WHAT DOES STATIC REWARD-MODEL EVALUATION IDENTIFY ABOUT DOWNSTREAM OPTIMIZATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

In reinforcement learning from human feedback (RLHF), static evaluations are commonly used to select reward models (RM) before policy optimization. We ask when such evaluations are sufficient to determine which RM yields a better policy. The question matters because training a policy for every candidate is expensive, while RMs with similar static accuracy are known to produce very different policies. Two RMs can rank every response identically and have the same reward variance, yet one improves the policy while the other makes it worse. Ranking and variance indicate how strongly an RM prefers some responses over others, not whether those preferences agree with the target's. The question is therefore not how many scores to observe, but whether they are taken on responses that the target itself distinguishes. We formalize this for reward models that are linear in a fixed set of features. Under a norm-bounded linear reward class, we derive the sharp range of initial effects consistent with exact score observations, showing that the pairwise comparison can be identified without recovering every reward score. This local comparison need not persist as the policy distribution changes. We derive sufficient conditions under which it remains identified after a finite policy change, including at matched KL distance. Experiments with pretrained RM scores show that partial observations can identify the comparison after finite optimization or leave both signs attainable. Static evaluation can thus guide RM selection, but only when it observes what the target cares about, and only for limited optimization.

1 INTRODUCTION

In a common reinforcement learning from human feedback (RLHF) pipeline, a reward model (RM) is trained from preference data to assign scalar scores to model responses. The policy is then optimized to increase the expected RM score, typically under a KL constraint or penalty relative to a reference policy. Since downstream evaluation requires optimizing and evaluating a policy for each candidate RM, RMs are commonly compared beforehand using static metrics such as pairwise preference accuracy (Lambert et al., 2025; Frick et al., 2025). However, similar static accuracy can correspond to substantially different downstream policy performance (Wen et al., 2025). For RM selection, this means that similar static evaluations need not reveal which candidate will yield higher downstream utility after optimization.

Existing work has studied how the predictive value of static RM metrics depends on the evaluation responses (Wen et al., 2025), and how reward properties beyond accuracy, such as reference-policy reward variance, affect optimization dynamics (Razin et al., 2025). Separately, work on partial identifiability in reward learning and robust reward learning studies what can be concluded for downstream applications when the reward is not fully determined by the available feedback (Skalse et al., 2023; Lazzati & Metelli, 2026). The central question is when static RM evaluation is sufficient to determine the effect of RM optimization on target utility. We further ask whether that conclusion persists after finite optimization.

We fix a target reward r_G to define policy utility and study KL-regularized optimization of candidate RMs from a common reference policy π_0. A static evaluation need not uniquely determine a reward: multiple rewards can be consistent with the same observed evaluation, and the set of such rewards forms the feasible set (Skalse et al., 2023; Lazzati & Metelli, 2026). We say that the evaluation identifies a downstream quantity if that quantity takes the same value for every reward in the feasible

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.

set. We first study how optimizing a single candidate RM changes target utility. For RM selection, we then compare these utility changes across candidate RMs. A strong empirical association between a static metric and downstream performance across evaluated RMs does not by itself establish that these downstream quantities are identified.

To understand what static RM evaluation can establish about finite optimization, we separate two effects. At the common reference policy, all candidate rewards are evaluated under the same response distribution, so we can ask whether the ambiguity allowed by the static observations already leaves the initial target-utility effect undetermined. During finite optimization, different rewards induce different policy distributions, adding a second source of variation as the policy moves. We therefore use RQ1 and RQ2 to characterize what static evaluation determines at the reference policy, and RQ3 to study whether those conclusions persist under finite policy optimization. Figure 1 summarizes this two-stage analysis.

RQ1: Do complete response ordering and reward variance determine whether reward optimization initially increases or decreases target utility? No. We construct rewards with identical complete response ordering and reward variance under the same reference policy, yet with opposite initial effects on target utility. Complete ordering fixes which reward gaps are positive or negative, while reward variance summarizes their overall magnitude; neither determines how those gaps align with the target-reward gaps. This alignment determines the initial utility effect M_0. Thus, even complete ordinal information together with reward variance is insufficient to determine the direction in which reward optimization initially moves target utility.

RQ2: Which static observations are sufficient to determine the initial effect of reward optimization on target utility? RQ1 shows that ordering and reward variance are not sufficient, but this does not imply that the full reward must be recovered. For a specified linear reward class, we show that score observations identify the initial effect exactly when their span contains the target-relevant reward direction; otherwise, a norm bound gives the sharp identified interval. The same analysis gives a sharp interval for the pairwise difference, so RM selection can be identified even when neither individual effect is point identified. Which scores are observed can therefore matter more than how many.

RQ3: When does a comparison determined at the reference policy remain valid after finite optimization? A comparison determined at the reference policy need not remain valid after the policy moves. As optimization changes response probabilities, the effect of a candidate reward on target utility can change along the optimization path. We bound the deviation between the finite utility difference and its local approximation, giving sufficient conditions under which the comparison determined at the reference policy remains valid after a finite policy change. These conditions are not automatic: we also construct feasible rewards that agree on the local comparison but yield opposite finite utility differences. We analyze this question both at a common optimization strength and at a matched KL distance, since reward rescaling can change how far different rewards move the policy at the same optimization strength.

Experiments with pretrained RM scores show these distinctions in practice. Identical ordering and variance can still produce opposite utility changes, while the choice of observed scores determines whether local and finite comparisons are identified. A full-parameter REINFORCE Leave-One-Out (RLOO) run further shows that RMs with similar static accuracy can change their relative target utility over training.

2 Preliminaries and problem setup

2.1 Reward models and downstream policy quality

Let $x \sim D_X$, let $\pi_0(y \mid x)$ be an initial policy, let $r(x, y)$ be a candidate RM, and let $r_G(x, y)$ denote the fixed target reward. We evaluate a policy by

$$U(\pi) = \mathbb{E}_x \mathbb{E}_{y \sim \pi(\cdot|x)}[r_G(x, y)]. \tag{1}$$

The target r_G defines policy quality and enters the identification results through its covariance direction and score range. We instantiate r_G with ArmoRM's native scores (Wang et al., 2024).

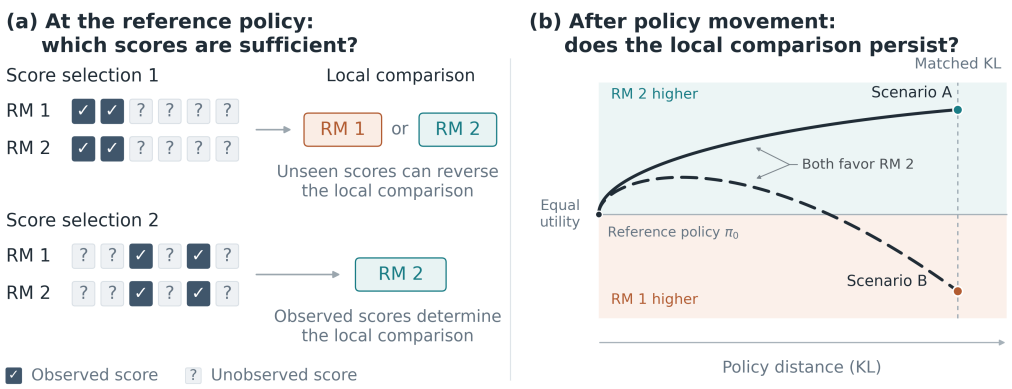


Figure 1: What static RM evaluation can determine at the reference policy and after policy movement. (a) With the same number of observed scores, the local comparison may or may not be determined depending on which scores are observed. (b) Two reward assignments consistent with the same observed scores can agree on the local comparison, yet yield opposite target-utility comparisons at the same matched KL distance after policy movement.

The candidate class is $\mathcal{R}$. For almost every prompt, response sums and extrema use the finite reference support $\mathcal{Y}_x^0 = \{y : \pi_0(y \mid x) > 0\}$. Finite reward values ensure a well-defined Gibbs normalizer.

2.2 Static RM evaluation and its response distribution

Let $e = \mathcal{E}_{D_{\text{eval}}}(r)$ denote the observation produced by a static evaluation. The rewards in $\mathcal{R}$ consistent with e form the feasible set (Lazzati & Metelli, 2026),

$$\mathcal{R}(e) = \{r \in \mathcal{R} : \mathcal{E}_{D_{\text{eval}}}(r) = e\}. \tag{2}$$

A downstream quantity is identified if it has the same value for every reward in $\mathcal{R}(e)$. Its range over $\mathcal{R}(e)$ is the identified set; point identification means this set is a singleton (Skalse et al., 2023).

For pairwise evaluation, let $X \sim D_X$ and $(Y, Y') \sim D_{\text{eval}}(\cdot, \cdot \mid X)$, and write $\Delta r_G = r_G(X, Y) - r_G(X, Y')$ and $\Delta r = r(X, Y) - r(X, Y')$. Ignoring target ties or using a fixed tie convention,

$$\text{Acc}_{D_{\text{eval}}}(r) = \Pr(\Delta r_G \Delta r > 0 \mid \Delta r_G \neq 0). \tag{3}$$

Pairwise comparisons retain only $\text{sign}(\Delta r)$, whereas score observations also retain gap magnitudes. Equation 3 uses strict agreement. Experimental tables give half credit to candidate ties and exclude target ties (Appendix B).

D_{eval} determines which response pairs are observed. For the local downstream effect studied in Section 3, we take $Q = \pi_0$. On-policy evaluation draws $Y, Y' \overset{\text{i.i.d.}}{\sim} \pi_0(\cdot \mid x)$, so the evaluation and local-effect distributions coincide. Our exact Gibbs experiments use responses generated once by the frozen initial policy.

Reward variance. For any Q, define the reward variance

$$\text{RV}_Q(r) = \mathbb{E}_x \, \text{Var}_{Y \sim Q(\cdot|x)}[r(x, Y)], \qquad \text{RV}_0(r) = \text{RV}_{\pi_0}(r). \tag{4}$$

Variance measures within-prompt score dispersion under Q (Razin et al., 2025).

2.3 Fixed-reference KL-regularized reward maximization

For $t > 0$, we analyze the exact fixed-reference objective

$$\pi_t^r \in \arg\max_{\pi} \left\{ \mathbb{E}_{x,y\sim\pi}[r(x, y)] - \frac{1}{t} \mathbb{E}_x D_{\text{KL}}(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)) \right\}. \tag{5}$$

Table 1: Static accuracy and matched-KL utility gain across five pretrained RMs. Accuracy is pairwise agreement with ArmoRM; gain is the mean ArmoRM increase under exact Gibbs optimization at mean KL 0.2, in units of 10^{-3}. Qwen uses 32 responses per prompt; H4 SFT and OpenOrca use 16.

Reward model	Qwen		H4 SFT		OpenOrca	
	Accuracy (%)	Gain (10^{-3})	Accuracy (%)	Gain (10^{-3})	Accuracy (%)	Gain (10^{-3})
Skywork Qwen3-0.6B	73.3	8.122	74.0	8.176	67.9	4.911
Skywork Qwen3-1.7B	76.5	9.154	76.8	9.212	70.3	5.468
Skywork Qwen3-4B	78.2	9.575	78.7	9.594	73.4	5.995
Skywork Llama-3.2-1B	73.8	8.539	75.7	8.480	68.2	4.852
InternLM2-1.8B	70.5	7.129	70.3	6.386	66.8	4.102

Its pointwise optimum is the exponential tilt

$$\pi_t^r(y \mid x) = \frac{\pi_0(y \mid x) e^{t r(x,y)}}{Z_t^r(x)}, \qquad Z_t^r(x) = \sum_{y'} \pi_0(y' \mid x) e^{t r(x,y')}. \tag{6}$$

Let $\pi_0^r = \pi_0$ be the continuous extension at $t = 0$. Write $G_t(r) = U(\pi_t^r) - U(\pi_0)$ and $\rho_r(t) = \mathbb{E}_x D_{\mathrm{KL}}(\pi_t^r \| \pi_0)$. Equation 6 is the exact distribution-space Gibbs optimizer (Donsker & Varadhan, 1975; Rafailov et al., 2023). Section 4.4 matches policy distance rather than t; Section 4.5 examines parameterized training.

We evaluate five pretrained RMs on fixed response banks from 128 UltraFeedback prompts (Cui et al., 2024), using ArmoRM as the target. We compute exact Gibbs and Best-of-N outcomes on these banks; Appendix B gives model identifiers, response-generation settings, and information costs.

3 From static observations to local identification

3.1 The local effect of reward maximization

For any response distribution Q, define

$$M_Q(r) = \mathbb{E}_x \operatorname{Cov}_{Y \sim Q(\cdot|x)}(r_G(x, Y), r(x, Y)), \qquad M_0(r) = M_{\pi_0}(r). \tag{7}$$

Theorem 3.1 (Derivative of downstream utility under KL-regularized reward optimization)**.** *If r_G and r are bounded on the reference-policy support and the reward r does not vary with t, then*

$$\frac{d}{dt} U(\pi_t^r) = \mathbb{E}_x \operatorname{Cov}_{\pi_t^r}(r_G, r), \qquad G_t'(r)\big|_{t=0} = M_0(r). \tag{8}$$

Thus, $M_0(r)$ is the exact first-order effect of optimizing r on target utility at the reference policy. Its sign determines whether target utility initially increases or decreases.

3.2 What ordering and reward variance leave unresolved

For i.i.d. $Y, Y' \sim \pi_0(\cdot \mid x)$,

$$M_0(r) = \tfrac{1}{2}\mathbb{E}[\Delta r_G \Delta r], \qquad \mathrm{RV}_0(r) = \tfrac{1}{2}\mathbb{E}[(\Delta r)^2]. \tag{9}$$

Complete ordering determines the signs of Δr, while RV_0 determines $\mathbb{E}[(\Delta r)^2]$. Neither determines the cross-moment $\mathbb{E}[\Delta r_G \Delta r]$, which captures how candidate reward gaps align with target-reward gaps.

Using the fixed target scores, we construct three rewards on the empirical response bank that preserve complete response ordering and reward variance while changing their alignment with the target. The three rewards have identical orderings on all 63,488 within-prompt response pairs, identical prompt means, and identical raw reward variance RV_0 (Appendix B.3). Under fixed tie-breaking, the three rewards give identical Best-of-N gains. Their matched-KL Gibbs gains, however, take both positive and negative values.

The constructions attain both signs of $M_0(r)$. Complete ordering and reward variance therefore do not identify the initial direction of target-utility change.

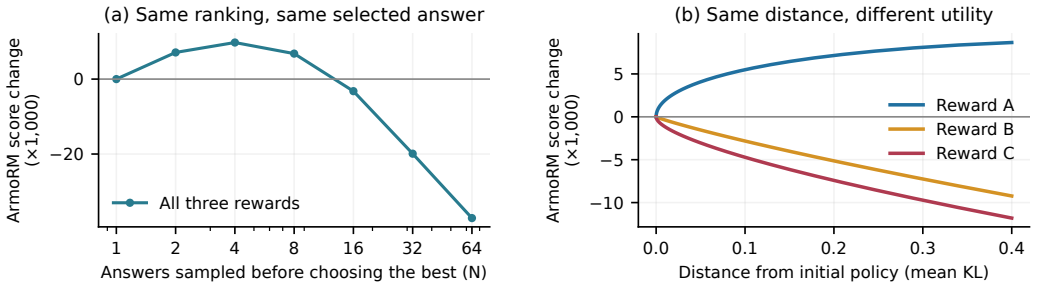


Figure 2: Identical ordering and reward variance do not determine Gibbs utility gain. Rewards A–C preserve the same response ordering, prompt means, accuracy, and RV_0, while varying the reward gap assigned to responses with low target scores. (a) Under fixed tie-breaking, Best-of-N, which depends only on the reward ordering, gives identical gains for all three rewards. (b) Exact Gibbs optimization depends on reward-gap magnitudes and yields both positive and negative target-utility gains at matched KL. Both vertical axes report mean ArmoRM gain relative to the initial policy, multiplied by 1,000 (Appendix B.3).

Static accuracy and downstream gain. Identification can fail even when static accuracy closely tracks downstream gain across the observed RMs. Across the five RMs, accuracy and matched-KL gain rank the models identically on Qwen and H4, and agree on nine of ten pairwise comparisons on OpenOrca (Table 1).

3.3 Score observations and local identification

Consider a linear reward class

$$r_w(x, y) = w^\top \phi(x, y) \tag{10}$$

and linear observations of RM scores $e = Aw$. For a fixed reference policy π_0, define

$$b = \mathbb{E}_x \mathbb{E}_{\pi_0}[(r_G - \mathbb{E}_{\pi_0}[r_G \mid x])\phi]\,. \tag{11}$$

Then $M_0(w) = b^\top w$.

Theorem 3.2 (Identification under linear score observations)**.** *$M_0(w)$ is identified from Aw if and only if $b \in \mathrm{row}(A)$, equivalently $\ker A \subseteq \ker b^\top$.*

Since $Aw = e$ leaves $\ker A$ unobserved, M_0 can still be identified without identifying the reward itself, exactly when those directions are orthogonal to b.

Partial identification. When M_0 is not point identified, a norm bound on the unobserved reward directions yields a sharp identified interval. Let

$$\mathcal{V}_R = \{v \in \ker A : \|v\|_2 \le R\}, \qquad \Gamma_{\mathrm{id}} = \|P_{\ker A} b\|_2. \tag{12}$$

Theorem 3.3 (Sharp identified interval)**.** *For every $v \in \mathcal{V}_R$,*

$$M_0(w + v) \in [M_0(w) - R\Gamma_{\mathrm{id}},\ M_0(w) + R\Gamma_{\mathrm{id}}]. \tag{13}$$

The interval is sharp. In particular, $|M_0(w)| > R\Gamma_{\mathrm{id}}$ if and only if $M_0(w + v)$ has the same strict sign for every $v \in \mathcal{V}_R$.

Γ_{id} therefore measures the part of the target-relevant direction that remains unconstrained by the observations, and $R\Gamma_{\mathrm{id}}$ is the exact worst-case variation in M_0 over the feasible set. The sign of the local effect is identified exactly when the interval excludes zero.

Pairwise comparison. For two candidates optimized from the same reference policy, define

$$\Delta M_0 = M_0(w_2) - M_0(w_1), \tag{14}$$

with independent feasible perturbations $v_i \in \mathcal{V}_{R_i}$.

Corollary 3.4 (Sharp identified interval for pairwise local comparison)**.** *The pairwise local effect lies in*

$$\begin{aligned} &\Delta M_0 \in [\Delta M_{\mathrm{ref}} - (R_1 + R_2)\Gamma_{\mathrm{id}},\ \Delta M_{\mathrm{ref}} + (R_1 + R_2)\Gamma_{\mathrm{id}}]\,, \\ &\Delta M_{\mathrm{ref}} = b^\top (w_2 - w_1). \end{aligned} \tag{15}$$

and this interval is sharp. Thus $\Delta M_{\mathrm{ref}} > (R_1 + R_2)\Gamma_{\mathrm{id}}$ *if and only if every pair of feasible rewards satisfies* $M_0(w_2 + v_2) > M_0(w_1 + v_1)$.

Because the comparison depends only on the difference, its identified interval can exclude zero even when neither individual local effect is point identified. Thus RM selection can require strictly less information than identifying either candidate reward or either candidate's M_0.

Observation design. The interval gives a direct criterion for choosing score observations: informative observations are those that reduce Γ_{id}, the target-relevant component left in the null space. We instantiate this criterion in the five-dimensional span of prompt-centered RM scores. A target-directed design greedily selects within-prompt score differences that reduce Γ_{id}^2 using orthogonal least squares (Soussen et al., 2013); a random dispersed design queries the same number of responses without using the target direction. Both retain one within-prompt score difference for every two queried responses. Details of the class, norm bound, target direction, and query construction are given in Appendix B.5.

Table 2: Target-directed observations constrain the target-relevant direction more strongly than random observations at the same score count and observation rank. The residual $\Gamma_{\mathrm{id}}/\|b\|$ measures the fraction of the target-relevant direction left in the unobserved subspace. Identified comparisons count determined signs of ΔM_0 over ten RM pairs and ten query selections.

Scores per RM	Rank	$\Gamma_{\mathrm{id}}/\|b\|$		Identified comparisons	
		Random	Directed	Random	Directed
4	$2/5$	0.83654	0.06793	0/100	89/100
6	$3/5$	0.64869	0.00485	3/100	100/100
8	$4/5$	0.35156	0.00003	36/100	100/100

At every budget, the target-directed design leaves a smaller target-direction residual and identifies more pairwise local comparisons; with six scores, $\Gamma_{\mathrm{id}}/\|b\|$ decreases from 0.64869 to 0.00485, while identified comparisons increase from $3/100$ to $100/100$.

Dependence on the target. The identification condition is target specific because b depends on r_G. For a fixed target, only b must lie in $\mathrm{row}(A)$, so other nonconstant reward directions may remain unobserved. Requiring identification for every bounded target is much stronger: it leaves only prompt-wise additive constants unobserved, which preserve the entire Gibbs path (Corollaries D.2–D.3; Appendix D.1). Our certificates therefore assume a fixed cardinal target; ordinal preference labels alone do not generally determine b. Holding the candidate class and score observations fixed while replacing ArmoRM with InternLM as the target changes which local comparisons are identified (Appendix B.10).

4 From local effects to finite optimization

In Section 3, we characterized what static observations determine about the local effect at the reference policy. Moving to finite optimization introduces a second source of uncertainty: the policy distribution itself changes as optimization proceeds. We first study this transition at a common optimization strength t, and then at a common KL distance in Section 4.4.

4.1 WHY AN INITIAL DERIVATIVE IS NOT A FINITE GUARANTEE

Even if $M_0(r)$ is point identified, it determines only the initial direction of target-utility change. For any $T > 0$,

$$G_T(r) = \int_0^T M_t(r)\, dt, \qquad M_t(r) = \mathbb{E}_x \operatorname{Cov}_{\pi_t^r}(r_G, r). \tag{16}$$

Thus $G_T(r)$ depends on the covariance along the entire optimization path. Because this covariance is evaluated under π_t^r, it can change as optimization reweights responses. Consequently, $M_0(r) > 0$ does not imply $G_T(r) > 0$; Appendix A.4 shows that target utility can initially increase but eventually fall below $U(\pi_0)$.

4.2 QUANTIFYING THE LOCAL-TO-FINITE GAP

Assume throughout this section that r_G and all candidate rewards are uniformly bounded on the reference support. Let

$$\dot{\pi}_0^r(y \mid x) = \pi_0(y \mid x)\big(r(x,y) - \mathbb{E}_{\pi_0}[r \mid x]\big) \tag{17}$$

denote the first-order tangent of π_t^r at $t = 0$. Define the deviation from this linear approximation by

$$d_t^r = \pi_t^r - \pi_0 - t\dot{\pi}_0^r, \qquad C_{t,x}(r) = \tfrac{1}{2}\|d_t^r(x,\cdot)\|_1, \tag{18}$$

and let

$$B_G(x) = \max_y r_G(x,y) - \min_y r_G(x,y) \tag{19}$$

be the within-prompt range of the target reward.

Theorem 4.1 (Finite-optimization remainder bound)**.** *For bounded r_G and r on the reference support,*

$$|G_t(r) - tM_0(r)| \le \mathbb{E}_x[B_G(x)C_{t,x}(r)]. \tag{20}$$

The bound is worst-case sharp over target rewards constrained only by their per-prompt ranges.

Equation (20) bounds the error of the local approximation $tM_0(r)$ by the deviation of π_t^r from its first-order approximation, weighted by the target-reward range $B_G(x)$.

If the within-prompt range of r is at most B_r for every prompt, then, by Proposition A.1 (Appendix A.4),

$$C_{t,x}(r) \le \frac{B_r^2 t^2}{8}. \tag{21}$$

Thus, over a feasible set with uniformly bounded rewards, uncertainty already present in M_0 contributes at order t, whereas the additional error caused by policy movement is $O(t^2)$ (Corollary A.2). Section 4.3 combines these two terms to bound the finite utility difference between two candidate RMs.

4.3 PAIRWISE COMPARISON AFTER FINITE OPTIMIZATION

For two candidate rewards optimized from the same reference policy at a common strength t, let

$$\Delta G_t = G_t(r_2) - G_t(r_1). \tag{22}$$

Suppose the static observations imply $\Delta M_0 \in [\delta_-, \delta_+]$, and let B_t^Δ uniformly bound $|\Delta G_t - t\Delta M_0|$ over feasible reward pairs; Corollary A.3 provides such a bound.

Corollary 4.2 (Pairwise finite-optimization outer bound)**.**

$$\Delta G_t \in [t\delta_- - B_t^\Delta,\ t\delta_+ + B_t^\Delta], \qquad t\delta_- > B_t^\Delta \implies \Delta G_t > 0. \tag{23}$$

The interval gives a sufficient condition for a locally determined comparison to remain valid after finite optimization; if it crosses zero, the bound is inconclusive.

In the linear setting of Section 3, Corollaries 3.4 and 4.2 give

$$|\Delta G_t - t\Delta M_{\mathrm{ref}}| \le t(R_1 + R_2)\Gamma_{\mathrm{id}} + B_t^\Delta. \tag{24}$$

Equation (24) makes the two sources of uncertainty explicit: $t(R_1 + R_2)\Gamma_{\mathrm{id}}$ comes from unobserved reward directions, while B_t^{Δ} captures the additional effect of policy movement.

Using the uniform bounds in Appendix B.6, partial score observations can determine some common-t comparisons without point identifying either reward (Appendix Table 6). Conversely, even observing each candidate's exact M_0 need not determine the finite comparison: Appendix Figure 4 gives feasible rewards with identical initial effects but opposite finite outcomes.

4.4 RM COMPARISON AT MATCHED KL DISTANCE

Positive reward rescaling changes the policy reached at a common optimization strength, but not at a common KL distance. For bounded finite-support rewards with $\mathrm{RV}_0(r) > 0$, $\rho_r(t)$ is strictly increasing for $t > 0$. Let $t_r(\rho)$ denote its inverse on the attainable KL interval and define $G_\rho(r) = G_{t_r(\rho)}(r)$. As $\rho \downarrow 0$,

$$G_\rho(r) = \sqrt{2\rho}\, M_{\mathrm{KL}}(r) + O(\rho), \qquad M_{\mathrm{KL}}(r) = \frac{M_0(r)}{\sqrt{\mathrm{RV}_0(r)}} \tag{25}$$

(Theorem A.4, Appendix A.5). Thus $\Delta M_{\mathrm{KL}} \neq 0$ determines the sign of the gain difference for sufficiently small ρ, but not at a specified finite KL distance.

For a single prompt, $M_{\mathrm{KL}} = \sigma_G \,\mathrm{Corr}_{\pi_0}(r_G, r)$ when both variances are nonzero. Across prompts, the same identity uses jointly distributed, prompt-centered scores (Appendix A.5). Hence M_{KL} measures target alignment independently of global positive rescaling and prompt-wise additive shifts.

Finite matched-KL certificates. At a fixed KL distance ρ, different rewards consistent with the same static observations generally reach ρ at different optimization strengths. A finite-ρ certificate must therefore control the matched-KL gain uniformly over each candidate's entire feasible set.

For candidate i, any feasible coefficient vector can be written as

$$w_{0,i} + Nz, \qquad Z_i = \{z : \|z\|_2 \le R_i\}, \tag{26}$$

where the columns of N form an orthonormal basis for $\ker A$. Thus z parameterizes exactly the reward directions that remain unconstrained by the observed scores, and

$$r_{i,z} = \Phi(w_{0,i} + Nz). \tag{27}$$

We approximate Z_i by a finite set $\widehat{Z}_i$ such that every $z \in Z_i$ lies within distance δ_i of some cover point. Appendix A.6 derives uniform bounds for transferring both KL distance and target-utility gain from these cover points to the full feasible set. We denote the resulting Lipschitz bounds by $K_i(T)$ and $H(T)$.

Theorem 4.3 (Uniform comparison at finite matched KL)**.** *Fix $\rho > 0$. For each candidate i, suppose a finite $T_i > 0$ satisfies*

$$\min_{z_j \in \widehat{Z}_i} \rho_{r_{i,z_j}}(T_i) - K_i(T_i)\delta_i > \rho. \tag{28}$$

Then every $z \in Z_i$ has a unique $t_{i,z}(\rho) \in (0, T_i)$ and hence a well-defined gain $G_\rho(r_{i,z})$. Suppose the computed gains at the cover points satisfy

$$\left|\widehat{G}_{i,j} - G_\rho(r_{i,z_j})\right| \le \varepsilon_{i,j}. \tag{29}$$

Then, uniformly over Z_i,

$$G_\rho(r_{i,z}) \in [\ell_i(\rho), u_i(\rho)]. \tag{30}$$

where

$$\ell_i(\rho) = \min_j \left(\widehat{G}_{i,j} - \varepsilon_{i,j}\right) - H(T_i)\delta_i, \qquad u_i(\rho) = \max_j \left(\widehat{G}_{i,j} + \varepsilon_{i,j}\right) + H(T_i)\delta_i. \tag{31}$$

Consequently, for every feasible reward pair,

$$G_\rho(r_{2,z_2}) - G_\rho(r_{1,z_1}) \in [\ell_2 - u_1,\ u_2 - \ell_1]. \tag{32}$$

If this interval excludes zero, the pairwise comparison is identified without point identifying either reward.

Table 3: Matched-KL comparison outcomes at the same eight-score budget and mean KL 0.2. Percentages cover ten RM pairs under ten query selections per bank. Certified means that $\Delta G_{0.2}$ has a fixed sign over all feasible reward pairs; both signs attained means that feasible reward pairs attain both positive and negative $\Delta G_{0.2}$; unresolved means neither conclusion is established.

	Certified (%)		Both signs attained (%)		Unresolved (%)	
Response bank	Directed	Random	Directed	Random	Directed	Random
Qwen	33	10	62	77	5	13
H4 SFT	34	6	59	93	7	1
OpenOrca	38	15	60	71	2	14

The result turns finitely many cover-point evaluations into a guarantee over the entire feasible set, while allowing each reward to use its own optimization strength to reach ρ. Appendix B.8 gives a concrete Qwen example in which the uniform gain intervals are disjoint at $\rho = 0.2$, determining the pairwise comparison despite uncertainty in both rewards.

At the same eight-score budget and mean KL 0.2, target-directed observations certify 33–38% of comparisons across the three response banks, compared with 6–15% for random observations (Table 3); Appendix B.8 reports sensitivity to the norm bound.

One Qwen pair has $\Delta M_{\mathrm{KL}} \in [0.000263, 0.004816]$, yet feasible rewards yield gain differences $+0.003315$ and -0.00003193 at $\rho = 0.2$; the other 61 directed cases with opposite finite signs already admit both small-KL signs (Appendix Table 7).

4.5 Horizon dependence in parameterized policy training

We also examine horizon dependence under full-parameter RLOO (Ahmadian et al., 2024), training two policies from the same SFT initialization with an RM pair of similar static accuracy (0.6986/0.7044); details are in Appendix D.4. This complements the exact-Gibbs analysis by asking whether the horizon dependence in RQ3 also appears in parameterized policy training, where trajectories are compared at matched update counts rather than matched KL distances.

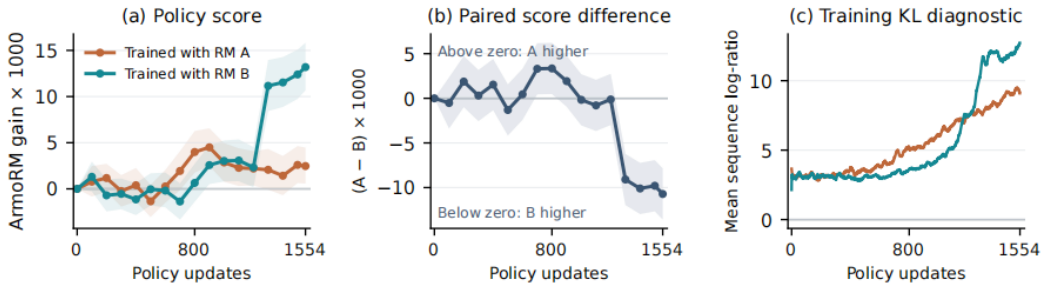


Figure 3: Relative target utility changes sign over one paired RLOO run. (a) ArmoRM gain from initialization. (b) A-minus-B gain, with pointwise 95% bands in (a) and simultaneous bands in (b). (c) Mean sequence log-ratio to the SFT reference. Policies are aligned by update count; see Appendix D.4.

5 Conclusion

Static RM evaluation need not determine downstream RM selection: complete ordering and reward variance can leave the comparison unresolved, while partial score observations can suffice when they constrain the target-relevant directions, with additional conditions needed after finite optimization. These results suggest that RM evaluation should be designed around its downstream use, with the target and optimization regime determining which reward information is relevant for selection. A practical next step is to use these conditions to decide when static evaluation is sufficient for RM selection and when downstream policy evaluation is necessary, while extending the analysis from a fixed target reward to the set of targets consistent with available preference evidence.

AI USE STATEMENT

Generative AI tools assisted with literature review, translation, and language editing as well as code debugging. The authors take responsibility for the final content of this work.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

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*, pp. 9722–9744, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975. doi: 10.1002/cpa.3160280102.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30039–30069, 2023. doi: 10.52202/075280-1308.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, and Joar Max Viktor Skalse. The perils of optimizing learned reward functions: Low training error does not guarantee low regret. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17306–17377, 2025. URL https://proceedings.mlr.press/v267/fluri25a.html.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. González, and Ion Stoica. How to evaluate reward models for RLHF. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2e01083b381b4865919b4915ef32e3d2-Abstract-Conference.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025. doi: 10.18653/v1/2025.findings-naacl.96.

Max Lamparth, Daniel Fein, Andreas Haupt, Marcel Hussing, and Mykel J. Kochenderfer. Reward bias substitution: Single-axis bias mitigations redirect optimization pressure. *arXiv preprint arXiv:2605.27996*, 2026. URL https://arxiv.org/abs/2605.27996.

Filippo Lazzati and Alberto Maria Metelli. Robustness in the face of partial identifiability in reward learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/af179b3066cae40d9e57241a75b9db2c-Abstract-Conference.html.

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

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ea4fe0a56d02c93401902b5b4c6b12da-Abstract-Conference.html.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004. doi: 10.1007/978-1-4419-8853-9.

Qwen Team. Qwen2.5: A party of foundation models! https://qwenlm.github.io/blog/qwen2.5/, 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. doi: 10.52202/075280-2338.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What makes a reward model a good teacher? an optimization perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1981.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058, 2023. URL https://proceedings.mlr.press/v202/skalse23a.html.

Charles Soussen, Rémi Gribonval, Jérôme Idier, and Cédric Herzet. Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares. *IEEE Transactions on Information Theory*, 59(5): 3158–3174, 2013. doi: 10.1109/TIT.2013.2238606. URL https://arxiv.org/abs/1111.0522.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking reward modeling in preference-based large language model alignment. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7423902b5534e2b267438c85444a54b1-Abstract-Conference.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://aclanthology.org/2024.findings-emnlp.620/.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/93f250215e4889119807b6fac3a57aec-Abstract-Conference.html.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067, 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

A Proofs and constructions

A.1 Target-utility derivative and an ordinal counterexample

For any bounded f, differentiating the exponential-family form of π_t^r (Wainwright & Jordan, 2008) gives

$$\frac{d}{dt}\pi_t^r(y \mid x) = \pi_t^r(y \mid x)\left(r(x,y) - \mathbb{E}_{\pi_t^r}[r \mid x]\right),$$

and hence

$$\frac{d}{dt}\mathbb{E}_{\pi_t^r}[f] = \mathrm{Cov}_{\pi_t^r}(f, r).$$

Setting $f = r_G$ and averaging over prompts proves Theorem 3.1.

A reward perturbation can leave the observed evaluation unchanged while changing this covariance. If h satisfies

$$\mathcal{E}_{D_{\mathrm{eval}}}(r + \delta h) = \mathcal{E}_{D_{\mathrm{eval}}}(r) \tag{33}$$

but

$$\mathbb{E}_x \,\mathrm{Cov}_{\pi_0}(r_G, h) \neq 0, \tag{34}$$

then

$$M_0(r + \delta h) = M_0(r) + \delta \mathbb{E}_x \,\mathrm{Cov}_{\pi_0}(r_G, h). \tag{35}$$

Thus, identification of M_0 requires every feasible perturbation that leaves the evaluation unchanged to leave this covariance unchanged.

For the ordinal counterexample, consider one prompt and three equiprobable responses. Let

$$r_G = (2, -5, 3), \qquad r^{(1)} = (0, 1, 3), \qquad r^{(2)} = (0, 2, 3). \tag{36}$$

The candidate rewards have the same strict ordering. Their centered values are

$$\left(-\frac{4}{3}, -\frac{1}{3}, \frac{5}{3}\right) \quad \text{and} \quad \left(-\frac{5}{3}, \frac{1}{3}, \frac{4}{3}\right), \tag{37}$$

so both variances equal $14/9$. Since $\mathbb{E}[r_G] = 0$,

$$\mathrm{Cov}(r_G, r^{(1)}) = \frac{1}{3}(0 - 5 + 9) = \frac{4}{3} > 0, \tag{38}$$

whereas

$$\mathrm{Cov}(r_G, r^{(2)}) = \frac{1}{3}(0 - 10 + 9) = -\frac{1}{3} < 0. \tag{39}$$

Thus, identical complete response ordering and reward variance do not identify the sign of M_0.

A.2 Identification from linear score observations and partial identification

If $Aw = Aw'$, then $v = w' - w \in \ker A$. Therefore $b^\top w$ is determined by Aw if and only if

$$b^\top v = 0 \quad \text{for every } v \in \ker A, \tag{40}$$

which is equivalent to $b \in (\ker A)^\perp = \mathrm{row}(A)$.

For $v \in \ker A$,

$$b^\top v = (P_{\ker A} b)^\top v, \tag{41}$$

and Cauchy–Schwarz yields

$$|b^\top v| \le \|P_{\ker A} b\|_2 \|v\|_2 \le R\Gamma_{\mathrm{id}}. \tag{42}$$

If $P_{\ker A} b \neq 0$, the choices

$$v = \pm R \frac{P_{\ker A} b}{\|P_{\ker A} b\|_2} \tag{43}$$

attain the endpoints; continuous scaling attains the entire interval. If the projection vanishes, the interval degenerates to one point. This is the support-function calculation for a Euclidean ball (Boyd & Vandenberghe, 2004). It proves Theorem 3.3; the preceding null-space argument proves Theorem 3.2.

For two candidates,

$$\Delta M_0 = b^\top (w_2 - w_1) + b^\top (v_2 - v_1). \tag{44}$$

Because v_1 and v_2 vary independently over their respective norm-bounded feasible sets,

$$|b^\top (v_2 - v_1)| \le (R_1 + R_2)\Gamma_{\mathrm{id}}, \tag{45}$$

with equality when v_1 and v_2 align with $P_{\ker A} b$ in opposite directions.

Identification across multiple target rewards. Let $\mathcal{G}$ be a set of target rewards. For $g \in \mathcal{G}$, define

$$b(g) = \mathbb{E}_x \mathbb{E}_{\pi_0} \left[(g - \mathbb{E}_{\pi_0}[g \mid x])\phi \right], \qquad M_0^g(w) = b(g)^\top w. \tag{46}$$

The same argument gives identification for every $g \in \mathcal{G}$ if and only if

$$\mathrm{span}\{b(g) : g \in \mathcal{G}\} \subseteq \mathrm{row}(A). \tag{47}$$

If a bounded reward perturbation h satisfies $\mathbb{E}_x \mathrm{Cov}_{\pi_0}(g, h) = 0$ for every bounded g, taking $g = h$ gives

$$\mathbb{E}_x \mathrm{Var}_{\pi_0}(h \mid x) = 0. \tag{48}$$

Hence $h(x, y) = c(x)$, $\pi_0(\cdot \mid x)$-almost surely for almost every x. The converse follows directly. Such prompt-wise additive constants leave the entire Gibbs policy path unchanged because

$$\pi_t^{r+c}(y \mid x) = \frac{\pi_0(y \mid x) e^{tr(x,y)} e^{tc(x)}}{e^{tc(x)} Z_t^r(x)} = \pi_t^r(y \mid x). \tag{49}$$

A.3 Stability under a common distribution shift and a fixed-support counterexample

Write $h = r_2 - r_1$. For any response distribution R, $\Delta M_R = \mathrm{Cov}_R(r_G, h)$, with covariance taken within each prompt and then averaged. For bounded f, g,

$$|\mathrm{Cov}_P(f, g) - \mathrm{Cov}_Q(f, g)| \le 6\|f\|_\infty \|g\|_\infty \mathrm{TV}(P, Q). \tag{50}$$

Indeed, the change in $\mathbb{E}[fg]$ is at most $2\|f\|_\infty \|g\|_\infty \mathrm{TV}(P, Q)$, and the change in the product of means is at most four times the same factor. Applying this per prompt and averaging proves Equation 124. Pinsker's inequality (Tsybakov, 2009) and Jensen's inequality give

$$d_{\mathrm{TV}}(P, Q) \le \mathbb{E}_x \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P(\cdot \mid x) \| Q(\cdot \mid x))} \tag{51}$$

$$\le \sqrt{\tfrac{1}{2} \mathbb{E}_x D_{\mathrm{KL}}(P(\cdot \mid x) \| Q(\cdot \mid x))}, \tag{52}$$

which proves Equation 126.

For a counterexample with unchanged support, consider one prompt with

$$r_G = (-1, 0, 1), \quad h = (1, -3, 1), \quad Q = (0.1, 0.2, 0.7), \quad P = (0.3, 0.6, 0.1). \tag{53}$$

Both distributions assign positive mass to all responses. Direct calculation yields

$$\mathrm{Cov}_Q(r_G, h) = 0.48 > 0, \qquad \mathrm{Cov}_P(r_G, h) = -0.48 < 0. \tag{54}$$

Taking $r_1 \equiv 0$ and $r_2 = h$ gives the claimed reversal. Since the supports agree, P can also be represented as an exponential tilt of Q.

A.4 Reversal and remainder bounds after finite optimization

We first control the error of the first-order policy approximation and then translate this error into an outer bound on the finite utility gain $G_t(r)$. We subsequently extend the same argument to the difference between two candidate rewards.

Proposition A.1 (Second-order remainder of the policy approximation)**.**

$$C_t(r) := \mathbb{E}_x C_{t,x}(r) \le \int_0^t (t-s)\,\mathrm{RV}_s(r)\,ds. \tag{55}$$

If $\operatorname{osc}_y r(x,\cdot) \le B_r$ *for every prompt, then*

$$C_{t,x}(r) \le \frac{B_r^2 t^2}{8}, \qquad C_t(r) \le \frac{B_r^2 t^2}{8}. \tag{56}$$

Hence $C_t(r) = O(t^2)$.

The second derivative of the exponential-tilt probabilities is quadratic in the centered reward, and its L_1 norm is controlled by $\mathrm{RV}_t(r)$. Taylor's integral remainder then gives the bound above, which is $O(t^2)$ under the bounded-range condition.

Suppose the static observation e implies

$$M_0(r) \in [m_-, m_+] \qquad \text{for every } r \in \mathcal{R}(e), \tag{57}$$

and let $B_t(e)$ be a uniform remainder bound satisfying

$$\mathbb{E}_x[B_G(x) C_{t,x}(r)] \le B_t(e) \qquad \text{for every } r \in \mathcal{R}(e). \tag{58}$$

Corollary A.2 (Outer bound on finite utility gain)**.** *For every* $r \in \mathcal{R}(e)$*,*

$$G_t(r) \in [tm_- - B_t(e),\ tm_+ + B_t(e)]. \tag{59}$$

If $tm_- > B_t(e)$*, then* $G_t(r) > 0$ *for all* $r \in \mathcal{R}(e)$.

This interval is an outer bound; its endpoints need not be attained by feasible rewards. If every reward in $\mathcal{R}(e)$ has within-prompt range at most B_R and $B_G(x) \le B_G$, Proposition A.1 allows

$$B_t(e) = \frac{B_G B_R^2 t^2}{8}. \tag{60}$$

The resulting interval has width $t(m_+ - m_-) + O(t^2)$ near $t = 0$.

Pairwise remainder. For two rewards optimized separately from the same reference policy at the same strength t, define

$$\Delta G_t = G_t(r_2) - G_t(r_1), \quad d_t^\Delta = (\pi_t^{r_2} - \pi_t^{r_1}) - t(\dot\pi_0^{r_2} - \dot\pi_0^{r_1}), \quad C_{t,x}^\Delta = \tfrac{1}{2}\|d_t^\Delta(x,\cdot)\|_1. \tag{61}$$

Write $C_t^\Delta = \mathbb{E}_x C_{t,x}^\Delta$.

Corollary A.3 (Pairwise remainder bound)**.**

$$|\Delta G_t - t\Delta M_0| \le \mathbb{E}_x[B_G(x) C_{t,x}^\Delta]. \tag{62}$$

The bound is worst-case sharp over target rewards constrained only by the prescribed per-prompt ranges. Moreover,

$$C_t^\Delta \le C_t(r_1) + C_t(r_2) \le \frac{B_{r_1}^2 + B_{r_2}^2}{8} t^2. \tag{63}$$

Reversal after finite optimization. Let

$$A_x = \arg\max_y r(x,y). \tag{64}$$

On finite support, π_t^r converges as $t \to \infty$ to the reference distribution restricted and renormalized on A_x. Under uniform boundedness, dominated convergence gives

$$U(\pi_t^r) \to U(\pi_\infty^r). \tag{65}$$

If $M_0(r) > 0$ but $U(\pi_\infty^r) < U(\pi_0)$, then $G_t(r) > 0$ for all sufficiently small positive t, whereas $G_T(r) < 0$ for all sufficiently large T. Hence the target-utility gain changes sign along the Gibbs path. Moreover, the mean value theorem implies $M_s(r) < 0$ for some $s > 0$; since $M_0(r) > 0$ and $M_t(r)$ is continuous, $M_t(r)$ also crosses zero. For example, with one prompt and a uniform reference policy over three responses, let $r_G = (0, 2, -1)$ and $r = (0, 10, 11)$. Then $M_0(r) = 2/3 > 0$, while the Gibbs policy converges to the third response, giving $U(\pi_\infty^r) = -1 < U(\pi_0) = 1/3$.

Proof of Theorem 4.1. The definition of d_t^r gives

$$G_t(r) - tM_0(r) = \mathbb{E}_x \sum_y r_G(x,y) d_t^r(x,y). \tag{66}$$

Because $\sum_y d_t^r(x,y) = 0$, subtracting the midpoint

$$m_x = \frac{\max_y r_G(x,y) + \min_y r_G(x,y)}{2} \tag{67}$$

does not change the sum. Since $|r_G - m_x| \le B_G(x)/2$,

$$\left| \sum_y r_G d_t^r \right| \le \frac{B_G(x)}{2} \|d_t^r\|_1 = B_G(x) C_{t,x}(r). \tag{68}$$

For fixed d_t^r, equality is attained by assigning $r_G(x,y)$ its upper allowed value wherever $d_t^r(x,y) > 0$ and its lower allowed value wherever $d_t^r(x,y) < 0$. This proves worst-case sharpness over target rewards constrained only by the prescribed per-prompt ranges.

Combining Theorem 4.1 with $M_0(r) \in [m_-, m_+]$ and the uniform remainder bound $B_t(e)$ gives Corollary A.2.

Proof of Proposition A.1. Taylor's integral remainder gives

$$\pi_t^r - \pi_0 - t\dot{\pi}_0^r = \int_0^t (t-s) \ddot{\pi}_s^r \, ds. \tag{69}$$

Within a prompt,

$$\ddot{\pi}_s^r(y) = \pi_s^r(y) \left((r(y) - \bar{r}_s)^2 - \mathrm{Var}_{\pi_s^r}(r) \right), \tag{70}$$

so $\|\ddot{\pi}_s^r\|_1/2 \le \mathrm{Var}_{\pi_s^r}(r)$. Integration and averaging yield the first inequality in Equation 55. If $\operatorname{osc} r \le B_r$, Popoviciu's inequality gives $\mathrm{Var}(r) \le B_r^2/4$, hence

$$C_{t,x}(r) \le \int_0^t (t-s) \frac{B_r^2}{4} \, ds = \frac{B_r^2 t^2}{8}. \tag{71}$$

Proof of Corollary A.3. For each prompt, d_t^Δ has zero total mass. The same midpoint argument as in the proof of Theorem 4.1 therefore gives

$$|\Delta G_t - t\Delta M_0| \le \mathbb{E}_x \left[B_G(x) C_{t,x}^\Delta \right]. \tag{72}$$

This bound is worst-case sharp over target rewards constrained only by the prescribed per-prompt ranges. The triangle inequality gives

$$C_t^\Delta \le C_t(r_1) + C_t(r_2) \le \frac{B_{r_1}^2 + B_{r_2}^2}{8} t^2. \tag{73}$$

Combining a uniform feasible-set bound on this remainder with $\Delta M_0 \in [\delta_-, \delta_+]$ proves Equation 23.

A.5 Matched-KL expansion

At matched KL, the leading local effect is $M_0(r)$ normalized by the initial reward standard deviation. Along the exponential tilt,

$$\rho_r'(t) = t \, \mathrm{RV}_t(r), \qquad \rho_r(t) = \tfrac{1}{2} \, \mathrm{RV}_0(r) t^2 + O(t^3). \tag{74}$$

For bounded rewards on finite support with $\mathrm{RV}_0(r) > 0$, ρ_r is strictly increasing for $t > 0$. Let $t_r(\rho)$ be its inverse on the attainable KL interval and define $G_\rho(r) = G_{t_r(\rho)}(r)$.

Theorem A.4 (Small-KL expansion of downstream utility gain)**.** *As $\rho \downarrow 0$,*

$$G_\rho(r) = \sqrt{2\rho} \, M_{\mathrm{KL}}(r) + O(\rho), \qquad M_{\mathrm{KL}}(r) = \frac{M_0(r)}{\sqrt{\mathrm{RV}_0(r)}}. \tag{75}$$

For two rewards, $\Delta G_\rho = \sqrt{2\rho} \, \Delta M_{\mathrm{KL}} + O(\rho)$. If $\Delta M_{\mathrm{KL}} \neq 0$, its sign equals that of ΔG_ρ for all sufficiently small positive ρ.

Proof. Write $\mathrm{RV}_t = \mathbb{E}_x \operatorname{Var}_{\pi_t^r}(r \mid x)$. Differentiating the log-partition function (Wainwright & Jordan, 2008) gives

$$\rho_r'(t) = t\,\mathrm{RV}_t, \qquad \frac{d}{dt}\,\mathrm{RV}_t = \mathbb{E}_x \mathbb{E}_{\pi_t^r}[(r - \bar{r}_t)^3]. \tag{76}$$

Thus $\mathrm{RV}_t = \mathrm{RV}_0 + O(t)$ under the bounded finite-support assumptions, and

$$\rho_r(t) = \frac{1}{2}\,\mathrm{RV}_0\, t^2 + O(t^3). \tag{77}$$

If $\mathrm{RV}_0 > 0$, ρ_r is strictly increasing on a sufficiently small positive interval, with inverse

$$t_r(\rho) = \sqrt{\frac{2\rho}{\mathrm{RV}_0}} + O(\rho). \tag{78}$$

Theorem 4.1 gives $G_t = M_0 t + O(t^2)$. Substitution proves Equation 75; subtracting two expansions gives the pairwise result. If the leading pairwise coefficient is nonzero, it determines the sign for all sufficiently small positive ρ.

If $\operatorname{osc} r \le B_r$, then

$$\left|\frac{d}{dt}\,\mathrm{RV}_t\right| \le B_r\,\mathrm{RV}_t \le \frac{B_r^3}{4}, \qquad \left|\rho_r(t) - \frac{1}{2}\,\mathrm{RV}_0\, t^2\right| \le \frac{B_r^3 t^3}{12}. \tag{79}$$

This is an explicit control of the $O(t^3)$ remainder; the main text uses only its $O(\rho)$ consequence after locally inverting $\rho_r(t)$ near $t = 0$.

Attainable KL budgets and reward rescaling. For every finite t, the Gibbs policy has the same support as π_0. Thus $\mathrm{RV}_0(r) > 0$ implies $\mathrm{RV}_t(r) > 0$ and $\rho_r'(t) > 0$ for $t > 0$. With $A_x = \arg\max_y r(x, y)$, the limiting policy is π_0 conditioned on A_x. Pointwise convergence and monotone convergence of the prompt-wise KL give

$$\rho_\infty(r) = \lim_{t\to\infty} \rho_r(t) = \mathbb{E}_x[-\log \pi_0(A_x \mid x)] \in (0, \infty]. \tag{80}$$

A finite value of $t_r(\rho)$ exists for $0 < \rho < \rho_\infty(r)$, but not for arbitrary KL budgets. On a uniform K-draw bank, the upper limit is $\mathbb{E}_x \log(K/|A_x|)$. For global $a > 0$, $\pi_t^{ar+c(x)} = \pi_{at}^r$, so $t_{ar+c}(\rho) = t_r(\rho)/a$ and the policy reached at every attainable KL distance is unchanged. A prompt-dependent scaling generally changes the resulting policy when rewards are compared at the same mean KL.

Correlation interpretation. Set $\widetilde{r} = r - \mathbb{E}_{\pi_0}[r \mid x]$ and $\widetilde{r}_G = r_G - \mathbb{E}_{\pi_0}[r_G \mid x]$, and take joint expectations under $x \sim D_X$, $y \sim \pi_0(\cdot \mid x)$. Both centered variables have zero mean, with variances $\mathrm{RV}_0(r)$ and $\mathrm{RV}_0(r_G)$. Whenever both are positive,

$$M_{\mathrm{KL}}(r) = \sqrt{\mathrm{RV}_0(r_G)}\,\operatorname{Corr}_{D_X \pi_0}(\widetilde{r}_G, \widetilde{r}). \tag{81}$$

This follows from $M_0(r) = \mathbb{E}[\widetilde{r}_G \widetilde{r}]$. For a single prompt, the correlation in Equation 81 is the ordinary Pearson correlation across responses. With multiple prompts, the joint correlation generally differs from the average of prompt-wise correlations. The latter can also be undefined at individual constant-score prompts even when both aggregate variances are positive. If $\mathrm{RV}_0(r_G) = 0$, all target gains on the reference support vanish. This interpretation concerns the leading small-KL coefficient; it neither determines finite gains from the correlation alone nor changes the local quantity M_0 defined at the reference policy.

A.6 Uniform comparison at finite matched KL

We prove Theorem 4.3 from Section 4. All KL divergences use the same reference policy and fixed prompt weights. Let $\widehat{Z}_i \subset Z_i$ be a finite δ_i-net of the closed feasible ball Z_i. Different feasible rewards generally require different optimization strengths $t_{i,z}(\rho)$ to reach the same KL distance ρ. Therefore, a bound evaluated at one common strength t does not by itself control all feasible rewards at the requested KL distance.

To transfer bounds from the net points to the full feasible set, define

$$
\begin{aligned}
D_x &\coloneqq \max_{y,y'} \left\| N^\top \big(\phi(x,y) - \phi(x,y')\big) \right\|_2 , \\
\widehat{B}_{i,x} &\coloneqq \max_{y,y'} \Big\{ \left| \big(\phi(x,y) - \phi(x,y')\big)^\top w_{0,i} \right| + R_i \left\| N^\top \big(\phi(x,y) - \phi(x,y')\big) \right\|_2 \Big\}, \\
K_i(T) &\coloneqq \frac{T^2}{4} \sqrt{\mathbb{E}_x\, \widehat{B}_{i,x}^2\; \mathbb{E}_x\, D_x^2}, \\
H(T) &\coloneqq \frac{T}{4} \sqrt{\mathbb{E}_x\, B_G(x)^2\; \mathbb{E}_x\, D_x^2}.
\end{aligned} \tag{82}
$$

The quantity $\widehat{B}_{i,x}$ uniformly bounds the within-prompt reward range over all $z \in Z_i$.

Proof. For a fixed i, z, t, define the prompt-averaged covariance $\mathcal{C}_t(f,h) = \mathbb{E}_x \operatorname{Cov}_{\pi_t^{r_{i,z}}}(f,h)$ and set $\psi_h = \Phi N h$ for a unit direction h. Differentiating the Gibbs log-partition function with respect to z at fixed t gives

$$
\partial_h \rho_{r_{i,z}}(t) = t^2 \mathcal{C}_t(r_{i,z}, \psi_h). \tag{83}
$$

Using the covariance bound

$$
|\operatorname{Cov}(X,Y)| \le \frac{1}{4} \operatorname{osc}(X) \operatorname{osc}(Y), \tag{84}
$$

together with Cauchy–Schwarz across prompts, $|\partial_h \rho_{r_{i,z}}(T_i)| \le K_i(T_i)$, since $\operatorname{osc}_y \psi_h \le D_x$. The δ_i-net bound therefore extends Equation 28 from the net points to every $z \in Z_i$. In particular, no feasible reward is constant on every prompt's reference support. The identity $\rho_r'(t) = t \operatorname{RV}_t(r) > 0$ for $t > 0$, continuity, and $\rho_r(0) = 0$ imply existence and uniqueness of $t_{i,z}(\rho) \in (0, T_i)$.

Set $t = t_{i,z}(\rho)$ and $V_t = \mathcal{C}_t(r_{i,z}, r_{i,z}) > 0$. Implicit differentiation followed by the chain rule yields

$$
\begin{aligned}
\partial_h t_{i,z}(\rho) &= -t\, \frac{\mathcal{C}_t(r_{i,z}, \psi_h)}{V_t}, \\
\partial_h G_\rho(r_{i,z}) &= t \left\{ \mathcal{C}_t(r_G, \psi_h) - \frac{\mathcal{C}_t(r_G, r_{i,z}) \mathcal{C}_t(r_{i,z}, \psi_h)}{V_t} \right\}.
\end{aligned} \tag{85}
$$

Although $\partial_h t_{i,z}(\rho)$ contains $1/V_t$, the derivative of the gain can be bounded without this factor. The covariance form $\mathcal{C}_t(\cdot,\cdot)$ is positive semidefinite and induces an inner product after quotienting out prompt-wise constants. The bracket in Equation 85 is the covariance inner product obtained after projecting ψ_h onto the orthogonal complement of $r_{i,z}$. Orthogonal projection cannot increase the induced norm, so

$$
\sqrt{\mathcal{C}_t(r_G, r_G) \mathcal{C}_t(\psi_h, \psi_h)} \le \frac{1}{4} \sqrt{\mathbb{E}_x B_G(x)^2\; \mathbb{E}_x D_x^2}. \tag{86}
$$

Hence $\|\nabla_z G_\rho(r_{i,z})\|_2 \le H(T_i)$ for all $z \in Z_i$. Integrating along the segment from z to a nearest net point and adding the evaluation error gives Equation 31. Subtracting the two candidate intervals gives the pairwise bound.

A.7 Score-function projection and one-step parameter updates

This section proves Propositions D.6 and D.7 for differentiable parameterized policies. Assume differentiability and fixed support in a neighborhood of the parameter value of interest. For any function $f(x,y)$ that does not explicitly depend on θ, the score-function identity (Williams, 1992) gives

$$
\nabla_\theta \mathbb{E}_{x,y\sim\pi_\theta}[f] = \mathbb{E}_x \operatorname{Cov}_{\pi_\theta}(f, s_\theta), \qquad s_\theta = \nabla_\theta \log \pi_\theta. \tag{87}
$$

For one prompt, let

$$
\tilde{f} = f - \mathbb{E}[f], \qquad g = \mathbb{E}[s\tilde{f}], \qquad F = \mathbb{E}[ss^\top]. \tag{88}
$$

If $a \in \ker F$, then

$$
0 = a^\top F a = \mathbb{E}[(a^\top s)^2], \tag{89}
$$

so $a^\top s = 0$ almost surely and hence $a^\top g = 0$. Therefore

$$
g \in (\ker F)^\perp = \operatorname{range}(F). \tag{90}
$$

Table 4: Candidate RMs and target evaluator.

Hugging Face model ID	Role
`Skywork/Skywork-Reward-V2-Qwen3-0.6B`	Candidate
`Skywork/Skywork-Reward-V2-Qwen3-1.7B`	Candidate
`Skywork/Skywork-Reward-V2-Qwen3-4B`	Candidate
`Skywork/Skywork-Reward-V2-Llama-3.2-1B`	Candidate
`internlm/internlm2-1_8b-reward`	Candidate
`RLHFlow/ArmoRM-Llama3-8B-v0.1`	Target

With $\alpha = F^{+}g$, we have $F\alpha = g$ and

$$\mathbb{E}[s(\tilde{f} - \alpha^{\top} s)] = 0. \tag{91}$$

Thus $\alpha^{\top} s$ is the $L_2(\pi_\theta)$ orthogonal projection of $\tilde{f}$ onto the linear span of the score coordinates. The Pythagorean identity gives

$$\mathrm{Var}_{\pi_\theta}(f) = g^{\top} F^{+} g + R_{\perp}^2(f), \tag{92}$$

where $R_{\perp}^2(f)$ is the squared $L_2(\pi_\theta)$ norm of the residual orthogonal to this span. The residual does not contribute to $g = \mathbb{E}[s\tilde{f}]$. Hence total reward variance alone does not determine the first-order parameter gradient.

Let

$$J_r(\theta) = \mathbb{E}_{\pi_\theta}[r]. \tag{93}$$

Suppose at θ_0,

$$\langle \nabla U(\theta_0), \nabla J_r(\theta_0) \rangle \le -\delta < 0, \qquad \|\nabla J_r(\theta_0)\| \le \Gamma, \tag{94}$$

and ∇U is L_U-Lipschitz. Then for

$$0 < \eta \le \frac{\delta}{L_U \Gamma^2}, \tag{95}$$

the update $\theta_0 + \eta \nabla J_r(\theta_0)$ decreases target utility by at least $\eta\delta/2$. If J_r is also L_J-smooth and $\eta \le 1/L_J$, the same step increases J_r whenever $\nabla J_r(\theta_0) \neq 0$. These bounds follow from the standard descent/ascent lemmas for functions with Lipschitz gradients (Nesterov, 2004). The result concerns a single local gradient step and does not extend the exact finite-optimization certificates to an entire training trajectory.

B Experimental protocol and additional results

B.1 Models, prompts, and response sampling

We use five pretrained scalar RMs as candidate rewards (Table 4), four from Skywork-Reward-V2 (Liu et al., 2026). ArmoRM-Llama3-8B-v0.1 provides the target scores. Each RM is evaluated with its released tokenizer and chat template, and responses are scored without truncation.

The primary response bank uses 128 prompts from `HuggingFaceH4/ultrafeedback_binarized` (`test_prefs`), selected by a 512-token filter using the Pythia tokenizer (Biderman et al., 2023) and fixed before RM scoring; only the prompt text is used in our evaluation. `Qwen/Qwen2.5-1.5B-Instruct` (Qwen Team, 2024) generates 32 responses per prompt with temperature 1, top-p 1, top-k 0, and at most 320 new tokens, yielding 4,096 responses. Exact Gibbs optimization treats the sampled response indices uniformly, including duplicate texts, so the experimental reference distribution is the empirical response bank rather than the generator's full sequence distribution.

We independently repeat the Qwen sampling once with the same prompts and settings to obtain a second 4,096-response bank. For cross-generator checks, `HuggingFaceH4/mistral-7b-sft-beta` and `Open-Orca/Mistral-7B-OpenOrca`, both based on `mistralai/Mistral-7B-v0.1` (Jiang et al., 2023), each generate 16 responses per prompt with their native chat templates, temperature 1, top-p 1, top-k 0, and a 512-token generation limit, yielding 2,048 responses per bank. The candidate and target RMs are unchanged across banks.

All candidate, target, and response-generation checkpoints are fixed in Appendix B; parameterized policy updates are studied separately in Appendix D. ArmoRM serves as the fixed target evaluator throughout these experiments.

B.2 Static quantities, exact optimization, and uncertainty

Accuracy is computed over all unordered response pairs within each prompt. Target ties are excluded, and candidate ties receive half credit; this differs slightly from the strict convention in Equation 3. Accuracy therefore measures agreement with the target ordering, not response correctness.

We report both the raw within-prompt variance

$$\mathrm{RV}_0(r) = \frac{1}{NK}\sum_{x,k}(r_{xk} - \bar{r}_x)^2 \tag{96}$$

and the normalized variance used by Razin et al. (2025),

$$\overline{\mathrm{RV}}(r) = \frac{N^{-1}\sum_x (K-1)^{-1}\sum_k (r_{xk} - \bar{r}_x)^2}{s_r^2}, \tag{97}$$

where s_r^2 is the global sample variance. The former is the quantity used in M_{KL}; the latter is reported only for comparison with prior RM-variance analyses.

For matched-KL results, t is solved separately for each reward from $\rho_r(t) = \rho$. We use $\rho \in \{10^{-4}, 10^{-3}, 0.01, 0.03, 0.1, 0.2, 0.4\}$, all attainable for the original candidate RMs. Best-of-N samples are drawn with replacement from the empirical bank. After fixing the reward ordering and tie-break, a response of rank k is selected with probability

$$(k/K)^N - ((k-1)/K)^N, \tag{98}$$

so Best-of-N depends only on the induced ordering.

Uncertainty intervals for matched-KL gains use paired percentile bootstrap resampling over the 128 prompts (Efron, 1979). Each bootstrap replicate refits the KL multiplier on the resampled prompt distribution. These intervals condition on the sampled response bank; the controlled-reward experiment in Appendix B.3 instead keeps the full-bank multiplier fixed while resampling prompts.

B.3 Controlled rewards with identical ordering and reward variance

Let u_{xk} denote the target score of response k to prompt x, and let a_x index the response with the lowest target score, breaking ties by index. We first normalize target scores within each prompt,

$$v_{xk} = \frac{u_{xk} - \min_j u_{xj}}{\max_j u_{xj} - \min_j u_{xj}}, \tag{99}$$

setting $v_{xk} = 0$ when all target scores are equal. For $\delta \in \{1.01, 3, 10\}$, we replace the score of every sampled copy of the response text indexed by a_x with δ, obtaining $q_{\delta,xk}$. We then center within each prompt and rescale globally,

$$\widetilde{r}_{\delta,xk} = \frac{q_{\delta,xk} - \bar{q}_{\delta,x}}{\left[(NK)^{-1}\sum_{x,k}(q_{\delta,xk} - \bar{q}_{\delta,x})^2\right]^{1/2}}. \tag{100}$$

For every $\delta > 1$, the same lowest-target response is placed above all other responses under the constructed reward; changing δ changes score-gap magnitudes without changing the response ordering. Prompt centering leaves the Gibbs policy unchanged, while positive global rescaling only reparameterizes the optimization strength and therefore leaves the policy at a fixed KL distance unchanged. Consequently, the three rewards have identical complete response ordering and accuracy, zero prompt means, and $\mathrm{RV}_0 = 1$. This controlled construction establishes that the ambiguity can occur even when complete ordering and reward variance are fixed.

Table 5: Identical response ordering and reward variance allow opposite utility changes. All three constructed rewards have accuracy 0.91615, $\mathrm{RV}_0 = 1$, and normalized variance $\overline{\mathrm{RV}} = 1.03201$. The parameter δ controls the score assigned to the response with the lowest target score before centering and rescaling. M_0 is the local target-utility effect; gain is the change in mean ArmoRM score at matched KL 0.2. Intervals are 95% prompt-bootstrap intervals with the full-bank KL multiplier fixed.

δ	M_0	ArmoRM gain at KL 0.2	Gain: 95% bootstrap interval
1.01	0.014148	0.007154	$[0.006273,\ 0.008128]$
3.00	-0.000509	-0.005139	$[-0.005766,\ -0.004571]$
10.00	-0.006919	-0.007418	$[-0.008157,\ -0.006734]$

B.4 Empirical reward class and feasible sets

On the primary Qwen bank, let $S \in \mathbb{R}^{5\times 4096}$ collect the prompt-centered scores of the five candidate RMs. Since S has rank five, we use its five-dimensional row span as the empirical linear reward class. If V contains the corresponding right singular vectors, define

$$\Phi = \sqrt{4096}\, V, \qquad w_i = \Phi^\top S_i^\top / 4096. \tag{101}$$

Then $\Phi^\top \Phi/4096 = I$ and Φw_i exactly reconstructs candidate i's centered score vector. We use the norm-bounded class

$$\mathcal{R}_\Phi = \{r_w = \Phi w : \|w\|_2 \le B\}, \qquad B = \max_i \|w_i\|_2 = 3.573482, \tag{102}$$

the smallest origin-centered Euclidean ball in these coordinates that contains all five candidate RMs.

Because prompt-wise additive constants do not affect the Gibbs policy, the primary score observations retain only within-prompt differences. Let A contain the observed difference rows and $e_i = Aw_i$. Under the norm bound above, the feasible coefficient set for candidate i is

$$w_{0,i} + \{v \in \ker A : \|v\|_2 \le R_i\}, \qquad w_{0,i} = A^+ e_i, \qquad R_i = \sqrt{B^2 - \|w_{0,i}\|_2^2}. \tag{103}$$

The radius follows because the minimum-norm solution is orthogonal to $\ker A$. The target direction is $b = \Phi^\top(u - \bar{u}_x)/4096$.

We report the marginal query budget conditional on the fixed empirical class Φ, norm bound B, and target direction b. Scores used to construct these quantities are treated as setup information rather than counted as retained static observations.

B.5 Target-directed query selection

Given the empirical feature class and target direction b, we select score differences to reduce

$$\Gamma_{\mathrm{id}} = \|P_{\ker A} b\|_2. \tag{104}$$

At each step, we sample 256 distinct within-prompt response pairs from prompts not queried previously. For a proposed difference row a, let

$$a_\perp = P_{\ker A} a, \qquad b_\perp = P_{\ker A} b. \tag{105}$$

Among proposals with $a_\perp \neq 0$, we select the pair maximizing

$$\frac{(a_\perp^\top b_\perp)^2}{\|a_\perp\|_2^2}, \tag{106}$$

which is exactly the reduction in Γ_{id}^2 after adding that observation. This is the row-selection criterion underlying orthogonal least squares (Soussen et al., 2013), restricted here to within-prompt score differences. Each selected pair uses two scores per RM and contributes one difference observation. Four sequential selections therefore give nested budgets of 2, 4, 6, and 8 scores; we repeat the randomized proposal procedure ten times. Selection uses only Φ and b; it does not use candidate-specific M_0, finite gains, or certification outcomes.

As in Appendix B.4, the query budget excludes the scores used to construct Φ, B, and b. After its initial development, the target-directed selection rule was held fixed across all reported response banks. The selected observation matrices remain non-full-rank and leave a nonzero component of b in $\ker A$, so a pairwise local comparison can be identified without point identifying either candidate's M_0.

For a fixed response bank, Φ, B, and b are shared across candidates and query selections. A new candidate must lie in the specified class and norm ball; changing the target requires recomputing b and the target-dependent bounds, while changing the response bank requires reconstructing the empirical geometry.

Finite common-t results using the refined uniform remainder bound are reported in Appendix B.6. Matched-KL results for the same observation design are reported in the main text and Appendix B.8.

B.6 Computing uniform remainder bounds

We compute a uniform bound on the remainder

$$F_{i,t}(z) = G_t(r_{i,z}) - tM_0(r_{i,z}) \tag{107}$$

over each candidate's feasible set.

For candidate i, the maximum within-prompt reward range over its feasible set is

$$\widehat{B}_{i,x} = \max_{k,\ell} \left\{ |d_{x,k\ell}^\top w_{0,i}| + R_i \|P_{\ker A} d_{x,k\ell}\|_2 \right\}, \tag{108}$$

where $d_{x,k\ell} = \phi_{xk} - \phi_{x\ell}$. Proposition A.1 therefore gives the uniform range bound

$$C_i^{\mathrm{range}}(t) = \frac{t^2}{8}\, \mathbb{E}_x\left[B_G(x) \widehat{B}_{i,x}^2 \right]. \tag{109}$$

In the experiments we use the smaller of this bound and the grid-based bound in Proposition B.1.

Proposition B.1 (Uniform remainder bound on a one-dimensional feasible set)**.** *Suppose* $\ker A$ *has unit basis vector* n *and write* $r_z = \Phi(w_0 + zn)$*,* $|z| \le R$*,* $\psi = \Phi n$*. For fixed finite* $t \ge 0$*, define* $F_t(z) = G_t(r_z) - tM_0(r_z)$ *and*

$$L_t = t \left\{ \frac{1}{4} \mathbb{E}_x[B_G(x) \operatorname{osc}_y \psi(x,y)] + |b^\top n| \right\}. \tag{110}$$

On an equally spaced grid $z_1, \dots, z_m$ *including* $-R$ *and* R*,* $m \ge 2$*,*

$$\sup_{|z| \le R} |F_t(z)| \le \max_{1 \le j \le m} |F_t(z_j)| + \frac{L_t R}{m-1}. \tag{111}$$

Consequently, replacing each candidate's remainder by this bound remains sufficient to certify the pairwise comparison after finite optimization.

Proof. Differentiating with respect to z gives

$$\partial_z G_t(r_z) = t\, \mathbb{E}_x \operatorname{Cov}_{\pi_t^{r_z}}(r_G, \psi), \qquad \partial_z M_0(r_z) = b^\top n. \tag{112}$$

Using $|\operatorname{Cov}(X,Y)| \le \frac{1}{4} \operatorname{osc}(X) \operatorname{osc}(Y)$ gives $|\partial_z F_t(z)| \le L_t$. Every $z \in [-R, R]$ lies within $R/(m-1)$ of a grid point, so the result follows from the Lipschitz bound.

In the experiments we use the smaller of Equation 111 and $C_i^{\mathrm{range}}(t)$ for each candidate. Six-score directed designs use the two-dimensional cover in Equation 113. All results reported in Table 6 use the refined second-order Lipschitz bound in Equation 115.

Two unobserved directions. When $\ker A$ has dimension two, write $r_z = \Phi(w_0 + Nz)$ with $\|z\|_2 \le R$ for an orthonormal basis N of the null space. Define $\Psi = \Phi N$ and

$$D_x = \max_{y,y'} \|\Psi(x,y) - \Psi(x,y')\|_2, \qquad L_t^{(2)} = t\{\mathbb{E}_x[B_G(x) D_x]/4 + \|N^\top b\|_2\}.$$

Applying the same covariance bound gives $\|\nabla_z F_t(z)\|_2 \le L_t^{(2)}$. Hence any finite set covering the feasible disk within distance δ yields

$$\sup_{\|z\|_2 \le R} |F_t(z)| \le \max_{z_j} |F_t(z_j)| + L_t^{(2)}\delta. \tag{113}$$

This extension supports both outcomes: some comparisons can be certified without observing all reward directions, while other feasible reward pairs attain opposite finite signs.

Second-order Lipschitz control. The same interval and disk covers admit a correction that retains the cancellation at $t = 0$. For candidate i, write $r_z = \Phi(w_{0,i} + Nz)$, $\|z\|_2 \le R_i$, and let $\psi_h = \Phi N h$ for a unit null-space direction h. Direct differentiation gives

$$\partial_h F_{i,t}(z) = t\,\mathbb{E}_x\left[\mathrm{Cov}_{\pi_t^{r_z}}(r_G, \psi_h) - \mathrm{Cov}_{\pi_0}(r_G, \psi_h)\right]. \tag{114}$$

Along the fixed r_z tilt, the derivative in s of the first covariance is $\mathbb{E}_{\pi_s^{r_z}}[(r_G - \mathbb{E}r_G)(\psi_h - \mathbb{E}\psi_h)(r_z - \mathbb{E}r_z)]$, with all means conditional on the same prompt and policy. Its absolute value is at most $\widehat{B}_{i,x}\sqrt{\mathrm{Var}(r_G)\,\mathrm{Var}(\psi_h)} \le \widehat{B}_{i,x} B_G(x) D_x/4$. Here $D_x = \mathrm{osc}_y\,\psi(x,y)$ in one dimension and the projected feature diameter defined above in two dimensions. Integrating over $0 \le s \le t$ yields the uniform bound

$$\|\nabla_z F_{i,t}(z)\|_2 \le \widetilde{L}_{i,t} := \frac{t^2}{4}\mathbb{E}_x[B_G(x) D_x \widehat{B}_{i,x}]. \tag{115}$$

Thus either cover remains valid with $L_t^{(d)}$ replaced by $\min\{L_t^{(d)}, \widetilde{L}_{i,t}\}$. The second-order bound depends on the candidate's feasible center and radius and is valid over its entire feasible set. Only the bound between evaluated grid points changes; the observations, grid values, and feasible witnesses are fixed.

All reported common-t results in Table 6 use this refined bound.

Table 6: Common-t comparison outcomes on the primary Qwen bank. Each method uses ten query selections and ten RM pairs per selection. Certified means that ΔG_t has a fixed sign over all feasible reward pairs; both signs attained means that feasible pairs attain both signs; unresolved means neither conclusion is established. Median KL is measured across the five original candidate RMs. All results use the refined uniform remainder bound.

		8 scores, random			8 scores, directed			6 scores, directed		
	Median		Both signs			Both signs			Both signs	
t	KL	Certified	attained	Unresolved	Certified	attained	Unresolved	Certified	attained	Unresolved
0.01	0.000333	36	64	0	100	0	0	90	0	10
0.03	0.002941	30	64	6	100	0	0	70	0	30
0.1	0.030590	20	62	18	80	0	20	40	2	58
0.3	0.225272	9	61	30	39	16	45	10	21	69
1	1.193668	1	69	30	5	37	58	0	40	60

B.7 Opposite finite outcomes with identical local effects

We augment the score observations with each candidate's exact local effect $M_0(w_i)$,

$$A_+ = [A^\top,\ b/\|b\|_2]^\top, \qquad e_{+,i} = (e_i^\top,\ M_0(w_i)/\|b\|_2)^\top, \tag{116}$$

and recompute the feasible sets under the same global norm bound B. Thus the individual local effects, and hence the local pairwise comparison, are fixed by construction.

Nevertheless, finite optimization can reverse the comparison. Figure 4 shows one Skywork Qwen3-1.7B/Llama-3.2-1B pair for which four within-prompt score differences plus the exact M_0 values leave one unobserved reward direction. Two feasible reward assignments agree on all these observations but give opposite signs of the gain difference at $t = 0.3$: -0.002253 and $+0.0000685$. Both satisfy the same norm bound.

B.8 Matched-KL certification and sensitivity

We apply Theorem 4.3 and the bounds in Appendix A.6 to compute uniform gain intervals at fixed KL budgets.

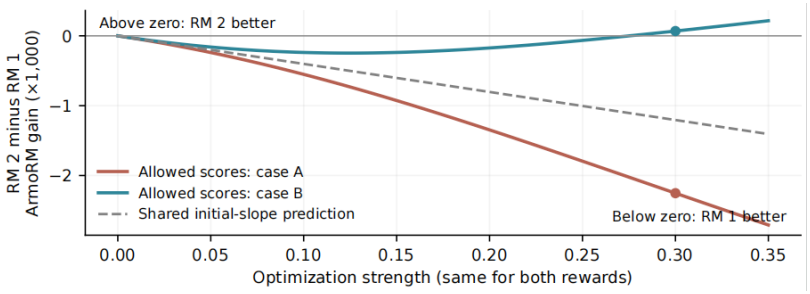


Figure 4: Identical local effects do not determine the comparison after finite optimization. Two feasible reward assignments agree on four queried score differences and both candidates' exact M_0 values, yet give opposite signs of the target-utility difference at $t = 0.3$. The dashed line shows the shared local prediction.

Computing uniform gain intervals. For a one-dimensional feasible set, we evaluate a uniform grid of $m = 1{,}025$ points, giving $\delta_i = R_i/(m-1)$. For each KL budget, we choose T_i from the fixed increasing grid $0.01, 0.03, 0.1, 0.3, 1, 3, 10$ as the first value satisfying the sufficient reachability condition in Equation 28.

At each grid point, bracketed inversion provides an interval containing the exact multiplier $t_{i,z}(\rho)$; this interval is converted into a gain-error bound using $|\partial_t G_t| \le J_i$. Numerical error bounds are included in every reported certificate. At $\rho = 0.2$, the largest pairwise numerical error bound is below 5.6×10^{-11}, whereas the smallest certified margin exceeds 2.1×10^{-6}.

Matched-KL comparison from partial score observations. The primary bank uses ten directed and ten random query selections, retaining four within-prompt differences from eight raw scores per RM. All twenty observation matrices have rank four in the same five-dimensional class, with no exact M_0 observation. Across the primary Qwen bank, the directed design certifies 34 of 100 comparisons for $\rho \in [0.001, 0.1]$ and 33 at $\rho = 0.2$; random selection certifies 10–11 over the same range. The remaining cases are predominantly ones for which both signs are attained.

A concrete certificate. At $\rho = 0.2$, one Qwen3-4B/Qwen3-1.7B comparison is certified although both candidates retain nonzero feasible radii. The gain intervals are $[0.00635, 0.00923]$ and $[0.009565, 0.009585]$, giving

$$\Delta G_{0.2} \in [0.000340, 0.003232]. \tag{117}$$

What changes under variance normalization? In the empirical orthonormal score coordinates, $\mathrm{RV}_0(w) = \|w\|_2^2$. Along a one-dimensional feasible set $w = w_{0,i} + zn$, with $w_{0,i} \perp n$ and $\|n\|_2 = 1$, write $a_i = b^\top w_{0,i}$, $\beta = b^\top n$, and $v_i = \|w_{0,i}\|_2^2 > 0$. Then

$$M_{\mathrm{KL}}(z) = \frac{a_i + \beta z}{\sqrt{v_i + z^2}}, \qquad \frac{dM_{\mathrm{KL}}}{dz} = \frac{\beta v_i - a_i z}{(v_i + z^2)^{3/2}}. \tag{118}$$

Its extrema occur at the interval endpoints or at the single interior stationary point when it exists. Across the ten directed query selections, all 100 signs of ΔM_0 are identified, but only 39 signs of ΔM_{KL} are identified. Of these 39, 33 are certified at $\rho = 0.2$, feasible rewards attain both finite signs in one case, and five remain unresolved. The other 61 comparisons already attain both signs of ΔM_{KL}. Thus, in most cases where the finite comparison is not identified, the sign of ΔM_{KL} is already not identified. Table 7 isolates the remaining case in which ΔM_{KL} has a fixed sign but the finite comparison does not.

Sensitivity to the norm bound. We keep the candidate class, target, reference, observations, and query selections fixed and enlarge the coefficient-norm bound to αB, for $\alpha = 1, 1.05, 1.10$, recomputing reachability and gain bounds over each enlarged feasible set.

Additional response banks. We reconstruct the empirical score class, target direction, and norm bound independently on the H4 SFT and OpenOrca banks and apply the same eight-score query

Table 7: A fixed positive sign of ΔM_{KL} does not determine the finite comparison. Both cases satisfy the same observed score differences and norm bound. Each difference is 4B minus 1.7B, multiplied by 1,000. Positive values favor 4B.

Case	$\Delta M_{\mathrm{KL}} \times 10^3$	$\Delta G_{0.2} \times 10^3$
A	$+4.809$	$+3.315$
B	$+0.276$	-0.032

Table 8: Matched-KL outcomes as the coefficient-norm bound is enlarged. The candidate class, target, observations, and query selections are fixed; enlarging B admits additional feasible rewards. Counts are over 100 comparisons at $\rho = 0.2$. For diagnostic purposes, unresolved cases are separated according to whether uniform KL reachability is verified; "KL unverified" means the sufficient reachability condition is not established.

Norm-bound multiplier	Query method	Certified	Both signs attained	Unresolved	KL unverified
1.00	Directed	33	62	5	0
1.05	Directed	17	76	7	0
1.10	Directed	7	84	9	0
1.00	Random	10	77	5	8
1.05	Random	5	85	2	8
1.10	Random	4	87	1	8

procedure at $\rho = 0.2$. The resulting outcomes are reported in Table 3. Because the response bank changes together with the empirical score geometry, this comparison does not isolate a generator or SFT effect.

B.9 Small-KL approximation and common-distribution reweighting

Small-KL approximation. Across the five original candidate RMs, the relative error of $\sqrt{2\rho}\, M_{\mathrm{KL}}$ increases with KL distance: the median error grows from 0.25% at $\rho = 10^{-4}$ to 14.9% at $\rho = 0.2$ and 22.7% at $\rho = 0.4$. Despite this magnitude error, all ten pairwise orderings agree with the exact Gibbs gains throughout the evaluated grid. This is an empirical observation for these five RMs, not a certificate over the feasible reward sets.

Common-distribution reweighting. We center and variance-normalize each candidate reward once, then evaluate every pair under the same shifted response distribution P_ρ, obtained by exponentially tilting the reference distribution with normalized Qwen3-0.6B scores. All ten local pairwise orderings persist through $\rho = 0.2$, and one reverses at $\rho = 0.4$. The sufficient bound in Theorem D.4 certifies none of these nonzero shifts, illustrating that the bound is conservative rather than necessary. Appendix A.3 separately shows that common reweighting can reverse a local comparison even when response support is unchanged.

B.10 Changing the target with a fixed candidate class

To isolate dependence on the target, we use the same four Skywork candidate RMs, the same six-score dispersed observations, and the same four-dimensional centered score class, while defining policy quality with either ArmoRM or InternLM. InternLM is excluded from the candidate set in both conditions. The two target directions have cosine similarity 0.938, yet the same observations identify 16 of 60 local pairwise comparisons under ArmoRM and only 8 of 60 under InternLM. Thus observations that constrain one target-relevant direction need not constrain another equally well. This experiment compares how identification changes across two learned target evaluators.

B.11 Static associations across response banks

To check whether the associations in Table 1 depend on measuring static metrics and optimized utility on the same responses, we separate the two measurements. Splitting the 128 prompts in half and measuring the metric on one half and matched-KL gain on the other gives accuracy–gain correlations between 0.860 and 0.999 across the two split directions on the four response banks.

Figure 5 provides a stronger separation by measuring each static metric on one response bank and optimized utility on another, using 16 responses per prompt in every bank. Accuracy and M_{KL} remain positively associated with matched-KL gain throughout this five-RM pool, whereas normalized reward variance has no consistent direction. Across this five-RM pool, the cross-bank correlations provide a descriptive robustness check.

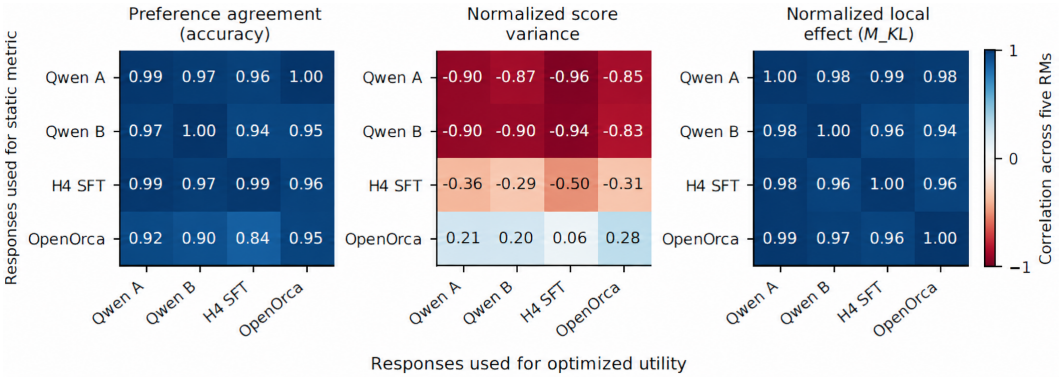


Figure 5: Cross-bank correlations between static RM metrics and matched-KL utility gain across five RMs. Rows indicate the response bank used to compute the metric and columns the bank used to compute ArmoRM gain at mean KL 0.2. Each cell is a Pearson correlation across five RMs; all banks are restricted to 16 responses per prompt. Panels show accuracy (preference agreement with ArmoRM), normalized reward variance $\overline{\mathrm{RV}}$, and the normalized local effect (M_{KL}).

C Related work

Partial identifiability in reward learning. Skalse et al. (2023) characterize the reward ambiguities left by several feedback sources through their invariances and analyze when these ambiguities matter for downstream tasks such as policy optimization. Lazzati & Metelli (2026) represent an incompletely identified target reward by a feasible set, quantify the resulting loss in downstream applications, and develop a worst-case robust approach with sample and iteration complexity guarantees. Our setting instead fixes the target utility and lets candidate RMs vary over the sets consistent with their static score observations. We ask whether those observations suffice to determine a pairwise downstream comparison, first at the reference policy and then after finite optimization.

Reward model evaluation and downstream outcomes. Wen et al. (2025) use controlled synthetic experiments to study how RM accuracy relates to optimized policy performance, showing that RMs with similar accuracy can yield substantially different downstream outcomes and that accuracy can fail to reflect overoptimization. Frick et al. (2025) construct proxy RM evaluations and study how their metrics correlate with crowd-sourced human preferences over post-DPO language models. Malik et al. (2026) similarly relate RewardBench 2 performance to downstream use in Best-of-N sampling and PPO. These works study predictive associations between static metrics and downstream outcomes. We instead ask whether the observed static information is sufficient to determine the downstream comparison for every reward consistent with it.

Reward error, ordering, and optimization. Gao et al. (2023) quantify reward overoptimization by tracking a gold reward while optimizing an imperfect proxy with reinforcement learning or Best-of-N. Fluri et al. (2025) formalize an error–regret mismatch in which low reward-model test error can coexist with large policy regret under distribution shift. Zhu et al. (2023) derive policy-learning

guarantees from pairwise and K-wise comparisons, showing that pessimistic reward estimation can yield good policies under coverage assumptions. Sun et al. (2025) argue that order consistency can suffice for response optimization in settings such as sampling-based optimization and contextual bandits. Our KL-regularized setting differs: identical response orderings need not induce the same optimized policy because the exponential tilt also depends on reward-gap magnitudes.

Reward variance and evaluation under policy shift. Razin et al. (2025) show that low reward variance under the initial policy can make policy-gradient reward maximization arbitrarily slow even for accurate RMs, and that the same RM can induce different variance across language models. They further contrast policy-gradient optimization with Best-of-N, for which perfect ranking accuracy is optimal. Lamparth et al. (2026) show that identical audit-distribution observables can be consistent with different downstream mitigation outcomes when the reference policy changes. They also show that ordinal benchmark functionals can be invariant to positive reward rescaling even as KL-regularized downstream return changes. Our analysis keeps the reference policy fixed and asks whether partial score observations determine the target-utility comparison; our counterexamples additionally hold complete ordering and reward variance fixed.

D Target information, reference changes, and parameterized updates

D.1 Identification across multiple target rewards

The identification direction depends on the target reward through b. Let $\mathcal{G}$ be a set of target rewards for which the required expectations are finite, and define

$$b(g) = \mathbb{E}_x \mathbb{E}_{\pi_0}\big[(g - \mathbb{E}_{\pi_0}[g \mid x])\phi\big], \qquad M_0^g(w) = b(g)^\top w. \tag{119}$$

Theorem D.1 (Identification across multiple target rewards)**.** *$M_0^g(w)$ is identified from Aw for every $g \in \mathcal{G}$ if and only if*

$$\operatorname{span}\{b(g) : g \in \mathcal{G}\} \subseteq \operatorname{row}(A). \tag{120}$$

Corollary D.2 (Identification for all bounded target rewards)**.** *For a bounded reward perturbation h,*

$$\mathbb{E}_x \operatorname{Cov}_{\pi_0}(g, h) = 0 \quad \textit{for every bounded } g \tag{121}$$

if and only if $h(x, y) = c(x)$, $\pi_0(\cdot \mid x)$-almost surely for almost every x.

Corollary D.3 (Invariance to prompt-wise additive constants)**.** *If two rewards differ by a prompt-wise additive constant $c(x)$, then*

$$\pi_t^{r+c} = \pi_t^r \qquad \textit{for all } t \geq 0. \tag{122}$$

The factor $e^{tc(x)}$ cancels in the Gibbs normalizer. Thus, with a fixed target, some nonconstant reward perturbations may leave its local effect unchanged. Requiring identification for every bounded target rules out all such perturbations except prompt-wise additive constants, which leave the entire Gibbs policy path unchanged. The information requirement therefore depends on whether the target is fixed or allowed to vary, consistent with the application-dependent invariances studied by Skalse et al. (2023). Proofs are given in Appendix A.2.

D.2 Stability under a common distribution shift

Let

$$\Delta M_Q = M_Q(r_2) - M_Q(r_1) \tag{123}$$

and define ΔM_P analogously for another response distribution P. This compares the same two rewards under a common response distribution; it does not compare their separately optimized policies.

Theorem D.4 (Stability of the pairwise local effect)**.** *If $|r_G| \leq R_G$ and $|r_2 - r_1| \leq B_r$ on the support, then*

$$|\Delta M_P - \Delta M_Q| \leq 6 R_G B_r \, d_{\mathrm{TV}}(P, Q), \tag{124}$$

where

$$d_{\mathrm{TV}}(P, Q) = \mathbb{E}_x \operatorname{TV}\big(P(\cdot \mid x), Q(\cdot \mid x)\big). \tag{125}$$

Hence,

$$|\Delta M_P - \Delta M_Q| \le 6R_G B_r \sqrt{\tfrac{1}{2}\mathbb{E}_x D_{\mathrm{KL}}\big(P(\cdot \mid x)\|Q(\cdot \mid x)\big)}. \tag{126}$$

If $\Delta M_Q = \gamma > 0$, the comparison remains positive whenever the right-hand side is smaller than γ. Failure of this sufficient condition is inconclusive.

Proposition D.5 (Sign reversal under response reweighting)**.** *There exist $P \neq Q$ with the same response support and rewards r_1, r_2 such that*

$$\Delta M_Q > 0, \qquad \Delta M_P < 0. \tag{127}$$

Thus changing response probabilities alone can reverse the local comparison without changing support. Proofs and the counterexample are given in Appendix A.3.

D.3 Score-function projection and one-step parameter updates

For a differentiable parameterized policy, let

$$s_\theta(x, y) = \nabla_\theta \log \pi_\theta(y \mid x) \tag{128}$$

denote the score function. For any $f(x, y)$ independent of θ, the score-function identity (Williams, 1992) gives

$$\nabla_\theta \mathbb{E}_{x,y\sim\pi_\theta}[f] = \mathbb{E}_x \operatorname{Cov}_{\pi_\theta}(f, s_\theta). \tag{129}$$

For each prompt, define

$$\tilde{f} = f - \mathbb{E}_{\pi_\theta}[f \mid x], \qquad g_x(f) = \mathbb{E}_{\pi_\theta}[s_\theta \tilde{f} \mid x], \qquad F_\theta(x) = \mathbb{E}_{\pi_\theta}[s_\theta s_\theta^\top \mid x]. \tag{130}$$

Let $F_\theta(x)^+$ denote the Moore–Penrose inverse and define

$$R^2_\perp(f; x, \theta) = \min_\alpha \mathbb{E}_{\pi_\theta}\big[(\tilde{f} - \alpha^\top s_\theta)^2 \mid x\big]. \tag{131}$$

Proposition D.6 (Score-function projection)**.**

$$\operatorname{Var}_{\pi_\theta}(f \mid x) = g_x(f)^\top F_\theta(x)^+ g_x(f) + R^2_\perp(f; x, \theta). \tag{132}$$

The first term is the squared $L_2(\pi_\theta)$ norm of the projection of $\tilde{f}$ onto the linear span of the score coordinates. The residual is orthogonal to this span and does not contribute to the first-order parameter gradient. Hence total reward variance alone does not determine that gradient.

Let

$$J_r(\theta) = \mathbb{E}_{\pi_\theta}[r]. \tag{133}$$

Proposition D.7 (One-step parameter update)**.** *Suppose at θ_0,*

$$\langle \nabla U(\theta_0), \nabla J_r(\theta_0) \rangle \le -\delta < 0, \qquad \|\nabla J_r(\theta_0)\| \le \Gamma, \tag{134}$$

and ∇U is L_U-Lipschitz in a neighborhood containing the update. For

$$0 < \eta \le \frac{\delta}{L_U \Gamma^2}, \qquad \theta_1 = \theta_0 + \eta \nabla J_r(\theta_0), \tag{135}$$

we have

$$U(\theta_1) - U(\theta_0) \le -\frac{\eta\delta}{2} < 0. \tag{136}$$

If J_r is also L_J-smooth, $\eta \le 1/L_J$, and $\nabla J_r(\theta_0) \neq 0$, then

$$J_r(\theta_1) - J_r(\theta_0) \ge \frac{\eta}{2}\|\nabla J_r(\theta_0)\|^2 > 0. \tag{137}$$

Thus a sufficiently small gradient step can increase the candidate reward while decreasing target utility. This is a one-step local statement and does not imply the same behavior over an entire training trajectory. Proofs are given in Appendix A.7.

D.4 LONG-HORIZON RLOO TRAINING

Training rewards and selection. We use two Pythia-2.8B reward models initialized from the AlpacaFarm-style SFT checkpoint (Dubois et al., 2023). RM A is trained for five epochs on 8,000 ArmoRM-relabeled UltraFeedback pairs, while RM B is trained for one epoch on 33,135 pairs, with 13.81% of pair labels reversed relative to the ArmoRM relabeling. Their final checkpoints are fixed throughout policy training. On a common fixed response bank, their pairwise accuracies are 0.6986 and 0.7044, and their exact Gibbs gains at KL 0.2 are 0.00777 and 0.00697. The RM pair was selected for similar static accuracy and differing exact-Gibbs outcomes on the fixed response bank. Figure 3 reports a subsequent long-horizon run of this pair and illustrates that relative target utility can change over training.

Policy training and evaluation. Both policies start from the same SFT initialization. We use the RLOO implementation of Razin et al. (2025), updating all policy parameters for 1,554 steps with 16 prompts and two responses per prompt, learning rate 10^{-7}, KL coefficient 0.05, and unnormalized rewards. Training samples at temperature 1 with a 512-token maximum.

Every 100 updates and at the endpoint, we evaluate the policies on 128 held-out prompts using eight fresh responses per prompt, temperature 1, top-p 1, and at most 320 new tokens. ArmoRM provides the target scores. Figure 3 reports pointwise bootstrap intervals for individual policy gains and simultaneous bootstrap bands for the paired difference. Prompts are resampled while keeping each prompt's eight responses together, so the bands quantify prompt-level evaluation uncertainty within this training run.

Policy-distance diagnostic. Panel (c) reports the mean sequence log-ratio to the fixed SFT reference, averaged over the preceding 25 updates. The late increase in policy B's ArmoRM score coincides with a sharp rise in this diagnostic. It is computed from training rollouts, whereas panels (a)–(b) use fresh held-out responses; Panel (c) therefore provides a policy-distance diagnostic based on the training rollouts.