0.$$

The unanimous comparison is strictly reversed. This does not contradict unrestricted order preservation: the projection onto a proper score subspace is essential here. □

B.6 GEOMETRIC FEASIBILITY AND THE LCPO ALTERNATIVE

For this subsection assume that features are distinct and every voter ranking is linearly feasible. The LCPO rule of Ge et al. (2024a) lexicographically maximizes the sequence of Copeland scores over feasible strict rankings that respect every unanimous comparison. This admissible set is nonempty because it contains each submitted ranking. It is a ranking-selection rule, not a majority-weighted logistic estimator.

Proposition B.6 (Feasibility-aware Condorcet consistency)**.** *A candidate y_c is uniquely top-ranked by some linear reward if and only if ϕ_c is an exposed vertex of the feature convex hull. If y_c is a Condorcet winner, LCPO ranks it first whenever a feasible Pareto-respecting ranking with y_c first exists. If no such ranking exists, no linear rule can simultaneously respect PO and uniquely rank y_c first on that profile.*

Proof. A linear functional uniquely maximizes at ϕ_c over the finite features exactly when it uniquely maximizes there over their convex hull. This is the definition of an exposed vertex. For the second claim, y_c has Copeland score $n-1$ and every other candidate has score at most $n-2$. An admissible ranking starting with y_c is therefore lexicographically preferred to every admissible ranking with another first candidate. The last claim is immediate for strict rankings. It also covers a reward with ties among lower candidates: any parameter satisfying the strict winner and unanimous inequalities can be perturbed slightly to avoid the finitely many remaining equality hyperplanes, preserving those strict inequalities and producing an admissible strict ranking. □

For a small explicit obstruction, take $\phi_A = (2, 0)$, $\phi_B = (-1, 2)$, $\phi_C = (-1, -2)$, and $\phi_D = (0, 0)$. The voter parameters $(1, 1/10), (-1, 1), (-1, -1)$ induce, respectively,

$$A \succ D \succ B \succ C, \qquad B \succ D \succ C \succ A, \qquad C \succ D \succ B \succ A.$$

Candidate D beats each rival by two votes to one, and the majority order is $D \succ B \succ C \succ A$. But $\phi_D = (\phi_A + \phi_B + \phi_C)/3$, so no linear reward can uniquely rank D first. Thus even the existence of a strict majority order does not imply its linear feasibility.

B.7 A COMPLETE ONE-DIMENSIONAL POSITIVE RESULT

Proposition B.7 (Rank-one linear preference domain)**.** *Suppose $\mathcal{S}_\Phi = \mathrm{span}\{v\}$ for a centered vector with distinct coordinates, and every voter ranking is linearly feasible. Let ρ be the order of v and α the fraction of voters submitting ρ rather than its reverse. For standard BT and every $\lambda > 0$, the unique estimate is sv, where $\mathrm{sgn}(s) = \mathrm{sgn}(\alpha - 1/2)$. Consequently the reward-stage rule satisfies PO, MC, PMC, and CC on this restricted domain, including unanimous profiles.*

Proof. Every nonzero feasible score vector is a multiple of v, so the only strict feasible rankings are ρ and its reverse. Let $\Delta_{ij} = v_i - v_j > 0$ for $i \succ_\rho j$. The objective is

$$\sum_{i \succ_\rho j} [\alpha \log(1 + e^{-s\Delta_{ij}}) + (1 - \alpha) \log(1 + e^{s\Delta_{ij}})] + \frac{\lambda s^2}{2} \|v\|_2^2. \tag{B.11}$$

Its derivative is

$$\sum_{i \succ_\rho j} \Delta_{ij} [\sigma(s\Delta_{ij}) - \alpha] + \lambda s \|v\|_2^2. \tag{B.12}$$

This is strictly increasing, tends to opposite infinities at opposite ends of the real line, and at zero has the sign of $1/2 - \alpha$. Its unique zero therefore has the asserted sign, also when $\alpha = 0$ or 1. A strict majority orientation supplies both the majority order and the majority/Condorcet winner. Unanimity is preserved at the endpoints. If both orientations occur, no pair is unanimous; at $\alpha = 1/2$, the estimate is zero and none of the strict-majority axioms is triggered. If a strict output is required, choose either feasible orientation as a tie-breaking convention. □

C Proofs and Further Results for Polynomial Reward Models

This appendix supports Section 5. Throughout, the prompt, the $n \geq 2$ distinct features, and a complete profile are fixed, and the objective is the normalized form of Eq. (2.3) (Section 2), with L_q as in Appendix B. Many arguments of Appendix B hold for any closed subspace of $\mathcal{H}$ and apply here with $\mathcal{S}_k$ in place of $\mathcal{S}_\Phi$; we only prove what is specific to polynomial rewards. Appendices C.1 and C.2 prove Theorem 5.1 and Proposition 5.1, and the remaining subsections prove the statements in the paragraph "Axiomatic implications."

C.1 Proof of Theorem 5.1

Existence, centering, and projection. Because the constant monomial is present, $\mathbf{1} \in \mathrm{col}(V_k)$, so subtracting the mean of a representable reward vector preserves representability and leaves every reward difference unchanged. Since $\|r\|_2^2 = \|Hr\|_2^2 + n\bar{r}^2$ for $r = Hr + \bar{r}\mathbf{1}$, minimizing the score-penalized objective over $\mathrm{col}(V_k)$ automatically yields mean-zero rewards and is equivalent to Eq. (5.1). The reward vector is unique, although a coefficient vector realizing it need not be. Existence and uniqueness of $r^{(k)}$, the moment equation, and the Bregman characterization in Eq. (5.2) then follow exactly as in Appendix B.1, with $\mathcal{S}_\Phi$ and $H\Phi$ replaced by $\mathcal{S}_k$ and HV_k; the ordering of $r^\star$ by b is Theorem 3.1. The projected object is the unrestricted fitted reward vector, not the Borda or Copeland score vector, and the projection is not Euclidean in general.

Interpolation and its universal degree bound. Since $\mathbf{1} \in \mathrm{col}(V_k)$,

$$\mathrm{rank}(V_k) = 1 + \mathrm{rank}(HV_k).$$

Thus $\mathrm{rank}(V_k) = n$ is equivalent to $\mathcal{S}_k = \mathcal{H}$, in which case the constrained and unrestricted minimizers coincide. The spaces are nested as k increases. To show that full rank is eventually attained for any distinct features, define, for each $i \in [n]$,

$$L_i(z) = \prod_{j \neq i} \frac{(\phi_i - \phi_j)^\top (z - \phi_j)}{\|\phi_i - \phi_j\|_2^2}. \tag{C.1}$$

The denominators are nonzero. These polynomials have degree at most $n-1$ and satisfy $L_i(\phi_j) = \mathbb{1}[i = j]$. Therefore $p(z) = \sum_i a_i L_i(z)$ realizes any prescribed scores $(a_1, \ldots, a_n)$, proving $k_* \leq n - 1$.

Generic degree and sharpness. The dimension of the total-degree-k polynomial space is $D_k = \binom{d+k}{k}$. Evaluation functionals over all points of $\mathbb{R}^d$ span its dual: otherwise a nonzero polynomial would vanish at every point. We may therefore select D_k points with linearly independent evaluation functionals. Selecting fewer rows, or appending additional distinct points, shows that some evaluation matrix has rank $\min(n, D_k)$. Hence at least one maximal minor is a nonzero polynomial in the coordinates of the n feature vectors. Its zero set is a proper algebraic set. Outside that set, $\mathrm{rank}(V_k) = \min(n, D_k)$. Taking the finite union of the exceptional sets for the degrees up to the first k with $D_k \geq n$ establishes the generic formula in Eq. (5.3).

For sharpness, let the features lie on a line, $\phi_i = a + t_i v$, where $v \neq 0$ and the t_i are distinct. A degree-k multivariate polynomial restricts to a univariate polynomial of degree at most k. Conversely, a linear functional on this line recovers t, so every such univariate polynomial is a restriction of a multivariate one. The evaluation matrix consequently has rank $\min(n, k+1)$, and $k_* = n - 1$. This completes the proof.

C.2 Proof of Proposition 5.1

The norm bound and the pairwise bound are Appendix B.2 with $\mathcal{S}_k$ in place of $\mathcal{S}_\Phi$; in particular,

$$\left| (r_i^{(k)} - r_j^{(k)}) - (r_i^\star - r_j^\star) \right| \leq \sqrt{2}\, \varepsilon_k, \tag{C.2}$$

which proves the comparison and winner guarantees, and the voting-gap bound of Eq. (B.5) gives the explicit sufficient condition $\sqrt{2}\,\varepsilon_k < 1/(\lambda + n/4)$ for a Condorcet winner under majority labels to

remain the unique winner. For the monotonicity claim, $\mathcal{S}_k \subseteq \mathcal{S}_{k+1}$, and by Eq. (B.1) $r^{(k)}$ minimizes $D_{A_\lambda}(\cdot, r^\star)$ over $\mathcal{S}_k$, so minimizing over the larger space cannot increase the divergence. Neither the Euclidean error nor the number of correctly ordered pairs is asserted to decrease monotonically with k.

C.3 Axioms after interpolation and a persistent counterexample

When $k \geq k_*$, Theorem 5.1 identifies the fitted weak order with the order of b. Under standard labels b is the Borda count: if every voter prefers y_i to y_j, each ranks y_i above at least one more candidate than y_j, and averaging gives $b_i > b_j$, which establishes PO. Under majority labels b is the Copeland score, which satisfies PO, MC, PMC, and CC on complete profiles without pairwise majority ties (Appendix A.2); ties are handled by the half-point convention as in Appendix B.1.

Proposition C.1 (Majority failures at every positive degree)**.** *There is a complete profile with three candidates and linearly feasible voter rankings such that, for every $k \geq 1$ and every $\lambda > 0$, standard polynomial BT violates MC, PMC, and CC.*

Proof. Take $\phi_A = (1, 0)$, $\phi_B = (0, 1)$, and $\phi_C = (0, 0)$. Three voters submit $A \succ B \succ C$, induced by the linear utility vector $(2, 1)$; two submit $B \succ C \succ A$, induced by $(-1, 1)$. The Borda scores are

$$b_A = \frac{6}{5}, \qquad b_B = \frac{7}{5}, \qquad b_C = \frac{2}{5}.$$

Candidate A is ranked first by three of five voters and defeats both rivals by $3/5$. The majority relation is $A \succ B \succ C$, whereas the Borda order is $B \succ A \succ C$. The three feature vectors are affinely independent, so $k_* = 1$. Theorem 5.1 therefore gives the latter order for every $k \geq 1$ and $\lambda > 0$. All optima in this example are finite by the positive score penalty. □

C.4 Persistence of strict violations under small regularization

Lemma B.2 applies to $\mathcal{S}_k$ and transfers unregularized impossibility results to small positive penalties.

Corollary C.2 (Bounded-degree Pareto failures with a positive penalty)**.** *For each fixed $k \geq 1$, there exists a complete profile with distinct two-dimensional features and linearly feasible voter rankings, and a number $\lambda_0 > 0$, such that standard polynomial BT violates PO and PMC for every $0 < \lambda < \lambda_0$.*

Proof. Apply Theorem 4.1 and its construction in Hollender & Kraiczy (2026) to the logistic loss. The construction has a finite optimum with a strictly reversed unanimous comparison and a linearly feasible majority order. Lemma B.2 preserves that reversal for sufficiently small $\lambda > 0$, violating both PO and PMC. □

Here the quantifiers are $\forall k\, \exists(\phi, \boldsymbol{\tau}, \lambda_0)$: the bad instance changes with k, whereas Theorem 5.1 fixes the candidate set first, so there is no contradiction.

An extension to any fixed $\lambda > 0$. For the normalization in Eq. (5.1), the restriction to small λ can be removed if the candidate set may grow.

Proposition C.3 (Fixed-penalty impossibility by candidate replication)**.** *For every $k \geq 1$ and every $\lambda > 0$, there exist finitely many distinct two-dimensional candidate features and three voters with linearly feasible complete rankings such that degree-k polynomial BT with this λ violates PO and PMC.*

Proof. Use a base instance from Corollary C.2, with n_0 candidates, and let λ_0 be its positive stability threshold. The polynomial evaluation matrix of the cited construction has full column rank: its proof of Theorem 4.1 shows that a degree-k polynomial vanishing on its fixed anchor candidates is identically zero. Its two distinct voter directions admit a common strictly positive direction h, with one direction represented by two of the three voters. Choose an integer M such that $\lambda/M < \lambda_0$.

First replace each base candidate i by M labeled copies at the same feature vector. Each voter ranks the groups according to the base ranking and uses the same fixed strict ordering within each

group. At these coincident features every polynomial gives all copies of i the same reward r_i. Cross-group comparisons contribute $M^2 L_q(r)$, and the within-group comparisons contribute the constant $K = n_0\binom{M}{2}\log 2$. The score penalty contributes $\lambda M\|r\|_2^2/2$. Thus the replicated objective equals

$$M^2\left[L_q(r) + \frac{\lambda}{2M}\|r\|_2^2\right] + K. \tag{C.3}$$

Its minimizing base scores reverse a unanimous comparison because $\lambda/M < \lambda_0$.

To make the features distinct and all submitted rankings realizable, replace copy $a \in [M]$ of candidate i by $\phi_{i,a}(\eta) = \phi_i + \eta a h$, where $\eta > 0$ is sufficiently small. Every voter then strictly orders the copies by a, while all cross-group rankings remain those of the base instance. Choose the within-group order above to be this order. The comparison labels are therefore independent of η near zero. For small enough η, all feature vectors are distinct.

Let $\widetilde{V}(\eta)$ be the full monomial evaluation matrix on these copies, including the constant column. At $\eta = 0$ it is the base evaluation matrix with repeated rows, so it has full column rank. By continuity, its smallest singular value is bounded below by a positive constant in a neighborhood of zero. The coefficient objective

$$L_{\text{copies}}(\widetilde{V}(\eta)\theta) + \frac{\lambda}{2}\|\widetilde{V}(\eta)\theta\|_2^2$$

is consequently uniformly coercive and strictly convex in θ in that neighborhood. Comparing with $\theta = 0$ bounds all minimizers in a common compact set. Continuity of the objective and uniqueness then imply that the minimizing coefficients converge to their value at $\eta = 0$. The strict reversed comparison between copies of the unanimously ordered base pair persists for small positive η. This violates PO. Two voters still share one strict complete ranking, so their ranking is the strict majority order; the same reversed comparison also violates PMC. □

The replication identity uses the sum over candidate pairs and the unnormalized squared score norm in Eq. (5.1); if either is averaged with a candidate-count-dependent factor, the effective coefficient must be rescaled accordingly.

C.5 Quadratic rewards remove the individual-winner obstruction

Proposition C.4 (Every distinct candidate is top-feasible at degree two)**.** *For any fixed finite set of distinct features and every candidate y_c, there is a degree-two polynomial reward that ranks y_c uniquely first. This does not require* $\operatorname{rank}(V_2) = n$.

Proof. The polynomial $p_c(z) = -\|z - \phi_c\|_2^2$ satisfies $p_c(\phi_c) = 0 > p_c(\phi_j)$ for all $j \neq c$. Centering its evaluation vector preserves these inequalities and remains representable because the constant monomial is included. □

Thus, unlike a linear reward whose maximizer must be an exposed feature vertex, a quadratic reward can select a candidate inside the original convex hull. This is only a representability result: the fitted loss minimizer need not choose that polynomial. It also does not guarantee that the same polynomial satisfies every unanimous comparison or realizes a prescribed full ranking.

C.6 Why coefficient regularization is a different model

The interpolation conclusions depend on using the same score penalty for every degree. To see the distinction, put $Z_k = HV_k$ and instead minimize $L_q(Z_k\theta) + \lambda\|\theta\|_2^2/2$. For a fixed representable centered score vector r, the minimum squared coefficient norm among solutions of $Z_k\theta = r$ is

$$\min_{\theta: Z_k\theta = r} \|\theta\|_2^2 = r^\top (Z_k Z_k^\top)^\dagger r,$$

where † denotes the Moore–Penrose inverse. Indeed, a singular value decomposition of Z_k shows that the minimum-norm solution is $Z_k^\dagger r$, and its squared norm equals the displayed quadratic form. Coefficient ridge is therefore equivalent to

$$\min_{r \in \mathcal{S}_k} L_q(r) + \frac{\lambda}{2} r^\top (Z_k Z_k^\top)^\dagger r. \tag{C.4}$$

Even when $\mathcal{S}_k = \mathcal{H}$, this is generally an anisotropic, feature-dependent penalty, rather than $\lambda\|r\|_2^2/2$. Full evaluation rank alone therefore does not imply the Borda/Copeland ordering theorem for coefficient ridge. A sufficient condition is that $(Z_k Z_k^\top)^\dagger$ restrict to a positive scalar multiple of the identity on $\mathcal{H}$, since that only rescales λ.

Similarly, the full-rank theorem is about scores on the observed set. Different polynomial coefficients may produce the same optimal scores and disagree away from that set. Interpolation by itself supplies no out-of-sample guarantee.

C.7 Uniform original-space directions do not identify quadratic rewards

The uniform-direction PO theorem in Section 5 of Hollender & Kraiczy (2026) concerns linear rewards. A polynomial comparison depends on the pair of endpoints, not only their difference: $\psi_k(a) - \psi_k(b)$ is generally different from $\psi_k(a-b)$. The following population example makes the resulting limitation explicit.

Proposition C.5 (Uniform differences can leave all quadratic terms invisible)**.** *There is a distribution of comparison pairs with uniform unit differences in $\mathbb{R}^2$ and two distinct linear voters such that the unregularized quadratic BT population risk has finite minimizers, some of which violate PO on an unseen pair.*

Proof. Let U be uniform on the unit circle, and compare $a = U/2$ with $b = -U/2$. Thus $a - b = U$. Take equally weighted voters $v_1 = (1, 0)$ and $v_2 = (0, 1)$, who prefer a exactly when $\langle v_\ell, U\rangle > 0$; ties have probability zero. For any quadratic polynomial

$$p(z) = a_0 + w^\top z + z^\top Q z,$$

its training comparison is $p(U/2) - p(-U/2) = w^\top U$, independently of a_0 and Q. The risk is therefore the linear risk

$$R(w) = \frac{1}{2}\sum_{\ell=1}^{2} \mathbb{E}_U \log\Big(1 + e^{-\operatorname{sgn}(\langle v_\ell, U\rangle) w^\top U}\Big).$$

It has a finite minimizer. For completeness, on the open arcs $D = \{u : u_1 u_2 < 0\}$ the two voters disagree. For $t = w^\top u$, their average loss is at least $|t|/2$, so

$$R(w) \geq \frac{1}{2}\mathbb{E}_U[\mathbb{1}[U \in D]|w^\top U|] \geq c\|w\|_2$$

for some $c > 0$. The last bound follows because the expectation is continuous and strictly positive on the compact unit circle of directions $w/\|w\|_2$: no nonzero linear form vanishes on the open arcs D. Hence R is coercive and continuous, and attains its minimum at some finite $w^\star$.

For every real t, $p_t(z) = (w^\star)^\top z - t\|z\|_2^2$ is a quadratic population minimizer with the same training risk. Both voters prefer $z = (1, 1)$ to $z' = 0$, but

$$p_t(1, 1) - p_t(0) = w_1^\star + w_2^\star - 2t < 0$$

whenever $t > (w_1^\star + w_2^\star)/2$. Such a minimizer violates PO on this unseen pair. □

This example shows that uniform original-space differences alone do not force every unregularized polynomial minimizer to generalize PO. It does not establish a failure for the finite-set score-regularized estimator in Eq. (5.1).

D Social Choice Theory and Step 3 of RLHF

The main text studies the reward-modeling step. This appendix composes it with policy optimization: Appendix D.1 sets up the policy step, Appendix D.2 shows that it cannot repair the Borda violations, Appendix D.3 shows when it inherits reward-stage guarantees, Appendices D.4 and D.5 specialize this to linear and polynomial rewards, and Appendix D.6 reports experiments.

D.1 POLICY OPTIMIZATION

Let $\pi_\omega(\cdot|x)$ be the probability distribution over responses given a prompt x, where ω denotes the weights of the LLM. RLHF aims to maximize the reward with a KL penalty between the RLHF model and the reference model. The objective of policy optimization is to solve:

$$\max_{\omega} \mathbb{E}_{x\sim D_{\mathcal{X}}} \left[\mathbb{E}_{y\sim\pi_\omega(\cdot|x)}[r(x,y)] - \beta D_{\mathrm{KL}}(\pi_\omega(y|x)\|\pi_{\mathrm{ref}}(y|x))\right], \tag{D.1}$$

where $\beta > 0$ is a parameter controlling the deviation from the base reference policy π_{ref}. Above, the Kullback–Leibler divergence between $\pi_\omega(y|x)$ and $\pi_{\mathrm{ref}}(y|x)$ is defined as $\mathbb{E}_{y\sim\pi_\omega(\cdot|x)}\big[\log\pi_\omega(y|x) - \log\pi_{\mathrm{ref}}(y|x)\big]$.

Definition D.1 (Step 3: Policy Optimization)**.** The policy optimization stage is defined as a function $s_3 : \mathbb{R}^n \to \mathbb{S}(\mathcal{Y})$ that maps the reward vector $\boldsymbol{r}$ to a probability distribution $\boldsymbol{p} \in \Delta(\mathcal{Y})$ over the response set $\mathcal{Y}$, which subsequently induces a ranking $\in \mathbb{S}(\mathcal{Y})$.

In this setting, the policy optimization objective can be expressed as:

$$s_3(\boldsymbol{r}) = \arg\max_{\boldsymbol{p}\in\Delta(\mathcal{Y})} \sum_{i=1}^{n} p_i r_i - \beta D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{\pi}_{\mathrm{ref}}). \tag{D.2}$$

Definition D.2 (Composite RLHF Social Choice Function)**.** The complete RLHF process is represented as a composite social choice function $f : \mathcal{T}^m \to \mathbb{S}(\mathcal{Y})$ defined by:

$$f(\boldsymbol{\tau}) = (s_3 \circ s_2)(\boldsymbol{\tau}),$$

where f directly maps the human preference profile $\boldsymbol{\tau}$ to a final ranking.

For a full-support reference policy and $\beta > 0$, Eq. (D.2) has the unique solution

$$p_i = \frac{\pi_{\mathrm{ref}}(y_i)e^{r_i/\beta}}{\sum_j \pi_{\mathrm{ref}}(y_j)e^{r_j/\beta}}, \tag{D.3}$$

which follows from the Lagrange condition for the simplex constraint, or by completing the KL divergence to this distribution. Taking log ratios, for every pair,

$$p_i > p_j \quad\Longleftrightarrow\quad r_i - r_j > \beta\log\frac{\pi_{\mathrm{ref}}(y_j)}{\pi_{\mathrm{ref}}(y_i)}. \tag{D.4}$$

A uniform reference therefore preserves the reward weak order, and in general a strict reward comparison survives exactly when its margin exceeds the reference log-ratio. The rest of this appendix applies Eq. (D.4) to the estimators of the main text.

D.2 POLICY OPTIMIZATION CANNOT REPAIR THE BORDA VIOLATIONS

Since $s_2(\boldsymbol{\tau})$ violates Axioms (2)–(4) under standard BT (Section 3), one might wonder whether Step (3) of RLHF could reverse this outcome, allowing the composite function $(s_3 \circ s_2)(\boldsymbol{\tau})$ to satisfy them. The following result shows that it cannot.

Theorem D.1. *Assume $n \geq 3$ and fix any $\lambda > 0$. Given any full-support reference policy $\boldsymbol{\pi}_{\mathrm{ref}}$ over $\mathcal{Y} = \{y_1, \ldots, y_n\}$, there exists a preference profile $\boldsymbol{\tau}$ such that, for every $\beta > 0$, the regularized RLHF rule $f(\boldsymbol{\tau}) = (s_3 \circ s_2)(\boldsymbol{\tau})$ does not rank the majority winner first, does not output the PMC ranking, and does not rank the Condorcet winner first, even though all three exist for $\boldsymbol{\tau}$.*

Consequently, the RLHF process f is neither majority consistent, pairwise majority consistent, nor Condorcet consistent. This suggests that RLHF may struggle to faithfully align with human preferences.

Proof. Pick distinct $a, b, c \in \mathcal{Y}$ with $\pi_{\mathrm{ref}}(a) \geq \pi_{\mathrm{ref}}(b)$. Consider the profile with five voters: 3 voters have ranking $b \succ a \succ c$ and 2 voters have ranking $a \succ c \succ b$. If $n > 3$, rank all remaining candidates below $\{a, b, c\}$ in the same fixed order for every voter.

Then b is ranked first by $3/5 > 1/2$ of the voters, so b is the majority winner. Moreover, $P_{ba} = 3/5$, $P_{bc} = 3/5$, and $P_{ac} = 1$, so b is the Condorcet winner and the pairwise-majority-consistent ranking is $b \succ a \succ c$.

By the Borda characterization of the regularized reward-modeling step (Theorem 3.1 with $q_{ij} = P_{ij}$), $r_i > r_j$ if and only if $\mathrm{BC}_i > \mathrm{BC}_j$, where $\mathrm{BC}_i = \sum_{k \neq i} P_{ik}$. Here

$$\mathrm{BC}_b = \frac{3}{5} + \frac{3}{5} = \frac{6}{5}, \qquad \mathrm{BC}_a = \frac{2}{5} + 1 = \frac{7}{5}, \qquad \mathrm{BC}_c = 0 + \frac{2}{5} = \frac{2}{5}.$$

Hence $r_a > r_b > r_c$.

By Eq. (D.3), the induced ranking is the same as that of $r_i + \beta \log \pi_{\mathrm{ref}}(y_i)$. Since $r_a > r_b$ and $\pi_{\mathrm{ref}}(a) \geq \pi_{\mathrm{ref}}(b)$, we have

$$r_a + \beta \log \pi_{\mathrm{ref}}(a) > r_b + \beta \log \pi_{\mathrm{ref}}(b) \qquad \text{for all } \beta > 0.$$

Therefore the RLHF output ranks a above b for every $\beta > 0$.

Thus RLHF does not rank the majority winner first, does not rank the Condorcet winner first, and does not output the pairwise-majority-consistent ranking. Hence it is neither MC, nor PMC, nor CC. □

D.3 Transfer of reward-stage guarantees

Theorem D.2. *Consider the regularized RLHF process $f(\boldsymbol{\tau}) = (s_3 \circ s_2)(\boldsymbol{\tau})$, where $s_2(\boldsymbol{\tau}) = r^\star$ is the minimizer of Eq.* (2.3) *with balanced weights $q_{ij} + q_{ji} = 1$ and scores $b_k = \sum_{j \neq k} q_{kj}$, and $s_3(\boldsymbol{r})$ is given by Eq.* (D.2) *with a full-support reference policy. For a profile $\boldsymbol{\tau}$ with reward vector $\boldsymbol{r} = s_2(\boldsymbol{\tau})$, define*

$$\beta^*(\boldsymbol{\tau}) = \min_{i,j:\, b_i > b_j,\ \pi_{\mathrm{ref}}(y_i) < \pi_{\mathrm{ref}}(y_j)} \frac{r_i - r_j}{\log \pi_{\mathrm{ref}}(y_j) - \log \pi_{\mathrm{ref}}(y_i)},$$

with the minimum over an empty set understood as $+\infty$. If $\beta < \beta^*(\boldsymbol{\tau})$, then $f(\boldsymbol{\tau})$ ranks y_i above y_j for every pair with $b_i > b_j$. Consequently, for each axiom among (2)–(4), if the ranking by b satisfies that axiom on $\boldsymbol{\tau}$, then so does $f(\boldsymbol{\tau})$. In particular, with the Copeland loss, f satisfies MC, PMC, and CC on every complete profile without pairwise majority ties for which $\beta < \beta^*(\boldsymbol{\tau})$.

Proof. By Theorem 3.1, $r_i > r_j$ if and only if $b_i > b_j$. By Eq. (D.4), $p_i > p_j$ if and only if $\log \pi_{\mathrm{ref}}(y_i) - \log \pi_{\mathrm{ref}}(y_j) + (r_i - r_j)/\beta > 0$. Fix a pair with $b_i > b_j$, so that $r_i > r_j$. If $\pi_{\mathrm{ref}}(y_i) \geq \pi_{\mathrm{ref}}(y_j)$, both terms are nonnegative and the second is positive, so $p_i > p_j$ for every $\beta > 0$. If $\pi_{\mathrm{ref}}(y_i) < \pi_{\mathrm{ref}}(y_j)$, the pair enters the minimum defining $\beta^*(\boldsymbol{\tau})$, so $\beta < \beta^*(\boldsymbol{\tau})$ gives

$$\beta < \frac{r_i - r_j}{\log \pi_{\mathrm{ref}}(y_j) - \log \pi_{\mathrm{ref}}(y_i)}, \qquad \text{equivalently} \qquad \frac{r_i - r_j}{\beta} > \log \pi_{\mathrm{ref}}(y_j) - \log \pi_{\mathrm{ref}}(y_i),$$

which is again $p_i > p_j$. Hence $f(\boldsymbol{\tau})$ ranks y_i above y_j for every pair with $b_i > b_j$.

Each of Axioms (2)–(4) only requires certain strict comparisons: the majority winner or the Condorcet winner above every other response, or every comparison of the pairwise-majority ranking. If the ranking by b satisfies the axiom on $\boldsymbol{\tau}$, each required comparison $y_i \succ y_j$ has $b_i > b_j$, and is therefore preserved by $f(\boldsymbol{\tau})$. The final claim follows because the Copeland score satisfies Axioms (2)–(4) on complete profiles without pairwise majority ties (Appendix A.2). □

We do not claim a converse. For $\beta \geq \beta^*(\boldsymbol{\tau})$ the policy fails to rank y_i strictly above y_j for at least one pair with $b_i > b_j$, which may or may not break a given axiom on that profile, and a profile on which b violates an axiom may, for some β, have that violation undone by π_{ref}.

D.4 Linear rewards

Applied to the linear estimator, Eq. (D.4) shows that the optimal KL-regularized policy preserves $y_i \succ y_j$ precisely when

$$r_i^{\mathrm{lin}} - r_j^{\mathrm{lin}} > \beta \log \frac{\pi_{\mathrm{ref}}(y_j)}{\pi_{\mathrm{ref}}(y_i)}. \tag{D.5}$$

Eq. (B.3) yields the sufficient condition

$$r_i^\star - r_j^\star > \sqrt{2}\, \varepsilon_\Phi + \beta \log \frac{\pi_{\mathrm{ref}}(y_j)}{\pi_{\mathrm{ref}}(y_i)}.$$

Only comparisons required by the axiom in question need this bound. For example, CC requires the winner to outrank each rival. Reward-stage guarantees therefore transfer only when the relevant margins also dominate the reference-policy bias.

Corollary D.3 (End-to-end implementation of a feasible ranking)**.** *Let θ_ρ induce a strict feasible ranking ρ. For a full-support reference and any $\beta > 0$, define*

$$t^\star = \beta \max_{i \succ_\rho j} \frac{[\log(\pi_{\mathrm{ref}}(y_j)/\pi_{\mathrm{ref}}(y_i))]_+}{\langle \theta_\rho, \phi_i - \phi_j \rangle}. \tag{D.6}$$

Every $t > t^\star$ makes the policy induced by $t\theta_\rho$ follow ρ.

Proof. All denominators are positive. For each comparison, multiplying its reward gap by t makes it exceed the reference log-ratio threshold in Eq. (D.5). A nonpositive log ratio requires only $t > 0$; otherwise the corresponding fraction gives the exact threshold. Taking the maximum proves the claim. □

This is useful after selecting an LCPO ranking (Appendix B.6), provided reward scale is unconstrained. Rescaling a fitted BT reward generally changes its training objective and is not asserted to remain an estimator.

D.5 Polynomial rewards

Let $\pi^{(k)}$ denote the policy of Eq. (D.3) with rewards $r^{(k)}$. By Eq. (D.4),

$$\pi^{(k)}(y_i) > \pi^{(k)}(y_j) \quad \Longleftrightarrow \quad r_i^{(k)} - r_j^{(k)} > \beta \log \frac{\pi_{\mathrm{ref}}(y_j)}{\pi_{\mathrm{ref}}(y_i)}. \tag{D.7}$$

Combined with Eq. (C.2), this gives the sufficient condition

$$r_i^\star - r_j^\star > \sqrt{2}\,\varepsilon_k + \beta \log \frac{\pi_{\mathrm{ref}}(y_j)}{\pi_{\mathrm{ref}}(y_i)}. \tag{D.8}$$

To transfer an axiom, apply this condition to all comparisons required by that axiom, such as the winner against every rival for CC. At interpolation the error term vanishes. A sufficiently small β then preserves every strict reward comparison on the fixed finite set; this threshold can depend on the profile and the reference.

D.6 Experiments on the policy stage

To our knowledge, the composite rule $s_3 \circ s_2$ has not been evaluated empirically. We compose the unrestricted reward estimator under majority labels with the KL-regularized policy of Eq. (D.3), using the reference model's own likelihood of each response as π_{ref}. The predicted threshold $\beta^*(\boldsymbol{\tau})$ of Theorem D.2 is exact: over the entire β grid and both variants of π_{ref} there is not a single prompt where the predicted and realized preservation of the reward order disagree (Table 3). The Copeland guarantee is intact for $\beta \leq 0.1$ and collapses by $\beta = 10$ (Figure 2). The scale of the threshold, however, is set entirely by how the reference distribution over responses is defined: the median β^* is 3.99 under a length-normalized reference and 0.0089 under the raw sequence likelihood. A reward-stage axiom guarantee is thus not something the pipeline inherits for free.

E Additional Experimental Details and Results

This appendix gives the full setup and results of the experiments summarized in Section 6. Experiments on the policy stage are reported in Appendix D.6.

E.1 Controlled simulation

Setup. We draw $n = 10$ candidates with generic features in $\mathbb{R}^2$ and $m = 5$ voters whose rankings are induced by linear utilities, so every voter ranking is linearly feasible. Holding the profiles fixed, we vary only the reward model's view of the candidates: either the same generic features, for which

Table 3: The policy stage on 5,553 held-out prompts, with the unrestricted reward estimator and majority labels. β^* is the threshold $\beta^*(\boldsymbol{\tau})$ of Theorem D.2; "mismatches" counts prompts over the whole β grid where the predicted and realized preservation of the reward order disagree, so 0 means the threshold is exact. The last four columns give Condorcet consistency of the composite rule $s_3 \circ s_2$.

		β^* quantiles				CC of $s_3 \circ s_2$ at $\beta =$			
π_{ref}	finite	.25	.50	.75	mismatches	0.01	0.1	1	10
length-normalized	0.833	2.24	3.99	7.8	0	1.000	1.000	0.988	0.549
raw	0.910	0.00512	0.0089	0.0188	0	0.706	0.248	0.170	0.161

π_{ref} is a softmax over the reference model's own log-likelihood of each response, either length normalized or raw. The reward-stage guarantee survives policy optimization only below a prompt-specific threshold, and the scale of that threshold is set entirely by how the reference distribution over responses is defined – a median β^* of 3.99 against 0.0089.

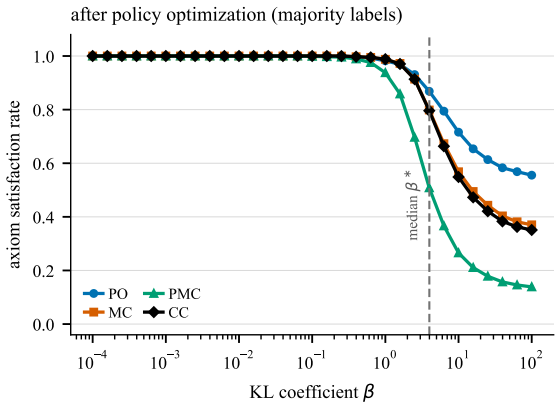


Figure 2: Axioms of the composite rule $s_3 \circ s_2$ under majority labels against the KL coefficient, with a length-normalized reference policy, on 5,553 held-out UltraFeedback prompts. The reward-stage guarantee survives only below a prompt-specific threshold, and the predicted β^* matches the realized one on every prompt.

$k_* = 3$, or collinear features, for which $k_* = n - 1 = 9$. Both losses are fitted at $\lambda = 0.1$ by Newton's method on an orthonormal basis of the feasible score space, with first-order residuals below 10^{-13}. We use $m = 5$ rather than a larger electorate because with ten candidates and $m = 21$ a unanimous pair never occurs and a majority winner occurs in 9% of profiles, so PO and MC would not be measurable at all; m is odd, so pairwise majority ties cannot arise. Every axiom is evaluated only on the profiles to which it applies, and applicability counts are reported.

The reward-stage results hold exactly. Table 4 checks each statement instance by instance over 13,000 constrained fits. The unrestricted minimizer orders the candidates by b in every instance, the projected first-order condition holds to solver precision, the recovery bound of Propositions 4.1 and 5.1 is never violated, the Bregman gap is monotone in the degree, and $r^{(k)} = r^\star$ for every $k \geq k_*$, with k_* taking the predicted values 3 and 9. The largest observed ratio $\|r^{(k)} - r^\star\|_2/\varepsilon_k$ is 0.23: the bound is valid but conservative, as expected of a global estimate that assumes no local curvature.

The interpolation degree is sufficient, not necessary. Panels (a) and (b) of Figure 3 show exact recovery of the target order against the degree. Recovery is complete at $k = k_*$ in both geometries, as the theorem requires, but it occurs well before the threshold for many individual profiles — already for 41% of profiles at $k = 1$ under majority labels in the generic geometry. The threshold governs representability of *every* score vector on the candidate set; it is not a per-profile requirement, and the gap between the two geometries is driven by feature geometry rather than by the number of coefficients.

Expressivity and the axioms move in opposite directions. Panel (c) and Table 5 isolate the paper's central point. Under majority labels, increasing the degree improves every axiom monotonically and

Table 4: Numerical verification of the reward-stage results on 13,000 constrained fits ($n = 10$ candidates, $d = 2$, $m = 5$ linearly feasible voters, $\lambda = 0.1$, 500 profiles per geometry). Each row is checked instance by instance; "pass rate" is the fraction of fits satisfying it.

Statement	Quantity checked	Pass rate
Unrestricted order = aggregation order	$r_i^\star > r_j^\star \iff b_i > b_j$	1.000
First-order condition on $\mathcal{S}_k$	$\|U_k^\top(\widehat{b}(r^{(k)}) - b + \lambda r^{(k)})\|_2 < 10^{-10}$	1.000
Recovery bound	$\|r^{(k)} - r^\star\|_2 \le \sqrt{1 + n/4\lambda}\,\delta_k$	1.000
Bregman monotonicity	$D_A(r^{(k+1)}, r^\star) \le D_A(r^{(k)}, r^\star)$	1.000
Interpolation threshold (generic)	$k_* = \min\{k : \binom{d+k}{k} \ge n\} = 3$	1.000
Interpolation threshold (collinear)	$k_* = n - 1 = 9$	1.000
Exact interpolation	$r^{(k)} = r^\star$ for $k \ge k_*$	1.000

Largest observed first-order residual $\max \|U_k^\top \nabla F\|_2 = 1.0e-13$; largest observed ratio $\|r^{(k)} - r^\star\|_2 / \varepsilon_k = 0.24$, so the bound holds with room to spare rather than being tight.

all four reach 1.000 at $k = k_*$. Under standard labels the same increase in expressivity makes matters *worse*: majority consistency falls from 0.854 at $k = 1$ to 0.714 at $k = 3$, and Condorcet consistency from 0.832 to 0.719, on an unchanged set of applicable profiles. A more expressive reward model converges faster to the unrestricted solution, and under standard labels that solution is Borda, which violates these axioms. Expressivity therefore controls how accurately the model realizes its target rule, and the loss controls which rule that is.

A restricted representation can break even Pareto optimality. With collinear features and $k = 1$, PO holds on only 0.134 of the applicable profiles under standard labels and 0.114 under majority labels, rising to 1.000 only at $k = k_* = 9$. This is the finite-sample counterpart of the impossibility results of Ge et al. (2024a): a representation too poor to express the unrestricted solution can reverse a unanimously preferred pair, and no choice of label repairs it.

The regularization boundary is real. Panel (d) restricts attention to instances whose strict majority order is linearly feasible and sweeps the score penalty. Majority-label fitting returns the majority order on every instance at $\lambda = 10^{-4}$, below the median value 1.3×10^{-3} of the threshold λ_ρ of Proposition B.4, and the rate decays to 0.843 for large λ. On the unanimous instance of Proposition B.5, Pareto optimality first fails at $\lambda \approx 0.05$, so the sufficient condition $\lambda \ge 100$ proved there is far from tight. Feasibility of the target ranking is thus not by itself enough once the estimator is required to be finite: this is exactly where a fixed positive penalty departs from the unregularized infimal rules analyzed by Ge et al. (2024a).

E.2 Reward models on UltraFeedback

Setup. We use the original UltraFeedback (Cui et al., 2024), in which each instruction carries four completions and each completion is scored on four annotation aspects. The four aspects act as four voters — a constitutional reading of the data, not four independent human annotators. An aspect that scores two completions equally splits its vote in half, and a prompt with any missing rating is dropped entirely, so every pair of a retained profile faces the same four-voter electorate; $55{,}527$ of the $63{,}967$ instructions survive. Prompt-response pairs are encoded by a frozen `Qwen2.5-1.5B-Instruct` (Qwen Team, 2024); a fixed random projection to d dimensions gives the reward features, and taking the first d columns of one projection makes the feasible score spaces nested, the real-data analogue of increasing the polynomial degree. Heads are fitted by L-BFGS on the objective in Eq. (2.3) summed over prompts, with $\lambda = 0.1$ penalizing candidate scores rather than coefficients, on an $80/10/10$ split by prompt. Freezing the encoder is essential: it is what makes the comparison one of reward expressivity over a fixed representation.

The aggregation rule is decisive in the unrestricted limit. Borda and Copeland disagree on 18.3% of prompts. On those contested prompts the unrestricted Borda solution returns the pairwise majority ranking in *none* of the 52 cases where one exists, while the unrestricted Copeland solution returns it in all 52; majority and Condorcet consistency are 0.964 and 0.971 under Borda against 1.000 under Copeland (Table 6b). Pooled over all held-out prompts these gaps almost vanish — Borda reaches

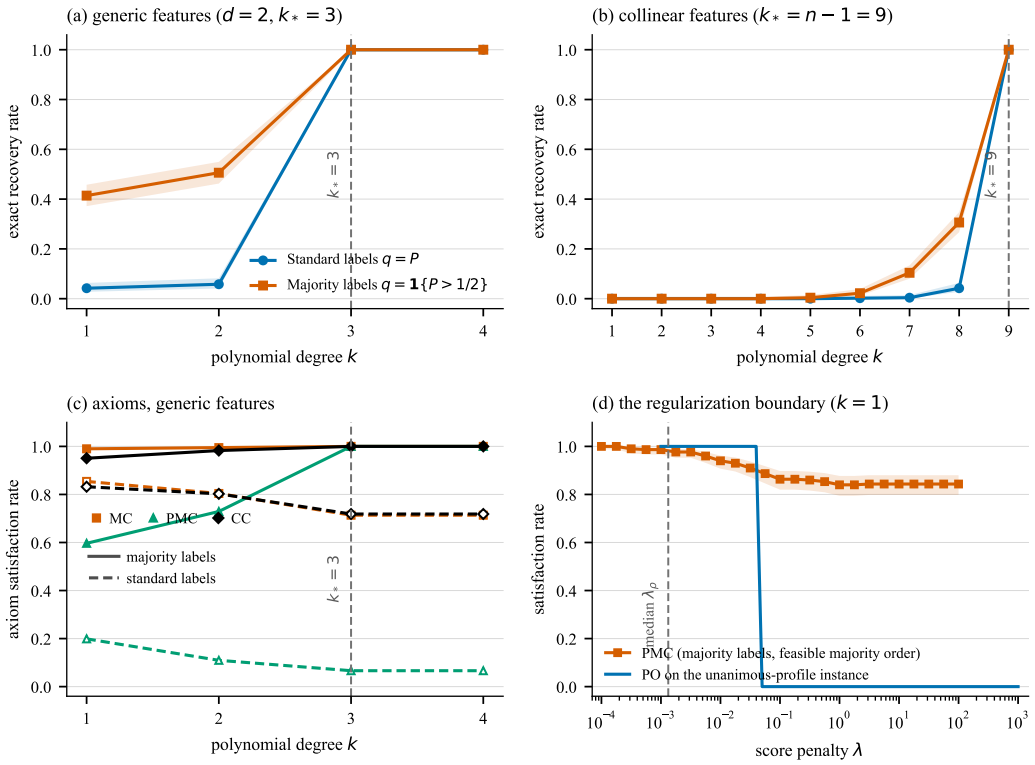


Figure 3: Controlled simulation with $n = 10$ candidates, $m = 5$ linearly feasible voters and score penalty $\lambda = 0.1$; shaded bands are Wilson 95% intervals over 500 profiles per geometry. **(a)** and **(b)** exact recovery of the target aggregation order (Borda under standard labels, Copeland under majority labels) against the polynomial degree, for a generic feature geometry ($k_* = 3$) and a collinear one ($k_* = n - 1 = 9$). Recovery is complete at k_*, as Theorem 5.1 requires, but frequently occurs earlier: the interpolation degree is sufficient for every score vector, not necessary for a given profile. **(c)** axiom satisfaction against degree in the generic geometry. Under majority labels every axiom improves to 1.000 at k_*; under standard labels majority and Condorcet consistency *deteriorate* as the model becomes more expressive, because the model is converging to Borda. Pareto optimality is omitted here because it is 1.000 for standard labels at every degree and at least 0.993 for majority labels; its interesting behavior is the collinear geometry (Table 5) and panel (d). **(d)** the regularization boundary at $k = 1$: on instances whose strict majority order is linearly feasible, majority-label fitting returns that order for small λ and fails increasingly often as λ grows, while on the unanimous instance of Proposition B.5 Pareto optimality fails from $\lambda \approx 0.05$ onwards.

0.978 on PMC — purely because the two rules coincide on the other 82% of prompts. Reporting only the pooled numbers would make the choice of loss look immaterial.

A representation-limited head cannot express either rule. The learned heads are far from their targets. The most expressive one, a linear head on the full 1536-dimensional representation, reaches pairwise accuracy 0.760 and reproduces the target order exactly on 13.5% of prompts, with Condorcet consistency 0.584 against the unrestricted 0.995. Figure 4 tracks this across the nested ladder: as d grows from 4 to 1536, PO rises from 0.428 to 0.718, MC from 0.334 to 0.612, PMC from 0.087 to 0.255 and CC from 0.329 to 0.584. Expressivity improves every axiom monotonically, exactly as representation error should.

But the two losses remain indistinguishable throughout. Across the whole ladder the largest difference between standard and majority labels on any axiom at any dimension is 0.0085. The same holds for the quadratic head, and it holds on the contested subset as well. This is the real-data instance of the low-expressivity regime of Appendix E.1: with collinear features at $k = 1$ the simulation also gives near-identical, and uniformly poor, axiom rates under the two losses (0.134 against 0.114 on

Table 5: Reward-stage behavior as polynomial expressivity grows. "Exact" is the fraction of profiles whose fitted reward induces the target aggregation order (Borda for standard labels, Copeland for majority labels); the axiom columns report the satisfaction rate among the profiles to which that axiom applies. $n = 10$, $d = 2$, $m = 5$, $\lambda = 0.1$, 500 profiles per geometry.

Features	Labels	k	Exact	PO	MC	PMC	CC	$\|r^{(k)} - r^\star\|_2$
generic ($k_* = 3$)	standard	1	0.042	1.000	0.854	0.199	0.832	0.45
		2	0.058	1.000	0.804	0.110	0.802	0.26
		3	1.000	1.000	0.714	0.066	0.719	0.00
	majority	1	0.414	1.000	0.990	0.597	0.951	2.02
		2	0.506	0.993	0.995	0.729	0.983	1.22
		3	1.000	1.000	1.000	1.000	1.000	0.00
collinear ($k_* = 9$)	standard	1	0.000	0.134	0.180	0.000	0.171	1.63
		5	0.000	0.450	0.366	0.000	0.405	1.12
		8	0.042	0.893	0.605	0.030	0.656	0.46
		9	1.000	1.000	0.659	0.048	0.698	0.00
	majority	1	0.000	0.114	0.176	0.000	0.173	6.73
		5	0.004	0.369	0.439	0.006	0.440	4.83
		8	0.306	0.812	0.805	0.458	0.796	2.02
		9	1.000	1.000	1.000	1.000	1.000	0.00

Applicable profiles (generic / collinear): PO 149/149, MC 199/205, PMC 347/334, CC 405/398 out of 500. With $m = 21$ voters instead of $m = 5$ a unanimous pair never occurs and a majority winner occurs in 9% of profiles, so PO and MC would not be measurable; $m = 5$ is odd, so pairwise majority ties cannot arise.

PO), and the two only separate once the fitted reward approaches its unrestricted target. Switching to majority labels is therefore necessary for the axioms but far from sufficient in practice — on a reward model this far from its unrestricted solution, representation error, not the aggregation objective, is what binds.

Two incidental observations. First, the quadratic head at $d = 64$ (2144 parameters) behaves like a linear head at a few hundred dimensions — pairwise accuracy 0.723 against 0.707 at $d = 128$ and 0.727 at $d = 256$, CC 0.512 against 0.499 and 0.526 — so what matters is the useful dimension of the feasible score space, not whether the extra directions come from more features or from monomials of existing ones, and the monomial directions buy very little per parameter. Second, sweeping the score penalty over $\lambda \in [10^{-3}, 10]$ shrinks $\|\theta\|$ from 1.22 to 0.08 while moving every rank-based metric by at most 0.005: with a head shared across tens of thousands of prompts the data term dominates and shrinkage barely rotates θ. The sharp λ-dependence of Appendix E.1 is a property of the per-prompt estimator the theory analyzes, and does not transfer automatically to an amortized reward model.

E.3 End-to-end and larger reward models

Implementation. The 7B experiments re-encode every prompt–response pair with a frozen `Qwen2.5-7B-Instruct` (Qwen Team, 2024) exactly as for the 1.5B model: the chat template is applied, sequences are truncated at $1{,}024$ tokens, and the last-token hidden state of the final layer ($d_{\text{model}} = 3584$) is stored in half precision. Projections take the first d columns of one fixed 3584×3584 Gaussian matrix scaled by $3584^{-1/2}$, and heads are fitted as in Appendix E.2. The LoRA reward model adds rank-16 adapters (Hu et al., 2022) ($\alpha = 32$, no dropout) to all attention and MLP projections of `Qwen2.5-1.5B-Instruct` and a bias-free scalar head on the last-token state, 18.5M trainable parameters in total, trained in bfloat16 with gradient checkpointing. Each minibatch holds 24 prompts, and the loss is the objective in Eq. (2.3) averaged over prompts, with $\lambda = 0.1$ applied to the centered scores of each prompt. We train for one epoch with AdamW at learning rate 10^{-4} and a one-cycle schedule with 5% warm-up. Weight decay is set to zero so that the only regularizer is the score penalty of the theory. For time reasons LoRA uses $20{,}000$ of the $44{,}421$ training prompts; the matched frozen head in Table 7 uses the same subset.

An end-to-end trained reward model behaves the same way. A frozen encoder invites the objection that the model is simply too weak, so we also fine-tune the backbone with LoRA, everything else held fixed. This is a much stronger model: pairwise accuracy rises from 0.760 for the best frozen

Table 6: Reward heads on frozen `Qwen2.5-1.5B-Instruct` representations of UltraFeedback, with the four annotation aspects acting as four voters (*principles as voters*, not four independent human annotators). The encoder is frozen and the projection fixed, so the two heads differ only in the expressivity of the reward model over an identical representation, while the two label sets differ only in the aggregation rule they target. The LoRA rows, where present, instead fine-tune the backbone end to end: that is what a practitioner ships, but it changes the representation itself, so those rows speak to practice rather than to the interpolation theorem. The "(7B)" rows repeat the frozen heads with the same projection construction and dimension on `Qwen2.5-7B-Instruct` ($d_{\text{model}} = 3584$), to check that the result is not an artefact of a small encoder. "Agree" is the fraction of held-out prompts whose predicted order matches the row's target order; axiom columns are rates among applicable prompts. $\lambda = 0.1$, projection dimension 64, 5,554 test prompts.

Head	Labels	Agree	Pair acc.	PO	MC	PMC	CC	BoN-CW	Rating
(a) all held-out prompts ($n = 5{,}554$)									
unrestricted	standard	1.000	0.998	1.000	0.995	0.978	0.995	0.998	4.657
	majority	1.000	1.000	1.000	1.000	1.000	1.000	1.000	4.654
linear	standard	0.096	0.692	0.600	0.482	0.173	0.464	0.467	4.153
	majority	0.078	0.692	0.592	0.487	0.168	0.470	0.474	4.151
quadratic	standard	0.111	0.723	0.647	0.536	0.210	0.512	0.515	4.224
	majority	0.097	0.723	0.643	0.543	0.209	0.521	0.524	4.225
linear (7B)	standard	0.152	0.778	0.785	0.636	0.279	0.592	0.595	4.377
	majority	0.128	0.780	0.783	0.642	0.281	0.598	0.601	4.377
quadratic (7B)	standard	0.173	0.798	0.819	0.666	0.322	0.620	0.623	4.403
	majority	0.145	0.798	0.817	0.669	0.318	0.623	0.626	4.402
LoRA (trained)	standard	0.199	0.824	0.849	0.707	0.366	0.665	0.669	4.432
	majority	0.168	0.822	0.844	0.710	0.364	0.668	0.671	4.432
(b) contested prompts, where Borda and Copeland disagree ($n = 1{,}004$)									
unrestricted	standard	1.000	0.988	1.000	0.964	0.000	0.971	0.987	4.470
	majority	1.000	1.000	1.000	1.000	1.000	1.000	1.000	4.450
linear	standard	0.096	0.675	0.605	0.481	0.077	0.444	0.447	3.928
	majority	0.006	0.672	0.596	0.488	0.058	0.450	0.454	3.930
quadratic	standard	0.096	0.705	0.656	0.577	0.154	0.529	0.532	4.010
	majority	0.013	0.703	0.654	0.592	0.135	0.542	0.545	4.008
linear (7B)	standard	0.144	0.764	0.768	0.645	0.154	0.587	0.591	4.123
	majority	0.017	0.765	0.767	0.659	0.154	0.592	0.595	4.131
quadratic (7B)	standard	0.165	0.780	0.801	0.693	0.192	0.631	0.635	4.168
	majority	0.020	0.778	0.797	0.679	0.192	0.615	0.618	4.147
LoRA (trained)	standard	0.183	0.802	0.835	0.710	0.192	0.646	0.651	4.181
	majority	0.020	0.801	0.835	0.720	0.250	0.653	0.657	4.179

"BoN-CW" is the best-of-4 hit rate on the Condorcet winner when one exists; "Rating" is the mean of the four aspect ratings of the selected response. Profile statistics over all 55,527 retained prompts: a Condorcet winner exists for 0.690, a majority winner for 0.517, the strict majority relation is a total order for 0.422, some pair is unanimous for 0.631, and Borda and Copeland disagree for 0.183. Pairwise majority ties occur in 0.578 of prompts (0.139 of pairs on average); such ties are split $1/2$–$1/2$, so the Copeland loss here is the tie-tolerant version and the no-ties hypothesis does not hold on every prompt. Applicable prompts, all / contested: PO 3510/587 MC 2921/414 PMC 2374/52 CC 3887/613. The LoRA rows train on 20,000 prompts rather than all 44,421; a frozen linear head trained on the same 20,000 reaches Kendall 0.332 and CC 0.463 against 0.331 and 0.464 on the full set, so the improvement from LoRA is attributable to the learned representation and not to the training-set size.

head to 0.824, and the Kendall distance to the target falls from 0.271 to 0.216. The gain is due to the learned representation, not to the data budget – a frozen head trained on the same $20{,}000$ prompts is indistinguishable from one trained on all $44{,}421$ (Kendall 0.332 against 0.331).

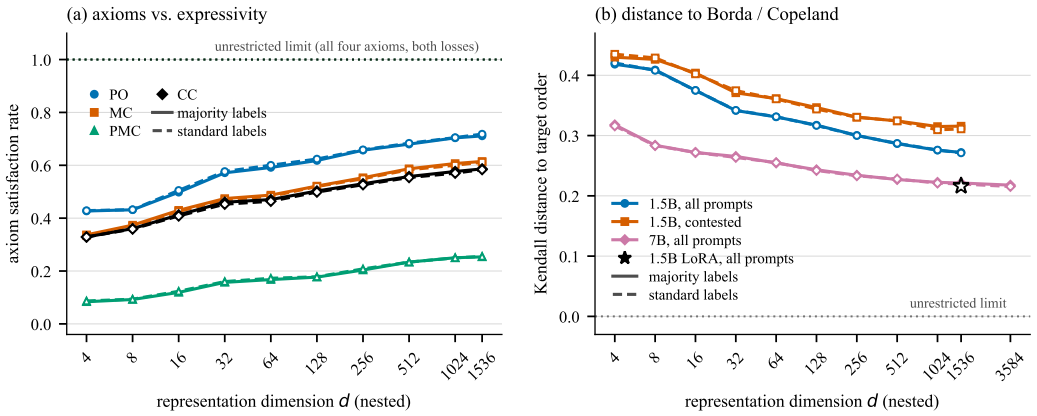


Figure 4: Reward heads on frozen `Qwen2.5-1.5B-Instruct` representations of UltraFeedback, with the four annotation aspects acting as four voters. Taking the first d columns of one fixed random projection gives nested feasible score spaces, the real-data analogue of the nested polynomial spaces of Section 5; 5,554 held-out prompts. **(a)** axiom satisfaction against d. Every axiom improves monotonically with expressivity, but the two label sets are indistinguishable at every dimension — the solid and dashed curves coincide to within 0.009 everywhere, so only one curve per axiom is visible. The dotted line marks the unrestricted limit, where the two rules separate completely (Table 6). **(b)** tie-aware Kendall distance to the row's target order (Borda for standard labels, Copeland for majority labels), on all held-out prompts and on the contested ones where the two rules disagree. Even the full $d = 1536$ linear head remains far from either target. The purple curve repeats the ladder on frozen `Qwen2.5-7B-Instruct` features up to its full width $d = 3584$, and the star is the 1.5B model fine-tuned end to end with LoRA: a larger encoder and a trained backbone each move the curve down by the same amount, to Kendall 0.216, and in neither case do the two label sets separate. Expressivity is what moves these curves; the loss is what the curves would be converging to, and at this level of representation error that difference is not yet visible in the output.

The two losses nevertheless remain indistinguishable: the largest difference on any axiom is 0.006 over all held-out prompts and 0.058 on the contested ones. This is what the out-of-sample capacity ladder of Appendix E.4 predicts, and the prediction is quantitative. A Kendall distance of 0.216 corresponds to roughly $C = 64$ on the in-sample ladder, where the contested gap is 0.115; a model that must also generalize lands below that, at 0.058. The capacity curve, fitted on frozen features, therefore forecasts the behavior of a model trained end to end.

So does a larger encoder. The complementary objection is that 1.5B parameters are simply too few. We therefore re-encode every response with a frozen `Qwen2.5-7B-Instruct` ($d_{\text{model}} = 3584$) and repeat both the table configuration and the full dimension ladder unchanged (Tables 6 and 7, Figure 4b). The larger encoder helps every axiom — at $d = 64$ the linear head's pairwise accuracy rises from 0.692 to 0.778 and CC from 0.464 to 0.592 — and at its full width it lands exactly where LoRA does (pairwise accuracy 0.824, Kendall 0.215 against 0.216). The gap between the two label sets does not open: it is at most 0.011 on any axiom at any dimension over all held-out prompts, and at most 0.058 on the contested ones. That largest contested gap is three of the 52 PMC-applicable prompts, and it has the opposite sign from the equally large LoRA gap, so it is sampling noise rather than an emerging effect. Scaling the encoder and training the backbone buy the same improvement, and neither brings the aggregation objective into view.

E.4 Prompt-specific capacity

When does the aggregation rule become visible? The pooled comparison above is dominated by representation error, so we sweep the one quantity the theory says should matter: how close the fitted reward can get to its unrestricted target. Partitioning prompts into C clusters and fitting one head per cluster interpolates between the amortized head ($C = 1$) and one head per prompt, which recovers the per-prompt unrestricted solution; the objective is separable, so all C heads are fitted in

Table 7: Larger encoders and end-to-end training on UltraFeedback ($5{,}554$ held-out prompts). The frozen 7B ladder uses the same nested projection construction as the 1.5B ladder of Figure 4; over all eleven dimensions, standard and majority labels differ by at most 0.011 on any axiom. Frozen rows use linear heads. The "20k" rows train on the same $20{,}000$ prompts, so the gap between the frozen $d = 64$ head and LoRA is due to the learned representation rather than to the amount of data. Columns as in Table 2.

Model	Setting	Labels	Pair acc.	PO	MC	PMC	CC	PMC (con.)	Kendall
7B frozen	$d = 4$	Standard	0.707	0.677	0.517	0.194	0.480	0.077	0.316
		Majority	0.707	0.673	0.521	0.192	0.484	0.077	0.318
	$d = 16$	Standard	0.759	0.753	0.591	0.245	0.547	0.096	0.272
		Majority	0.760	0.752	0.600	0.245	0.558	0.096	0.272
	$d = 64$	Standard	0.778	0.785	0.636	0.279	0.592	0.154	0.255
		Majority	0.780	0.783	0.642	0.281	0.598	0.154	0.255
	$d = 256$	Standard	0.803	0.831	0.672	0.310	0.625	0.173	0.233
		Majority	0.803	0.825	0.675	0.313	0.631	0.154	0.234
	$d = 1024$	Standard	0.817	0.850	0.696	0.346	0.649	0.192	0.221
		Majority	0.817	0.844	0.700	0.346	0.651	0.135	0.222
	$d = 3584$	Standard	0.824	0.864	0.707	0.363	0.659	0.173	0.215
		Majority	0.822	0.854	0.710	0.365	0.662	0.154	0.218
1.5B frozen	$d = 64$, 20k	Standard	0.691	0.598	0.482	0.168	0.463	0.077	0.332
		Majority	0.691	0.591	0.488	0.164	0.470	0.077	0.332
	$d = 1536$	Standard	0.760	0.718	0.612	0.255	0.584	0.115	0.271
		Majority	0.759	0.711	0.615	0.254	0.586	0.135	0.272
1.5B LoRA	20k	Standard	0.824	0.849	0.707	0.366	0.665	0.192	0.216
		Majority	0.822	0.844	0.710	0.364	0.668	0.250	0.217

one batched Newton solve. Table 8 and Figure 1a show the result: on the contested prompts the gap between standard and majority labels on pairwise majority consistency rises from 0.019 at $C = 1$ to 0.115 at $C = 64$, 0.519 at $C = 256$ and 0.981 at $C = 1024$, in lockstep with the Kendall distance to the target falling from 0.331 to 0.007. The mechanism the theory predicts is therefore measured, not inferred: the loss decides the outcome only once the model is close enough to realize it.

The capacity that makes the loss visible does not generalize. Figure 1b repeats the sweep out of sample, clustering and fitting on the training prompts and assigning held-out prompts to the nearest centroid. The distance to the target bottoms out at 0.306 around $C = 64$ and then deteriorates to 0.42, and the standard-vs-majority gap never exceeds 0.038 at any capacity. Reward models that must generalize across prompts sit firmly in the regime where the aggregation rule is invisible. Indeed at $C = 1$ the gap, averaged over the four axioms, is 0.003 against a standard deviation of 0.012 across five draws of the random projection: the choice of rule is below the noise floor of an arbitrary modeling choice.

E.5 Chatbot Arena

The construction above treats annotation aspects as voters. Chatbot Arena (Chiang et al., 2024) removes that interpretive step: candidates are language models, voters are anonymous human users, and the public leaderboard is itself a Bradley–Terry fit of the battles. We use the public `arena-human-preference-100k` release and take the $42{,}539$ votes among the twenty most-battled models.

Subsampling every pair to the same 63 votes makes the comparisons balanced, so Theorem 3.1 applies exactly — and it holds exactly: the regularized BT order coincides with the Borda order in all twenty independent subsamples (Table 9). On the full unbalanced data, which is what the live leaderboard uses, the two come apart (Kendall distance 0.026), so the balance hypothesis is not innocuous.

Table 8: Prompt-specific capacity interpolates between one shared reward head ($C = 1$) and one head per prompt, which recovers the per-prompt unrestricted solution. Columns 3–6 give the gap between standard and majority labels on the contested prompts, where Borda and Copeland disagree; the last two columns repeat the exercise out of sample. The gap only opens as the fitted reward approaches its unrestricted target, and it never opens at all for a model that has to generalize.

	in-sample (approximation)					out-of-sample	
C	Kendall	PO	MC	PMC	CC	Kendall	max gap
1	0.331	0.007	0.022	0.019	0.018	0.331	0.009
4	0.315	0.002	0.007	0.019	0.007	0.320	0.009
16	0.289	0.012	0.024	0.038	0.021	0.311	0.009
64	0.240	0.020	0.034	0.115	0.054	0.306	0.008
256	0.104	0.022	0.036	0.519	0.070	0.318	0.013
1024	0.007	0.000	0.034	0.981	0.026	0.376	0.013
4096	0.007	0.000	0.034	0.981	0.028	0.429	0.012
5554	0.007	0.000	0.034	0.981	0.028	0.406	0.012

Contested prompts applicable per axiom: PO 587, MC 414, PMC 52, CC 613. The one-head-per-prompt row ($C = 5{,}554$) does not reach Kendall 0 exactly because 3.6% of prompts contain two responses with identical features – duplicated completions – for which even a per-prompt head cannot realize an arbitrary score vector. At $C = 1$ the standard-vs-majority gap is 0.0033 on average, against a standard deviation of 0.0121 across 5 draws of the random projection: the choice of aggregation rule is below the noise floor of an arbitrary modeling choice.

Table 9: Borda and Copeland on 42,539 anonymous Chatbot Arena votes over the 20 most-battled models. Candidates are language models and voters are human users; the public leaderboard is itself a Bradley–Terry fit of these battles. In the *balanced* design every pair is subsampled to the same 63 votes, so the comparisons are balanced and Theorem 3.1 applies exactly; the *full* design keeps every battle, with the real unequal per-pair counts. Balanced figures are means over 20 independent subsamples.

	full (unbalanced)	balanced (63/pair)
Votes used	42,539	11,970
BT order equals Borda order	no	yes, 20/20 subsamples
Kendall distance, BT vs. Borda	0.026	0.000
Kendall distance, BT vs. Copeland	0.037	0.063
Condorcet winner exists	yes	0.50 of subsamples
… and BT ranks it first	yes	0.90
Strictly decided pairs	187	186
… reversed by BT	12	21.4
Pairs with a statistically clear majority	88	50
… reversed by BT	**0**	**0.1**
Largest reversed head-to-head margin $\lvert P - \frac{1}{2}\rvert$	0.059	0.098
Median margin over strictly decided pairs	0.070	0.076

Ties (both "tie" and "tie, both bad") count as half a vote each, matching the convention used throughout. "Statistically clear" means the head-to-head majority is significant at $z = 1.96$ under a normal approximation. On the full data the Borda and Copeland leaderboards differ in the top eight (BT: chatgpt-4o-latest, gemini-1.5-pro-exp-0801, gpt-4o-2024-05-13, gpt-4o-mini-2024-07-18, …; Copeland: chatgpt-4o-latest, gemini-1.5-pro-exp-0801, gpt-4o-2024-05-13, gpt-4o-2024-08-06, …), yet not one of the 88 statistically clear majorities is reversed.

Borda and Copeland do give different leaderboards: they disagree on 3.7% of pairs, and the two orders agree only on the first three places — two adjacent pairs are transposed and the eighth entry is a different model altogether — so five of the top eight positions change. But every disagreement sits on a near-tie. The largest head-to-head margin that the BT order reverses is $\lvert P_{ij} - \frac{1}{2}\rvert = 0.059$, and of the 88 pairs whose majority is statistically clear, *none* is reversed. On real preference data, the axiom violations of Borda thus concentrate where the electorate is close to indifferent, which is one reason why they may cost little in practice.

Table 10: Robustness of the UltraFeedback construction. "no finite MLE" is the fraction of prompts whose win graph is not strongly connected, so that the *unregularized* Bradley–Terry likelihood has no finite maximizer and the penalty $\lambda > 0$ is a necessity rather than a convenience. The right-hand blocks give the axioms of the unrestricted estimator under each label set; Pareto optimality is 1.000 in every cell and is omitted. The qualitative conclusion does not depend on how within-aspect ties are handled.

					unrestr. Borda			unrestr. Copeland		
tie convention	no MLE	B≠C	Cond.	tie	MC	PMC	CC	MC	PMC	CC
half-vote (used)	0.348	0.183	0.690	0.578	0.995	0.978	0.995	1.000	1.000	1.000
drop tied aspects	0.808	0.176	0.690	0.578	0.987	0.936	0.990	1.000	1.000	1.000
strict tie-break	0.808	0.104	0.947	0.154	0.978	0.962	0.991	0.990	1.000	1.000

$\Pr(\text{Borda order} \neq \text{Copeland order})$ for linearly feasible voters, 1500 profiles per cell. The distinction is uncommon with four candidates but near-certain with ten, which is the range real annotation pipelines use; an even electorate blunts it, because 2–2 splits make the two rules agree.

$n\backslash m$	3	4	5	7	11	21	51
3	0.14	0.05	0.19	0.22	0.24	0.24	0.26
4	0.27	0.13	0.38	0.42	0.47	0.46	0.48
5	0.43	0.20	0.54	0.60	0.65	0.66	0.67
6	0.48	0.27	0.69	0.72	0.77	0.78	0.78
8	0.64	0.38	0.82	0.86	0.92	0.92	0.91
10	0.66	0.45	0.88	0.93	0.97	0.97	0.96
12	0.70	0.45	0.90	0.96	0.98	0.99	0.99

E.6 Robustness and scope

Robustness. Table 10 reports two audits. First, the unregularized BT likelihood has no finite maximizer on 34.8% of UltraFeedback prompts under the tie convention we use, and on 80.8% under the two alternatives, so the penalty $\lambda > 0$ is a necessity rather than a convenience. Second, all qualitative conclusions survive the two alternative tie conventions: the unrestricted Copeland estimator satisfies every axiom to within 0.01 in each case, while the unrestricted Borda estimator does not. The same table records how often the two rules differ at all as a function of the electorate and the number of candidates: rarely with four candidates, almost always with ten, which is the range real annotation pipelines use.

Scope. A head shared across prompts and evaluated on held-out prompts is an empirical generalization test, not an instance of the fixed-candidate interpolation theorem: with four candidates in 64 or more dimensions the per-prompt design is generically full rank, so the threshold result has nothing to say here. Pairwise majority ties, which the Copeland result excludes, occur in 57.8% of prompts (13.9% of pairs on average) and are split evenly, so the Copeland loss used here is the tie-tolerant version.

F Other Related Work

Social choice and pluralistic alignment. Beyond the axiomatic analyses discussed in the main text, social choice informs alignment in several further ways. Policy aggregation applies voting rules such as the Borda count and approval voting to the state-action occupancy polytope of a Markov decision process (Alamdari et al., 2024); generative social choice combines representation guarantees with language models that generate and evaluate alternatives (Fish et al., 2024); and representative social choice studies aggregation when the issues and individuals are too numerous to consider directly (Qiu, 2024). A related line of work on pluralistic alignment (Sorensen et al., 2024) addresses heterogeneous preferences directly: fine-tuning language models to find statements that diverse groups agree with (Bakker et al., 2022), interpolating models fine-tuned on different rewards (Ramé et al., 2023), aligning a single policy to heterogeneous user types (Shirali et al., 2025), and eliciting truthful feedback from strategic labelers (Kleine Buening et al., 2025). Xiao et al. (2025b) show that KL-regularized RLHF can collapse onto the majority preference and propose a preference-matching

regularizer. Earlier, Noothigattu et al. (2018) aggregated learned preferences with voting rules for ethical decision making.

Preference fine-tuning. Several methods have been developed to improve preference fine-tuning in RLHF. Li et al. (2024) demonstrated that proximal policy optimization (PPO) (Schulman et al., 2017) may not fully leverage RLHF's capacity to align large language models (LLMs) with human preferences. The work by Tang et al. (2024) explored discrepancies between online and offline learning paradigms for alignment, while Ye et al. (2024) proposed an online iterative RLHF method based on a general preference structure. Additionally, Li et al. (2025) investigated how model diversity can be maintained during the supervised fine-tuning (SFT) stage. Outside of fine-tuning, model editing has emerged as a complementary strategy to modify LLM behavior across tasks (Jin et al., 2025).

Several extensions of direct preference optimization (DPO) (Rafailov et al., 2023) have been introduced to address its limitations (Liu et al., 2024; Azar et al., 2024; Chang et al., 2024; Gorbatovski et al., 2025; Rafailov et al., 2024; Yang et al., 2024). Nevertheless, recent analyses indicate that DPO may underperform compared to reward-optimized RLHF methods in terms of alignment effectiveness (Li et al., 2023; Xu et al., 2024; Tajwar et al., 2024). This performance gap has been partly attributed to representational mismatches inherent in DPO, as discussed by Li et al. (2023) and Xu et al. (2024), which constrain its robustness relative to reinforcement learning techniques like PPO. Furthermore, the on-policy nature of reward-based fine-tuning helps alleviate distribution shift between training and deployment scenarios, thereby boosting LLM reliability (Tajwar et al., 2024). The calibration challenges introduced by RLHF have also been studied by Xiao et al. (2025a).

An alternative line of work introduces Nash learning from human feedback (NLHF), which aims to align LLMs under general preference models (Munos et al., 2024; Liu et al., 2025; Shi et al., 2025). In this direction, Wang et al. (2025) examined the convergence behavior of NLHF and showed that the last iterate converges to a Nash equilibrium.