Reward Bias Substitution: Single-Axis Bias Mitigations Redirect Optimization Pressure

Max Lamparth[*]
Stanford University

Daniel Fein
Stanford University

Andreas Haupt
Stanford University

Marcel Hussing
University of Pennsylvania

Mykel J. Kochenderfer
Stanford University

Abstract

Single-axis mitigations of reward-model biases (e.g., reducing proxy reliance on length, sycophancy, or style) can rotate optimization pressure onto correlated proxies rather than eliminate it, a failure mode we call reward bias substitution. The failure is enabled by a measurement-versus-optimization gap between audit and policy-induced distributions during mitigation evaluation and policy training. We formalize mitigation outcomes into a regime taxonomy and prove that successful mitigation, bias substitution, and overcorrection produce identical observables under any audit-distribution scoring, including ranking accuracy and win-rate, even when granted oracle access to the true reward. Across published preference-learning mitigation work, no method we survey reports the evidence needed to certify successful mitigation. Augmenting evaluation with policy-induced distributions while tracking multiple biases provably closes the gap, and we translate this into actionable prescriptions for mitigation methods and benchmarks. We demonstrate bias substitution in language model RLHF, where a length penalty during GRPO training compresses responses as intended yet redirects optimization pressure onto confidence calibration, driving the policy into overconfidence while factual free-form accuracy falls. We also show a published length-debiasing operator that zeroes reward–length correlation on the audit distribution but reintroduces bias under best-of-N selection on three of four SOTA reward models, and a length–sycophancy coupling whose direction reverses under human–LLM judge disagreement.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) and related preference-learning methods use human annotations to shape model behavior on non-verifiable objectives, but the resulting reward models and policies commonly learn spurious correlations. A recurring example is reward-length correlation in learned rewards and verbosity in the resulting policies, across both RLHF and DPO [Singhal et al., 2024, Park et al., 2024]. Many mitigations target such features directly, for example by penalizing length during training [e.g., Shen et al., 2023, Chen et al., 2024b, Bu et al., 2025] or controlling for length in evaluations [Dubois et al., 2024, Li et al., 2024]. But *removing a feature does not remove the pressure to exploit it.* Spurious features are partially informative about quality and correlated with one another, so suppressing one redirects optimization onto the rest, similar to shortcuts in supervised learning [Li et al., 2023]. Figure 1 illustrates the symptom for language model RLHF. A length penalty during GRPO [Shao et al., 2024] training compresses responses as intended, yet the freed optimization pressure rotates onto an untargeted axis, driving the policy into overconfidence and degrading free-form factual accuracy while multiple-choice accuracy holds (see

[*]Correspondence to: lamparth@stanford.edu

Preprint.

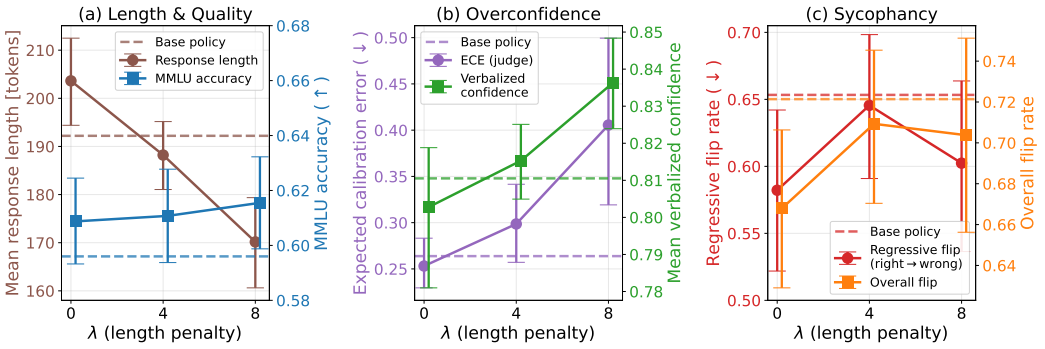


Figure 1: RLHF of *Llama-3.2-3B-Instruct* with *Skywork-Reward-V2-Llama-3.1-8B* using GRPO ($\beta = 2\text{e}{-2}$, LR $= 3\text{e}{-5}$, 600 steps) on *UltraFeedback* with four random seeds for each $\lambda \in \{0, 4, 8\}$, 95% bootstrapped confidence intervals. To mitigate length, we modify the reward $\tilde{R}$ in the RLHF objective as $\tilde{R}(x, y) = R_{\text{RM}}(x, y) - \lambda\, n_{\text{tok}}/100$. Training with $\lambda = 0$ does not break confidence calibration. As length is decreased the optimization pressure rotates onto the calibration axis, while MMLU accuracy and sycophantic behavior are preserved. See Section B.1 for details.

Section B.1). This phenomenon is not specific to length, nor RLHF, but a general failure mode we call **reward bias substitution**, in which a single-axis mitigation applied to a reward model or the implicit rewards of a policy relocates optimization pressure onto correlated proxies (see Figure 2).

Crucially, *bias substitution is invisible* to how mitigations are evaluated, a *measurement-versus-optimization gap* between audit and policy-induced distributions. On the audit distribution, a successful mitigation, a relocation of the bias onto another feature, and an overcorrection lowering true reward can all produce the same observable scores. Across the published preference-learning mitigation methods we survey, none report the evidence needed to certify that they removed a bias rather than relocated or overcorrected it, and the resulting substitution is routinely miscategorized as an isolated anomaly, a capability trade-off, or judge noise (Section 4, Section D). Existing reward-hacking definitions [Skalse et al., 2022, Laidlaw et al., 2025] do not consider mitigation operators and cannot separate removing a bias from relocating it. Left unaddressed, bias substitution risks silently trading one bias for another in deployment, leading to potential harm in user-facing interactions.

We prove this blindspot is structural, as no audit-distribution score can tell successful mitigation, bias substitution, and overcorrection apart, no matter how the benchmark is enriched and even when granted oracle access to the true reward. Similarly, multi-axis mitigation cannot help while still validated away from the policy-induced distribution, and causal identification would demand assumptions that preference data cannot grant. However, we prove that evaluating on the policy-induced distribution while tracking off-target features is sufficient to separate the three outcomes for explicit-reward RLHF and direct preference optimization. Concretely, a reward bias mitigation must be validated against at least two fixed policies under best-of-N or policy optimization, reporting the induced drift in length, confidence, and sycophancy alongside the usual audit score and correctness.
This paper contributes the following:

- **We formalize reward bias substitution**, and identify the measurement-versus-optimization gap as the structural mechanism making it invisible to prior reward-hacking definitions.
- **We introduce a regime taxonomy** classifying all distinct single-axis mitigation outcomes: successful mitigation (R0, with a contaminated sub-case R0_{cont}), bias substitution (R1), overcorrection (R2), silent non-op (R3), and audit-distribution sensitivity (R4).
- **We prove a matched impossibility-sufficiency pair**. No benchmark functional over audit-distribution observables can separate successful mitigation from substitution or overcorrection, while augmenting with policy-induced distributions provably can, and we turn the sufficient condition into prescriptions for methods and benchmarks.
- **We provide empirical evidence for our framework.** We validate substitution end-to-end in GRPO RLHF (length traded for overconfidence, R1), the measurement-versus-optimization gap on a length-debiasing operator across five reward models, and the first systematic characterization of length-sycophancy dependence across eight models from

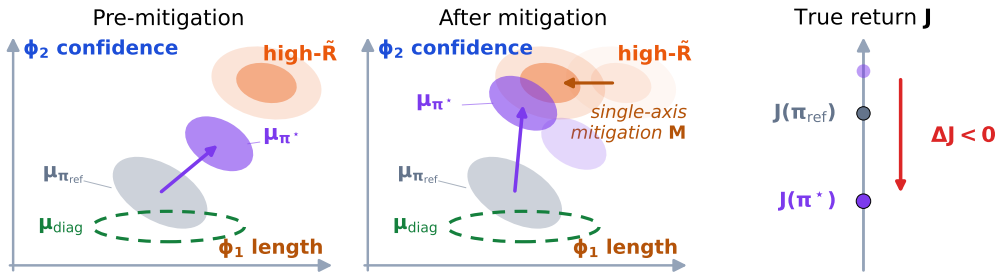


Figure 2: Reward bias substitution mechanism from Figure 1 (schematic). The correlated spurious axes are ϕ_1 (length) and ϕ_2 (confidence). Before mitigation (left), the optimized policy $\mu_{\pi^\star}$ sits in a high proxy reward region. A single-axis length mitigation M (middle) cuts length reliance and reroutes optimization pressure onto confidence. The rerouting leaves no trace at μ_{diag} and appears only at $\mu_{\pi^\star}$. For Figure 1, true return J falls even though the length bias looks removed (right). This gap is why audit-distribution scores alone cannot certify a real mitigation, formalized in Theorem 3.9.

different families, whose sign reverses under human–LLM judge disagreement (R4). We further reclassify previously isolated findings as R1 and R2 across the mitigation literature.

Our code for the experiments is available on GitHub[2] (MIT license).

2 Two Invariances Enabling Bias Substitution

We work in the single-turn contextual-bandit reduction of RLHF [Christiano et al., 2017, Stiennon et al., 2020, Rafailov et al., 2023, Kaufmann et al., 2025] with prompts $x \in \mathcal{X}$ drawn i.i.d. from a fixed context distribution $\mathcal{D}$, responses $y \in \mathcal{Y}$ sampled $y \sim \pi(\cdot \mid x)$, true reward $R(x, y)$, learned proxy $\tilde{R}(x, y)$, and the *plain return* $J(\pi, R) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(\cdot|x)}[R(x, y)]$ as the quantity we want to improve. The KL-regularized RLHF training objective is

$$J_{\text{RLHF}}(\pi; \tilde{R}) = \mathbb{E}_{x\sim\mathcal{D}}\Big[\mathbb{E}_{y\sim\pi(\cdot|x)}[\tilde{R}(x, y)] - \beta\, D_{\text{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big)\Big]. \tag{1}$$

For any reward $\bar{R}$ the KL-regularized optimum is the softmax policy [Peters and Schaal, 2007, Rafailov et al., 2023]

$$\pi^\star_\beta(\bar{R})(y \mid x) \propto \pi_{\text{ref}}(y \mid x) \exp\big(\bar{R}(x, y)/\beta\big). \tag{2}$$

Let $\mu_\pi(x, y) = \mathcal{D}(x)\pi(y \mid x)$ denote the policy-induced distribution on $\mathcal{X} \times \mathcal{Y}$ given a policy π (the one-step occupancy measure under $\mathcal{D}$). In addition to the π_{ref}-defined measure $\mu_{\pi_{\text{ref}}}$, we fix a *diagnostic measure* μ_{diag} on $\mathcal{X} \times \mathcal{Y}$, used for reliance estimation and correlation measurement (audit distribution). Natural choices include $\mu_{\pi_{\text{ref}}}$ itself (coupling measurement and optimization) and annotator-conditioned audit distributions $\mu^{\text{human}}_{\text{diag}}, \mu^{\text{LLM}}_{\text{diag}}$, reflecting that reliance estimates depend on who labels (see also Section A.3 for human-versus-LLM-judge gap). A standing regularity Assumption A.1 is invoked throughout for well-defined softmax policies and policy-level expectations.

Reward modeling in RLHF involves invariances at two levels. First, what preference data identifies about the reward (data-level), and second, how the KL-regularized policy responds to reward transformations (policy-level). At the *data level*, the reward identifiability problem leaves the learned reward underdetermined [Kim et al., 2021, Skalse et al., 2022, Tien et al., 2023, Casper et al., 2023]. Many reward functions explain the same preference data and comparisons identify the true reward only up to a prompt-only additive shift under the Bradley-Terry likelihood, leaving the cardinal scale loosely pinned [Skalse et al., 2023]. Under this underdetermination, a spurious feature can be indistinguishable from a structurally relevant one. At the *policy level*, KL-regularized optimization is sensitive to cardinal reward values, since $\tilde{R} \to c\tilde{R}$ at fixed β is equivalent to $\beta \to \beta/c$, so two reward models agreeing on all pairwise preferences can still induce different policies. Which reward transformations preserve the optimal policy is classical [Ng et al., 1999], but the relevant

[2] github.com/maxlampe/bias_substitution

invariance here is that same prompt-only additive shift, which the single-axis mitigations we study break. Combined with data-level identifiability, this means confounded features can be consistent with the observed rankings while still distorting the cardinal values that determine the optimized policy. A single-axis mitigation acts on exactly such a feature, so removing a bias, relocating it onto a correlated feature, and overcorrecting can all present the same audit-side evidence (Figure 2). Whether any audit-distribution benchmark can separate them is a claim about all benchmarks at once, not something one experiment can settle, so we prove it in Section 3.r

3 Formalizing Bias Substitution

We turn these invariances into a taxonomy of single-axis mitigation outcomes, characterizing what audit-distribution evaluation can and cannot certify, and deriving prescriptions for mitigation methods and benchmarks. We state the framework for explicit rewards, see Section A.12 for implicit rewards.

3.1 Feature Map and Spurious vs. Structurally Relevant Features

In this work, we are interested in categorizing and analyzing the effects of bias mitigation strategies given a multiplicity of axes that we may care to optimize for. We will refer to these axes as (surface) features. Before we can disentangle their interactions with rewards, we need to define features.

Definition 3.1 (Feature map)**.** *A* feature map *is a measurable function $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ capturing an interpretable surface attribute of a response. To form a set of features that we care about, we fix a finite ordered tuple of feature maps $\phi_1, \ldots, \phi_K$ and collect them into the vector-valued map $\Phi = (\phi_1, \ldots, \phi_K)^\top : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$.*

We use Φ both as the column-vector-valued map $(\phi_1, \ldots, \phi_K)^\top$ and, by slight abuse, as the set $\{\phi_1, \ldots, \phi_K\}$ when membership ($\phi_i \in \Phi$) or partitions ($\Phi = \Phi_{\mathrm{sp}} \sqcup \Phi_{\mathrm{struct}}$) are more natural.

To instantiate our feature sets, we will make the natural assumption that the chosen feature sets capture distinct response attributes under the diagnostic distribution. For example, features such as length, sycophancy, politeness, or hedging may be correlated, but they should not collapse into the same measurement across diagnostic responses. This rules out duplicate or exactly redundant features and ensures that each feature represents a separately identifiable axis of response behavior.

Assumption 3.2 (Non-degeneracy)**.** *The Gram matrix $G := \mathbb{E}_{\mu_{\mathrm{diag}}}[\Phi\Phi^\top] \in \mathbb{R}^{K \times K}$, with entries $G_{ij} = \mathbb{E}_{\mu_{\mathrm{diag}}}[\phi_i \phi_j]$, is positive definite.*

With these distinct axes in place, we can ask which attributes the reward actually depends on.

Definition 3.3 (Spurious vs. structurally relevant at μ_{diag})**.** *Treat R as a function on $\mathcal{X} \times \mathcal{Y}$ and assume feature realizability (Assumption A.2). Feature ϕ_i is* spurious with respect to R at μ_{diag} *if*

$$\mathbb{E}_{\mu_{\mathrm{diag}}}\big[R(x,y) \,\big|\, x,\ (\phi_j(x,y))_{j \neq i}\big] \text{ does not depend on } \phi_i(x,y) \text{ for } \mu_{\mathrm{diag}}\text{-a.e.} \tag{3}$$

Otherwise ϕ_i is structurally relevant at μ_{diag}. *Let $\Phi_{\mathrm{sp}} = \{\phi_i \in \Phi : \phi_i$ spurious w.r.t. R at $\mu_{\mathrm{diag}}\}$ and $\Phi_{\mathrm{struct}} = \Phi \setminus \Phi_{\mathrm{sp}}$ for the induced partition of Φ.*

See Section A.4 for the equivalent $o(\varepsilon)$ formulation under ε-mixture perturbations, capturing the intuition that R has no first-order dependence on ϕ_i at μ_{diag}. The corresponding causal reading and partial-identifiability relationship are given in Section A.1. Neither spuriousness nor structural relevance is fully identifiable from preference data alone [Skalse et al., 2023]. Under partial informativeness, natural features like length act as mediators of multiple mechanisms. The binary partition of Φ from Definition 3.3 may therefore conservatively classify them as Φ_{struct} when they correlate with R under μ_{diag} (Section A.3). This partition of Φ also matches existing single-axis mitigations acting on whole features rather than within-feature decompositions [e.g. Fein et al., 2026, Papadatos and Freedman, 2024, Chen et al., 2024a, Huang et al., 2025]. Finer partitions require stronger assumptions (interventions, gold labels, distributional invariance) outside our regime.

3.2 Single-Axis Mitigation and Measurement-Versus-Optimization Gap

Several prominent reward-model mitigations identify a feature direction, estimate $\tilde{R}$'s reliance on it, and subtract its contribution as, e.g., linear probes [Fein et al., 2026, Papadatos and Freedman, 2024],

non-linear calibration [Huang et al., 2025], or architectural disentanglement [Chen et al., 2024a]. To abstract away from these implementation details, we first define a population-level notion of reliance, given by the coefficients of $\tilde{R}$'s best linear approximation in Φ under the diagnostic distribution.

Definition 3.4 (Linear reliance)**.** *The* linear reliance *of $\tilde{R}$ on Φ at μ_{diag} is*

$$g(\tilde{R};\, \mu_{diag}) = \left(\mathbb{E}_{\mu_{diag}}[\Phi\Phi^\top]\right)^{-1}\mathbb{E}_{\mu_{diag}}[\Phi\, \tilde{R}] \in \mathbb{R}^K, \tag{4}$$

yielding the $L^2(\mu_{diag})$-orthogonal decomposition $\tilde{R} = \sum_i g_i\phi_i + \tilde{R}_\perp$ with $\mathbb{E}_{\mu_{diag}}[\phi_i\tilde{R}_\perp] = 0$.

Note g is a *statistic* of the triple $(\tilde{R}, \Phi, \mu_{\text{diag}})$, not a property of $\tilde{R}$ alone and evaluating at $\mu_{\pi^\star} \neq \mu_{\text{diag}}$ yields a different vector, driving later regime distinctions in our taxonomy.

Definition 3.5 (Single-axis mitigation)**.** *A* single-axis mitigation targeting feature i *is an operator $M_i : \tilde{R} \mapsto \tilde{R}'$ with $|g_i(\tilde{R}';\, \mu_{diag})| < |g_i(\tilde{R};\, \mu_{diag})|$. The canonical projection instance is*

$$M_i(\tilde{R})(x, y) \;=\; \tilde{R}(x, y) - g_i(\tilde{R};\, \mu_{diag})\, \phi_i(x, y). \tag{5}$$

We single out the projection instance above as canonical, because it zeros g_i at μ_{diag} exactly. The canonical M_i depends on μ_{diag} through $g_i(\tilde{R}; \mu_{\text{diag}})$. We write $M_i^{\mu_{\text{diag}}}$ when this dependence matters.

Lemma 3.6 (Single-axis identity at μ_{diag})**.** *For the canonical M_i, $g_i(M_i(\tilde{R});\, \mu_{diag}) = 0$ and $g_j(M_i(\tilde{R});\, \mu_{diag}) = g_j(\tilde{R};\, \mu_{diag})$ for $j \neq i$.*

Proof. Direct computation using $\mathbb{E}_{\mu_{\text{diag}}}[\Phi\phi_i]$ as the i-th column of $\mathbb{E}_{\mu_{\text{diag}}}[\Phi\Phi^\top]$. □

The identity justifies calling M_i *single-axis*, as at μ_{diag}, mitigation moves $\tilde{R}$ along the i-th coordinate of g-space.[3] Note that M_i is an *associational* operator, as it removes ϕ_i-reliance at μ_{diag} in the projection sense, not the causal contribution of ϕ_i to R (see Appendix A.1). Our taxonomy classifies the resulting outcomes by what the *optimizer does at $\mu_{\pi^\star}$, not by what is observable at μ_{diag}*.

Measurement-versus-optimization gap. Single-axis diagnostics evaluate M_i at μ_{diag}, where $|g_i| = 0$ by construction (Lemma 3.6). The optimizing policy realizes M_i's effects at $\mu_{\pi^\star}$, where in general $g_i(M_i(\tilde{R}); \mu_{\pi^\star}) \neq 0$ and $g_j(M_i(\tilde{R}); \mu_{\pi^\star}) \neq g_j(\tilde{R}; \mu_{\pi^\star})$ for $j \neq i$. Mitigation moves the proxy along an axis defined at the audit distribution, but optimization responds to a vector defined at a different distribution (see Figure 2). This gap is the structural mechanism enabling bias substitution, and it is invisible to any diagnostic that operates at μ_{diag} alone. Section 3.4 shows it cannot be closed by any audit-distribution-only evaluation and a closed-form instance is given in Sections A.7 and A.9.

Gauge invariance. The linear reliance g is not invariant under prompt-only reward shifts $\tilde{R}(x, y) \mapsto \tilde{R}(x, y) + b(x)$, while the KL-regularized optimum is, which matters for PPO-style reward whitening [Lambert, 2025]. Section A.6 gives a gauge-invariant g_{cent} and verifies that our taxonomy transfers.
Scale invariance. Applying M_i also changes $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$, and KL-regularized policies are not invariant to reward scale [Skalse et al., 2023], so this scale change interacts with optimization non-trivially. Section A.5 gives the corrected scale identity and a scale-invariant variant M_i^{norm}.

3.3 Exploitation Shift and Bias Substitution

Under optimization, the measurement-versus-optimization gap produces several distinct failure modes, which we classify along two axes, the change in true reward at the optimized policy and whether optimization pressure rotates onto another spurious feature. Throughout, fix $\beta > 0$, let $\tilde{R}$ be a proxy reward, let $\phi_i \in \Phi_{\text{sp}}$ be the targeted feature, and let M_i be a single-axis mitigation (Definition 3.5) targeting ϕ_i, with $\tilde{R}$ and $M_i(\tilde{R})$ satisfying the regularity Assumption A.1.

Definition 3.7 (Regime Taxonomy)**.** *Let $\pi = \pi^\star_\beta(\tilde{R})$ and $\pi' = \pi^\star_\beta(M_i(\tilde{R}))$ be the pre- and post-mitigation KL-regularized optima, and write $\Delta_j = \mathbb{E}_{\mu_{\pi'}}[\phi_j] - \mathbb{E}_{\mu_\pi}[\phi_j]$ for the induced drift in feature ϕ_j and $\Delta J = J(\pi', R) - J(\pi, R)$ for the change in true reward. Then M_i is in regime **R0**, **R0**$_{cont}$, **R1**, **R2**, or **R3** according to the rotation predicate on $\Phi_{\text{sp}} \setminus \{\phi_i\}$ and the sign of ΔJ,*

[3] The canonical M_i can also be seen as the population Frisch-Waugh-Lovell partial-out of ϕ_i from $\tilde{R}$ at μ_{diag}.

True reward change	*No rotation* ($\Delta_j = 0$ *on* $\Phi_{\text{sp}} \setminus \{\phi_i\}$)	*Rotation* ($\Delta_j \neq 0$ *for some* $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$)
$\Delta J > 0$	*R0 Successful mitigation*	*R0$_{cont}$ Contaminated success*
$\Delta J = 0$	*R3 Silent non-op*	*R1 Bias substitution (neutral)*
$\Delta J < 0$	*R2 Overcorrection*	*R1 Bias substitution (harmful)*

where R1 spans both rotation cells with $\Delta J \leq 0$*, neutral when* $\Delta J = 0$ *and harmful when* $\Delta J < 0$*.*

In the GRPO example in Figure 1 and Section B.1, penalizing length (the targeted ϕ_i) rotates pressure onto expressed confidence while factual accuracy drops, placing the mitigation in the bottom-right cell (harmful R1). Section A.3 verifies the taxonomy is exhaustive and Section A.8 gives the ε-banded versions of Δ_j and ΔJ for finite-sample studies that we use throughout Section 4 to classify published methods under noise-floor uncertainty. The regimes classify mitigations by quantities depending on R and Φ_{sp}, neither fully identifiable from preference data alone, and the rotation predicate references only Φ_{sp} because rotation onto structurally relevant features is not penalized.

Remarks. R0_{cont} is the outcome audit-distribution evaluation most easily mistakes for R0, where the rotation is active but its cost is outweighed by the gain from reducing ϕ_i, so evaluation registers the improvement but not the rotation. R1 is the regime the measurement-versus-optimization gap specifically enables, because $|g_i(M_i(\tilde{R}); \mu_{\text{diag}})| = 0$ holds at the audit distribution by construction (Lemma 3.6) while $\Delta J \leq 0$ at the optimized policy, so audit diagnostics targeting ϕ_i register apparent success regardless of the sign of ΔJ. R2 has two origins: scale overshoot, fixable by rescaling, and target misspecification, which removes a genuinely informative component of a partially informative feature and generically is not. Section A.13 distinguishes them via $\Delta J(c)$ across partial mitigations M_i^c. R3 is silent because the mitigation alters the proxy in ways the audit distribution may register while the optimized policy does not express them, the generic outcome when the targeted feature's optimization footprint was already small, leaving $D_{\text{KL}}(\pi' \,\|\, \pi)$ small at the relevant β.

The taxonomy in Definition 3.7 fixes a single audit distribution μ_{diag}, but cannot cover failure modes where the regime label changes when μ_{diag} does, e.g., across human and LLM-judge audits. Thus:

Definition 3.8 (R4 Audit-distribution sensitivity)**.** *A mitigation construction* M_i *exhibits* audit-distribution sensitivity *if, holding* $(\tilde{R}, R, \Phi_{sp}, \beta)$ *fixed, there exist* $\mu^{(1)}_{diag} \neq \mu^{(2)}_{diag}$ *such that*

$$\pi^\star_\beta\big(M_i^{(1)}(\tilde{R})\big) \quad \textit{and} \quad \pi^\star_\beta\big(M_i^{(2)}(\tilde{R})\big) \quad \textit{fall into different R0–R3 regimes, where } M_i^{(\ell)} := M_i^{\mu^{(\ell)}_{diag}}.$$

R4 is regime-labeled (rather than treated as a property of the deployment setting, like π_{ref}-sensitivity) because μ_{diag} is a designer-controlled input to the mitigation pipeline[4] and formalizes the observation that failure-mode classification is a property of $(R, \tilde{R}, M_i, \mu_{\text{diag}})$. See Figure 3 for example transition.

Instantiations. Definitions 3.7 and 3.8 cover the mechanistic outcome space, with closed-form instantiations of all regimes given in the linear-Gaussian and quadratic-nonlinear settings of Sections A.7 and A.9. The R0–R3 conditions reference only Δ_j and ΔJ, both invariant under the prompt-only shifts. The canonical M_i is not gauge-invariant, but M_i^{cent} (Section A.6) restores classification invariance. See also Section 4 for how published mitigations map onto these regimes.

3.4 Impossibility and Sufficiency for Audit-Distribution Evaluation

The measurement-versus-optimization gap raises the question whether any benchmark functional defined at the audit distribution can distinguish successful mitigation (R0) from contaminated success (R0_{cont}), substitution (R1), or overcorrection (R2). We prove that they cannot and show that augmenting evaluation with policy-induced distributions suffices (see also Sections A.10 and A.11).

Benchmark class. Let $\mathcal{B}$ denote the class of benchmark functionals depending only on the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag}, with μ_{diag} taken as the empirical evaluation measure. This class subsumes ranking accuracy, pairwise win-rate, preference-prediction calibration, reward–target correlation, the linear-reliance statistic g of Definition 3.4, and any composite of

[4] μ_{diag} is non-performative in the sense of Perdomo et al. [2020].

these. RewardBench [Lambert et al., 2025], RewardBench2's headline accuracy [Malik et al., 2026], AlpacaEval [Dubois et al., 2024], and Chatbot Arena [Chiang et al., 2024] all lie in $\mathcal{B}$. Including R as an input strengthens our impossibility result, since R appears identically across the four instances and standard benchmarks lack oracle access to it.

Theorem 3.9 (Audit-distribution insufficiency)**.** *Fix $\beta > 0$. There exist four choices of π_{ref} with μ_{diag}, $\tilde{R}$, M_i, R, Φ, Φ_{sp}, and β held fixed s.t.*

(i) *the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag} is identical across the four instances, so B takes the same value on all four for every $B \in \mathcal{B}$,*

(ii) *the optimized policies $\pi^\star_\beta(M_i(\tilde{R}))$ fall into regimes R0, R0$_{cont}$, R1, and R2 respectively, in the sense of Definition 3.7.*

The proof (Section A.10) constructs four reference policies in a linear-Gaussian setting with fixed Gram, so audit-side observables coincide while policy-side first moments place the optimized policies in distinct regimes. Since only π_{ref} varies, the same reward model can land in any of R0, $\mathrm{R0}_{\mathrm{cont}}$, R1, R2 depending on the reference policy it is deployed against, matching the finding of Malik et al. [2026] that high audit scores can fail to transfer to PPO under RM-policy lineage mismatch.

Theorem 3.9 establishes the *distributional blindspot*. A second, independent blindspot, the *functional blindspot*, affects ordinal benchmarks: $\mathcal{B}_{\mathrm{ord}} \subseteq \mathcal{B}$ (functionals invariant under monotone transformations of $\tilde{R}$) is blind to cardinal-scale shifts that KL-regularized optimization tracks (Corollary A.5). Since mitigation generically induces rescaling (Section A.5), ordinal-only evaluation cannot rule out scale-driven regime shifts even when the proxy is unchanged in ordinal content. We realize a non-linear instance of post-hoc length calibration [Huang et al., 2025] in Section B.2.

We show that augmenting the benchmark input with the policy-induced distributions provably suffices. Let $\mathcal{B}^+$ denote the class of benchmark functionals depending on the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under each of $\mu_{\mathrm{diag}}, \mu_{\pi^\star_\beta(\tilde{R})}, \mu_{\pi^\star_\beta(M_i(\tilde{R}))}$.

Theorem 3.10 (Audit-distribution sufficiency)**.** *Fix $\beta > 0$ and let $\pi = \pi^\star_\beta(\tilde{R})$, $\pi' = \pi^\star_\beta(M_i(\tilde{R}))$. With Δ_j and ΔJ as in Definition 3.7 and Section 3.3, define $B^\star \in \mathcal{B}^+$ by*

$$B^\star = \begin{cases} \mathrm{R0} & \text{if } \Delta J > 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}, \\ \mathrm{R0}_{\mathrm{cont}} & \text{if } \Delta J > 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}, \\ \mathrm{R1} & \text{if } \Delta J \leq 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}, \\ \mathrm{R2} & \text{if } \Delta J < 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}. \end{cases} \tag{6}$$

Under Assumptions A.1, 3.2, and A.2, with $M_i(\tilde{R})$ likewise satisfying Assumption A.1, for every tuple $(\pi_{\mathrm{ref}}, \tilde{R}, M_i, R, \Phi, \Phi_{\mathrm{sp}})$ with $(\pi, \pi') \in \mathrm{R0} \cup \mathrm{R0}_{\mathrm{cont}} \cup \mathrm{R1} \cup \mathrm{R2}$ in the sense of Definition 3.7, $B^\star$ returns the correct regime label.

Proof (Section A.11): Δ_j and ΔJ are first moments under μ_π and $\mu_{\pi'}$, both in $\mathcal{B}^+$. The four branches are mutually exclusive by Definition 3.7, where the sign of ΔJ separates R0 and $\mathrm{R0}_{\mathrm{cont}}$ from R1 and R2, and the rotation condition separates within each pair. A fifth branch ($\Delta J = 0 \wedge \Delta_j = 0$) extends the classifier to R3, with $D_{\mathrm{KL}}(\pi' \| \pi)$ separating policy-relevant from policy-irrelevant R3. A finite-sample version under noise-floor ε-bands follows from the same construction (Section A.8).

What this pair establishes. Theorems 3.9 and 3.10 tell us that audit-only inputs cannot separate R0, $\mathrm{R0}_{\mathrm{cont}}$, R1, and R2, **even when R itself is granted as an oracle input.** Augmenting with policy-induced distributions and the spurious-feature partition closes the gap.

3.5 Implications for Mitigations Methods and Evaluations

We translate the findings of Theorems 3.9 and 3.10 into prescriptions for new reward bias mitigation works. We state the corresponding prescriptions for benchmark developers, regime-specific gaps not closed by $B^\star$, multi-axis recommendations, and operator-variant caveats in Section A.13.

Method paper prescriptions. A paper claiming successful reward bias mitigation should:

1. *Evaluate at policy-induced distributions and report* $(\Delta_j, \Delta J)$. By Theorem 3.10 this input separates R0, R0_{cont}, R1, and R2, as no audit-only input does (Theorem 3.9). Concretely: Run the mitigated reward model in BoN or short PPO/GRPO against a fixed reference policy.
2. *Instrument off-target* Δ_j *on every* $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$ *at* $\mu_{\pi^\star}$. Reporting only on-target $|g_i| \approx 0$ leaves R0 structurally indistinguishable from R0_{cont} and R1, regardless of audit-side strength. Concretely: At least track length, confidence, and sycophancy, as instantiated in Section 4.
3. *Report cardinal scale* $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ *pre/post mitigation.* By Corollary A.5, ordinal scores are invariant under $\tilde{R} \mapsto c\tilde{R}$ $(c > 0)$ while $J(\pi^\star_\beta(c\tilde{R}), R)$ varies with c. Since mitigation induces rescaling (Section A.5), ordinal-only evaluation misses scale-driven regime shifts.
4. *Document* π_{ref}*-sensitivity and gauge dependence.* Theorem 3.9 produces four distinct regimes by varying only π_{ref} for fixed $(\tilde{R}, M_i, \mu_{\text{diag}})$. The canonical M_i is also sensitive to prompt-only reward shifts. Validate across at least two reference policies of different lineages and report under M_i^{cent} (Section A.6) to remove the gauge degree of freedom.

Joint adoption. R0 certification is a joint property of the mitigation operator and the evaluation protocol. Method papers cannot unilaterally provide policy-distribution evaluation without supporting benchmark infrastructure and benchmarks cannot unilaterally provide off-target Δ_j measurement without method papers specifying Φ_{sp}. Joint adoption moves the field from R0 claims audit-side scores cannot certify (Theorem 3.9) to R0 claims this framework certifies sufficient (Theorem 3.10). Extended prescriptions, including benchmark developer recommendations and multi-axis operator guidance, are in Section A.13. We map published mitigation work onto these regimes in Section 4.

4 Bias Substitution in the Wild

Applied to the prescriptions of Section 3.5, no method we survey in the published RLHF and DPO mitigation literature reports the evidence Theorem 3.10 requires to certify successful mitigation (R0). Bias substitution is already pervasive yet routinely miscategorized as an anomaly, a capability trade-off, or judge noise, because no prior framework distinguishes it from successful mitigation.

Reward models exhibit correlated, partially informative spurious features. Bias substitution requires correlated, partially informative spurious features under μ_{diag}, both hold in deployed reward models (Fact C.1–C.17, C.20). Partial informativeness anchors R2's target-misspecification origin, and correlated axes provide directions for pressure rotation. We extend this with length-sycophancy coupling measurements across labeling regimes (Fact C.9, see Section B.3). Across LLM responses to Reddit prompts from Cheng et al. [2026a] on eight models from different families, the sycophantic effect on length is $+24.3$ characters under human labels, $+128.4$ under LLM judge labels, $+154.3$ under judge agreement, and $\mathbf{-43.1}$ under judge disagreement (all $p < 0.01$). The sign reversal documents R4's audit-dependent operator construction empirically. Section A.9 isolates the same mechanism in closed-form phase diagrams of regime transitions driven by these couplings.

Bias substitution arises end-to-end under RLHF training. We run GRPO [Shao et al., 2024] on *Llama-3.2-3B-Instruct* [Grattafiori et al., 2024] with the Skywork-*Reward-V2-Llama-3.1-8B* reward model [Liu et al., 2025], shaping the reward as $\tilde{R}(x, y) = R_{\text{RM}}(x, y) - \lambda\, n_{\text{tok}}(y)/100$ for $\lambda \in \{0, 4, 8\}$ with four seeds per value. This shaping has the form of the canonical single-axis mitigation M_i of Definition 3.5, with a fixed coefficient in place of the estimated reliance. As λ rises, mean response length falls from 204 to 170 tokens and multiple-choice MMLU accuracy is unchanged, yet expected calibration error climbs from 0.25 to 0.41, free-form factual accuracy on TriviaQA falls from 0.56 to 0.42, and confidence-correctness AUROC drops from 0.73 to 0.65 (Figure 1, Table 3). The $\lambda = 0$ run leaves calibration intact, so the degradation is caused by the penalty and not by RLHF itself. Optimization pressure rotates off length onto expressed confidence and the true reward proxy of free-form factual accuracy degrades, instantiating harmful R1 in a standard RLHF pipeline. Full setup, training curves, and per-metric results are in Section B.1.

Bias mitigations that pass audit can fail under optimization. We show that the measurement-versus-optimization gap fails to close with previously published mitigation operators in Section B.2.

We evaluate the post-hoc LOESS calibration of Huang et al. [2025] and the linear-probe operator of Fein et al. [2026] across five reward models under Best-of-N selection. The calibration drives pooled reward-length correlation from 0.316 to 0.037, an audit-side success by construction. All four SOTA reward models acquire negative within-prompt correlations under the calibration, with three exceeding their unmitigated baselines in absolute value. AlpacaEval LC win rate degrades below baseline on two cells and GSM8K BoN accuracy drops 3.6 points on one. Under the ε-banded reading of Section A.8, two cells satisfy R2_ε on the targeted axis, or mixed R1_ε+R2_ε with off-target axes unmeasured, both undetectable to any $B \in \mathcal{B}$ by Theorem 3.9.

Published methods cannot certify successful mitigations. Where direct evidence exists it points to R0_{cont}, R1, or R2, and otherwise the regime is undetermined, which is itself the failure mode Theorem 3.9 formalizes. Direct evidence for a failure regime appears in Bu et al. [2025] and Huang et al. [2025] for R2_ε on the targeted axis, Zhao et al. [2025] for R2 or R3, and Eisenstein et al. [2024] for R0_{cont} or R1. Every other surveyed method leaves the regime undetermined. The strongest mitigation candidates make the gap concrete. Fein et al. [2026] survives the within-prompt diagnostic of Section B.2 with $\Delta J > 0$ on four of five RMs but does not measure off-target Δ_j. Srivastava et al. [2026] engages multi-axis mitigation, BoN, and transformation robustness, but its LLM-oracle counterfactuals inherit unmeasured R4 sensitivity. Cai et al. [2026] includes PPO results but reports no off-target Δ_j. Umer et al. [2026] meets more prescriptions than other surveyed methods and achieves $\Delta J > 0$ on four held-out benchmarks through multi-axis online RL, but its Φ_{sp} panel is implicit and off-panel substitution remains unaddressed. None jointly delivers the inputs Theorem 3.10 requires, and no current benchmark can close the gap, since every benchmark we surveyed [Lambert et al., 2025, Malik et al., 2026, Chiang et al., 2024, Yan et al., 2026] sits inside the impossibility class $\mathcal{B}$ of Theorem 3.9 (see Section A.13). We discuss these and all other surveyed works in detail in Section D.

5 Related Work

Prior definitions of reward hacking. Reward bias substitution is a Goodhart-type failure that survives mitigation, as suppressing the reliance of the proxy reward on a spurious feature relocates optimization pressure onto correlated proxies rather than removing it. This differs from reward overoptimization [Gao et al., 2023], which concerns pushing a fixed proxy too far, whereas our failure mode is introduced by a mitigation operator and arises even under moderate optimization. Existing definitions [Skalse et al., 2022, Kwa et al., 2024, Fluri et al., 2025, Laidlaw et al., 2025] take a single proxy as given and ask whether optimizing it produces an acceptable policy, with Liu et al. [2026] extending to the worst-case proxy in a correlation set. None consider mitigation operators, so prior definitions cannot distinguish a mitigation that reduces reward hacking from one that relocates it, overshoots, or is ineffective. The rotation-onto-correlated-proxies mechanism has supervised-learning precedents in shortcut learning [Geirhos et al., 2020], simplicity bias [Shah et al., 2020], and Whac-A-Mole [Li et al., 2023], with a structural analog in fairness gerrymandering [Kearns et al., 2018]. Our contribution provides the first formal taxonomy of these failures, with KL-regularized RLHF as the primary instantiation. Two failure modes absent under empirical risk minimization arise specifically from our work: cardinal-scale sensitivity (Corollary A.5) and π_{ref} dependence (Theorem 3.9).

Reward bias mitigation methods and benchmarks. We position our work with respect to prior mitigation methods and benchmarks in Section 4 and Section D. Our framework classifies what single-axis operators can certify rather than proposing a new one. Concurrent work addresses adjacent problems: probabilistic reward modeling [Chen et al., 2026], identifiability under correlated probit models [Cherapanamjeri et al., 2026], emergent misalignment from narrow SFT [Betley et al., 2026], heterogeneous safety drift under benign fine-tuning that standard proxies fail to predict [Khan et al., 2026], and a KL-minimal sycophancy correction [Shapira et al., 2026]. None formalize how single-axis mitigations redistribute optimization pressure onto correlated proxies.

6 Discussion

Limitations. The regime classification uses first-moment drift, which can miss tail shifts for bounded features such as sycophancy indicators. We argue that reward model biases admit to first order corrections to limit the practical severity of this gap (Section C). Definition 3.3 partitions Φ

into spurious and structurally relevant features rather than decomposing within-feature. However, our approach is conservative under partial informativeness, as it under-flags substitution rather than over-flagging clean mitigations (Section A.3). Also, as soon as anyone targets a feature for debiasing, they assert it belongs to Φ_{sp}, so our taxonomy and impossibility result apply regardless of whether Φ_{sp} is ground-truth identifiable from preference data.

Impact. Current improvement metrics for reward bias mitigation cannot distinguish a model that has become less reward-hacky from one that has merely shifted exploitation modes, and this ambiguity propagates directly into post-deployment alignment claims. Our taxonomy and impossibility-sufficiency pair apply to any preference-learning setting with single-axis mitigation, and we develop language RLHF in detail because it is the primary lever for non-verifiable behavior, where substitution is already active across published SOTA reward models and operators rather than hypothetical. However, the gap is closable. The checklist we provide turns R0 claims that audit-side scores cannot certify into claims our framework certifies, once jointly adopted by method and benchmark developers. In high-compute settings the same prescriptions scale to multi-axis mitigation, but only when evaluated at the policy-induced distribution rather than the audit distribution, and even then it certifies "no substitution within the measured panel," not "no substitution," since off-panel axes remain available for rotation. Length is the natural anchor, the central axis through which sycophancy and confidence couple (Facts C.9, C.11, C.12), and identifying the rest is the direction we most want pursued next.

Acknowledgments

Max Lamparth is supported through a grant from Coefficient Giving (formerly Open Philanthropy), Stanford's Hoover Institution Tech Policy Accelerator, and the Stanford Intelligent Systems Laboratory. Marcel Husing was partially supported by DARPA grant #HR00112420305. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views, position, or policy of DARPA or the US Government.

References

Anthropic. Models overview. Blog post, 2026. Accessed: 2026-05-06.

Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649:584–589, 2026. doi: 10.1038/s41586-025-09937-5.

Yuyan Bu, Liangyu Huo, Yi Jing, and Qing Yang. Beyond excess and deficiency: Adaptive length bias mitigation in reward models for RLHF. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3091–3098, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.169.

Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Yue Wang, Li Li, Wengang Zhou, and Houqiang Li. Disentangling length bias in preference learning via response-conditioned modeling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=hKxYESOzen`.

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

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: disentangled reward mitigates hacking in rlhf. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Longze Chen, Lu Wang, Renke Shan, Ze Gong, Run Luo, Jiaming Li, Jing Luo, Qiyao Wang, and Min Yang. Learning ordinal probabilistic reward from preferences. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0Vf5trUAVF.

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391:eaec8352, 2026a. doi: 10.1126/science.aec8352.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=igbRHKEiAs.

Yeshwanth Cherapanamjeri, Constantinos Costis Daskalakis, Gabriele Farina, and Sobhan Mohammadpour. Learning correlated reward models: Statistical barriers and opportunities. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TbEyl6krsY.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Brian Christian, Jessica A F Thompson, Elle, Vincent Adam, Hannah Rose Kirk, Christopher Summerfield, and Tsvetomira Dumbalska. Reward models inherit value biases from pretraining. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dT399j1Azv.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168. arXiv:2110.14168.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Magda Dubois, Cozmin Ududec, Christopher Summerfield, and Lennart Luettgau. Ask don't tell: Reducing sycophancy in large language models, 2026. URL https://arxiv.org/abs/2602.23971.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5u1GpUkKtG.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 893–900, 2025. doi: 10.1609/aies.v8i1.36598.

Daniel Fein, Max Lamparth, Violet Xiang, Mykel J. Kochenderfer, and Nick Haber. One bias after another: Mechanistic reward shaping and persistent biases in language reward models, 2026. URL https://arxiv.org/abs/2603.03291.

Xuan Feng, Bo An, Tianlong Gu, Liang Chang, Fengrui Hao, Peipeng Yu, and Shuai Zhao. C2po: Diagnosing and disentangling bias shortcuts in llms, 2025. URL https://arxiv.org/abs/2512.23430.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 17306–17377. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/fluri25a.html.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in RLHF. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=62A4d5Mokc.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence volume 2, pages665–673(2020)*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.121.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. Explaining length bias in LLM-based preference evaluations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6763–6794, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.358.

Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo Ponti, and Ivan Titov. Post-hoc reward calibration: A case study on length bias. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Iu8RytBaji.

Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Vishakh Padmakumar, Ilia Sucholutsky, Andrew Strait, Diyi Yang, Q. Vera Liao, and Umang Bhatt. Measuring and mitigating overreliance to build human-compatible ai, 2026a. URL https://arxiv.org/abs/2509.08010.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652:1159–1165, 2026b. doi: 10.1038/s41586-026-10410-0.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=f7OkIurx4b. Survey Certification.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018.

Emaan Bilal Khan, Amy Winecoff, Miranda Bogen, and Dylan Hadfield-Menell. Safety drift after fine-tuning: Evidence from high-stakes domains, 2026. URL https://arxiv.org/abs/2604.24902.

Hyeonji Kim, Sujeong Oh, and Sanghack Lee. Mitigating length bias in rlhf through a causal lens, 2025. URL https://arxiv.org/abs/2511.12573.

Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5496–5505. PMLR, 18–24 Jul 2021.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29.

Ashwin Kumar, Yuzi He, Aram H Markosyan, Bobbie Chern, and Imanol Arrieta-Ibarra. Detecting prefix bias in llm-based reward models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 3196–3206, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3715275.3732204.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: regularizing RLHF with KL divergence does not mitigate heavy-tailed reward misspecification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=UXuBzWoZGK.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

Nathan Lambert. Reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2504.12501.

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*, pages 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.96.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Gengxu Li, Tingyu Xia, Yi Chang, and Yuan Wu. Length-controlled margin-based preference optimization without reference model, 2025a. URL https://arxiv.org/abs/2502.14643.

Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. Does style matter? Disentangling style and substance in chatbot arena. Blog post, 2024. Accessed: 2026-05-06.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zhuo Li, Pengyu Cheng, Zhechao Yu, Feifei Tong, Anningzhe Gao, Tsung-Hui Chang, Xiang Wan, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Eliminating inductive bias in reward models with information-theoretic guidance, 2025b. URL https://arxiv.org/abs/2512.23461.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy, 2025. URL https://arxiv.org/abs/2507.01352. arXiv:2507.01352.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. Robust optimization for mitigating reward hacking with correlated proxies. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O3shkBWM2s.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. Eliminating biased length reliance of direct preference optimization via down-sampled KL divergence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1047–1067. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.emnlp-main.60.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fb0G86Dewb.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: simple preference optimization with a reference-free reward. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Curran Associates Inc., 2024.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. What's in my human feedback? learning interpretable descriptions of preference data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sC6A1bFDUt.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 278–287. Morgan Kaufmann Publishers Inc., 1999.

Ignavier Ng, Patrick Blöbaum, Siddharth Bhandari, Kun Zhang, and Shiva Kasiviswanathan. Debiasing reward models by representation learning with guarantees, 2025. URL https://arxiv.org/abs/2510.23751.

Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski,

Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. Nemotron-4 340b technical report, 2024. URL https://arxiv.org/abs/2406.11704.

OpenAssistant. OpenAssistant/reward-model-deberta-v3-large-v2. https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2, 2023. Hugging Face model card. Accessed: 2026-01-06.

Henry Papadatos and Rachel Freedman. Linear probe penalties reduce LLM sycophancy. In *Workshop on Socially Responsible Language Modelling Research*, 2024. URL https://openreview.net/forum?id=6N2yES22rG.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.297.

Judea Pearl and Elias Bareinboim. *External Validity: From Do-Calculus to Transportability Across Populations*, page 451–482. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. URL https://doi.org/10.1145/3501714.3501741.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 13–18 Jul 2020.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.findings-acl.847.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 745–750. Association for Computing Machinery, 2007. doi: 10.1145/1273496.1273590.

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 *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=magEgFpK1y.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How rlhf amplifies sycophancy, 2026. URL https://arxiv.org/abs/2602.01002.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=qq6ctdUwCX.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022.

Joar 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*, ICML'23. JMLR.org, 2023.

Ruike Song, Zeen Song, Huijie Guo, and Wenwen Qiang. Causal reward adjustment: Mitigating reward hacking in external reasoning via backdoor correction, 2025. URL https://arxiv.org/abs/2508.04216.

Pragya Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuveer, Karthikeyan Shanmugam, and Doina Precup. Robust reward modeling via causal rubrics. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oP99JQiDYp.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Curran Associates Inc., 2020.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S. Brown. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=R0Xxvr_X3ZA.

Muhammad Umer, Muhammad Ahmed Mohsin, Ahsan Bilal, Arslan Chaudhry, Andreas Haupt, Sanmi Koyejo, Emily Fox, and John M. Cioffi. General preference reinforcement learning, 2026. URL https://arxiv.org/abs/2605.18721.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, and Sinong Wang. Beyond reward hacking: Causal rewards for large language model alignment, 2025. URL https://arxiv.org/abs/2501.09620.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384. Association for Computational Linguistics, June 2024. doi: 10.18653/v1/2024.naacl-long.185.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=tLZZZIgPJX.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xJljiPE6dg.

Yuchen Yan, Jin Jiang, Zhenbang Ren, Yijun Li, Xudong Cai, Yang Liu, Xin Xu, Mengdi Zhang, Jian Shao, Yongliang Shen, Jun Xiao, and Yueting Zhuang. Verifybench: Benchmarking reference-based reward systems for large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JfsjGmuFxz.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=3GTtZFiajM.

Wenqian Ye, Guangtao Zheng, and Aidong Zhang. Rectifying shortcut behaviors in preference-based reward learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=m51t6RKfGH.

Yaowen Ye, Cassidy Laidlaw, and Jacob Steinhardt. Iterative label refinement matters more than preference optimization under weak supervision. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=q5EZ7gKcnW.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=rbBtoVnduo.

Danlong Yuan, Tian Xie, Shaohan Huang, Zhuocheng Gong, Huishuai Zhang, Chong Luo, Furu Wei, and Dongyan Zhao. Shorten after you're right: Lazy length penalties for reasoning rl, 2025. URL https://arxiv.org/abs/2505.12284.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. Demystify verbosity compensation behavior of large language models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 160–178, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.uncertainlp-main.14.

Kangwen Zhao, Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Dongyun Xue, Wengang Zhou, Li Li, and Houqiang Li. Bias fitting to mitigate length bias of reward model in rlhf, 2025. URL https://arxiv.org/abs/2505.12843.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc., 2023.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.198.

A Additional Formalizations

A.1 Causal Interpretation and Scope

In this appendix, we outline the causal frame of our formalization in Section 3 and delimit what the framework does and does not claim about the underlying causal structure of reward.

Underlying structural causal model. We posit, descriptively, that the true reward R and the features Φ are related through an underlying structural causal model $\mathcal{M}$ on (X, Y, R, Φ). We do not assume that $\mathcal{M}$ is identified from preference data $\mathcal{D}$. The reward identifiability characterization of Skalse et al. [2023] establishes that preference data identifies R only up to a prompt-only additive shift under the Bradley-Terry likelihood, leaving the cardinal scale of R and by extension the causal structure of $\mathcal{M}$ partially identified at best. The framework of Section 3 therefore operates within this partial-identification regime and it adopts causal vocabulary to state failure modes precisely, but its definitions and results do not require $\mathcal{M}$ to be recovered from $\mathcal{D}$.

Causal reading of Definition 3.3. The corresponding causal reading of Definition 3.3 is: ϕ_i is *causally spurious* with respect to R iff ϕ_i has no directed path to R in $\mathcal{M}$. Under positivity-style richness on the support of μ_{diag} (strengthening Assumption A.2) and faithfulness, the causal and observational readings coincide on the partition Φ_{sp} vs. Φ_{struct}. The causal reading is the primitive notion and the observational form its identifiable shadow. Thus, Φ_{sp} vs. Φ_{struct} is best read as the conservative observational projection of an underlying causal partition preference data does not fully identify [Skalse et al., 2023].

Partial informativeness as causal mediation. Partial informativeness (Sections A.3 and 3.1) corresponds causally to ϕ_i having both a direct path to R in $\mathcal{M}$ and one or more mediated paths through other features. Length is the canonical instance: a direct contribution via comprehensiveness, plus mediated effects of hedging, sycophancy, and verbosity compensation (see Sections C and 4). The conservative partition classifies such features as Φ_{sp}, under-flagging rotation onto causally relevant mediated paths.

Associational mitigation operators. The single-axis mitigations M_i of Definition 3.5 are associational: they target the linear reliance statistic g_i, which is an $L^2(\mu_{\text{diag}})$ regression coefficient, not a controlled direct effect. A causal counterpart would target the controlled direct effect of ϕ_i on R, identifiable only under interventions on $\mathcal{M}$ or strong conditional-independence assumptions that preference data does not warrant. Existing single-axis mitigations in the language reward modeling literature [e.g., Fein et al., 2026, Papadatos and Freedman, 2024, Huang et al., 2025] are associational in this sense, and the regime taxonomy of Section 3.3 characterizes their failure modes precisely because of the gap between associational construction and causal response of the optimized policy.

R2 origins in causal language. The two origins of R2 in Definition 3.7 restate causally as: scale overshoot (projection coefficient too large at $\mu_{\pi^\star}$ while ϕ_i remains causally spurious) versus target misspecification (ϕ_i has a non-zero direct path to R in $\mathcal{M}$ that the projection has zeroed at μ_{diag}). The rescalability test of Section A.13 discriminates these origins observationally, without identifying $\mathcal{M}$.

R4 and transportability. The audit-distribution sensitivity of Definition 3.8 is structurally analogous to the transportability question [e.g. Pearl and Bareinboim, 2022]. An identified mitigation at one audit distribution does not need to transport to another. R4 names this phenomenon under partial identification, without formally invoking transportability machinery.

A.2 Standing Assumptions and Optimization Setup

Assumption A.1 (Regularity)**.** *$R, \tilde{R} \in L^2(\mu_{\text{diag}}) \cap L^\infty(\mu_{\pi_{\text{ref}}})$ and $\phi_k \in L^2(\mu_{\text{diag}}) \cap L^\infty(\mu_{\pi_{\text{ref}}})$ for each $k = 1, \ldots, K$, where $\Phi = (\phi_1, \ldots, \phi_K)^\top$ is the feature map of Definition 3.1.*

The $L^2(\mu_{\text{diag}})$ inclusion makes $g(\cdot; \mu_{\text{diag}})$ of Definition 3.4 well-defined as a Gram-inner-product statistic at the diagnostic measure. The $L^\infty(\mu_{\pi_{\text{ref}}})$ conditions are natural both for bounded-output reward models and for the interpretable surface features this paper considers (length, formatting

counts, style indicators). The conditions imply that any $\bar{R}$ obtained as a finite $\mathbb{R}$-linear combination of $\tilde{R}$, R, and the ϕ_k also lies in $L^\infty(\mu_{\pi_{\text{ref}}})$, which covers every $\bar{R}$ used in this paper, including $M_i(\tilde{R})$ of Definition 3.5 and the scale-normalised variant M_i^{norm} introduced below. Since $\mu_{\pi_{\text{ref}}}$ is a probability measure, $L^\infty(\mu_{\pi_{\text{ref}}}) \subset L^2(\mu_{\pi_{\text{ref}}})$ holds automatically and no separate $L^2(\mu_{\pi_{\text{ref}}})$ clause is needed.

Optimization setup. Fix $\beta > 0$. For any $\bar{R}$ satisfying Assumption A.1, define the KL-regularized optimizer

$$\pi^\star_\beta(\bar{R}) \;=\; \arg\max_{\pi}\; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi}\big[\bar{R}(x,y)\big] \;-\; \beta\, D_{\text{KL}}\big(\pi \,\|\, \pi_{\text{ref}}\big),$$

which in the single-turn contextual-bandit regime admits the unique softmax closed form $\pi^\star_\beta(\bar{R})(y \mid x) \propto \pi_{\text{ref}}(y \mid x)\, \exp(\bar{R}(x,y)/\beta)$ recalled in the main text. Essential boundedness of $\bar{R}$ gives two-sided bounds on both $\exp(\bar{R}/\beta)$ and the normalizer $Z(x) = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[\exp(\bar{R}(x,y)/\beta)]$, so the density ratio $d\mu_{\pi^\star_\beta(\bar{R})}/d\mu_{\pi_{\text{ref}}} = \exp(\bar{R}/\beta)/Z(x)$ is essentially bounded. The chain $\bar{R} \in L^\infty(\mu_{\pi_{\text{ref}}}) \Rightarrow L^\infty(\mu_{\pi^\star_\beta(\bar{R})}) \subset L^2(\mu_{\pi^\star_\beta(\bar{R})})$ then follows immediately, and analogously for each R and each ϕ_k. The policy-level expectations $\mathbb{E}_{\mu_{\pi^\star}}[\phi_j]$ entering Δ_j in Definition 3.7 are then finite. Well-definedness of $g(\bar{R}; \mu_{\pi^\star_\beta(\bar{R})})$ additionally requires Gram non-degeneracy at $\mu_{\pi^\star}$. This condition is handled in Section A.3, where the relevant audit distributions are introduced.

A.3 Feature Realizability and the Diagnostic Measure

This section collects four items referenced from Section 3. A discussion of the choices available for μ_{diag} (pointed to from the preamble of Section 3), the Gram non-degeneracy condition for the decoupled case, the structural Assumption A.2 (feature realizability) on the feature map (invoked by Definition 3.3), and the expressiveness remark that motivates reading the regime language of Section 3.3 conservatively under partially informative features (pointed to immediately after Definition 3.3). The order below matches the narrative order of Section 3.

Remark (choice of diagnostic measure). The main text introduces μ_{diag} as an auxiliary distribution on $\mathcal{X} \times \mathcal{Y}$ used for reliance estimation and correlation measurement, distinct from the KL anchor π_{ref}. Two choices are load-bearing for this paper:

1. *The coupled case* $\mu_{\text{diag}} = \mu_{\pi_{\text{ref}}}$. Recovers the setting of Laidlaw et al. [2025] and makes diagnostic statistics directly comparable to their bounds, at the cost of coupling measurement and optimization reference points.
2. *An annotator-conditioned distribution* $\mu^{\text{human}}_{\text{diag}}$ *or* $\mu^{\text{LLM}}_{\text{diag}}$, reflecting that the same proxy $\tilde{R}$ induces different g-profiles depending on which annotator pool generated the audit data. This formalizes the human-vs-LLM-judge gap documented in Saito et al. [2023], Zheng et al. [2023], Chen et al. [2024a], Movva et al. [2026], and is what R4 (Definition 3.8) operationalizes.

Other valid choices, e.g., the preference-data distribution used to train $\tilde{R}$, curated audit sets, deployment distributions, and round-anchored distributions in iterative RLHF [Ye et al., 2025c], yield different g-profiles and shift substitution diagnostics accordingly. Substitution claims are always relative to the chosen diagnostic measure, and the choice should be stated explicitly in applications.

Where $g(\bar{R}; \mu_{\pi^\star_\beta(\bar{R})})$ is referenced in Section 3 (e.g., the measurement-vs-optimization gap following Lemma 3.6), well-definedness requires Gram non-degeneracy at $\mu_{\pi^\star}$, which follows from Assumption 3.2 (non-degeneracy) whenever $d\mu_{\pi^\star}/d\mu_{\text{diag}}$ is bounded above and below. In the coupled case ($\mu_{\text{diag}} = \mu_{\pi_{\text{ref}}}$), the bound follows from Assumption A.1 (regularity). In the decoupled case ($\mu^{\text{human}}_{\text{diag}}, \mu^{\text{LLM}}_{\text{diag}}$), it is a separate regularity condition assumed in the relevant statements.

Assumption A.2 (Feature Realizability)**.** *For μ_{diag}-almost every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and every feature index $i \in \{1,\dots,K\}$, there exists at least one $y' \in \mathcal{Y}$ with $\phi_j(x,y') = \phi_j(x,y)$ for all $j \neq i$ and $\phi_i(x,y') \neq \phi_i(x,y)$.*

Assumption A.2 is the non-vacuity condition for Definition 3.3. Without it, ϕ_i can be functionally determined by the prompt and the other features on $\mathcal{Y}$, so further conditioning on ϕ_i leaves the conditional expectation in Definition 3.3 unchanged and every feature is trivially classified as spurious. Assumption A.2 is therefore a richness condition on the response space *under* μ_{diag}, as $\mathcal{Y}$ must be

rich enough where the diagnostic measure places mass that each feature axis admits independent perturbation.

We treat g and Δ_j as observational statistics under μ_{diag} and $\mu_{\pi^\star}$ throughout, as an interventional reading along ϕ_i would require strengthening Assumption A.2 to a richness condition on counterfactual realization, which we do not invoke.

Even after weakening to the μ_{diag}-a.e. form, Assumption A.2 is fragile for natural-language features on Φ's own terms. Consider $\Phi = \{\phi_{\text{length}}, \phi_{\text{exclamation\_count}}\}$. Assumption A.2 demands that μ_{diag}-a.e. there exist a response y' matching y on exclamation count but differing in length, and conversely. The first direction is plausible, as long and short variants with the same exclamation count are typically realizable. The second is genuinely constrained, as at fixed length, the realizable range of exclamation counts is bounded above by length itself, and at very short lengths the counterfactual set may be empty. Even within Φ, the response space does not cleanly factor into independent feature axes. This failure mode is specific to features that count items within the response (exclamation marks, bullet points, headers), since their range is mechanically tied to length. Features operationalized as binary indicators or response-level scalars do not have this bound (e.g., sycophancy and confidence in Sections B.3 and 4), and Assumption A.2 holds approximately for these pairs.

Feature misspecification is itself a major failure mode in practice. Under Assumption A.2, Definition 3.3 cleanly separates spurious from structurally relevant features, and the canonical mitigation M_i of Definition 3.5 targets a well-defined coordinate of g-space. When Assumption A.2 fails, the definitions of Section 3.1 should be read as approximations, and the binary spurious/structurally-relevant partition as the best dichotomous summary of a genuinely continuous informativeness spectrum.

Remark (scope of the binary partition under partial informativeness). Definition 3.3 partitions Φ into spurious and structurally relevant features. The partition is scope-matched to the operator class deployed in the literature (whole-feature projections, see Section 3.1) and to what preference data identifies, but it does not separate within-feature components for partially informative features in the empirical setting surveyed in Section 4, where many natural features (foremost length) act as mediators for multiple mechanisms (sycophancy, epistemic uncertainty, informativeness) rather than as pure spurious proxies or pure structural drivers.

Under partial informativeness, the regime distinction of Section 3.3 still applies, but with a scoping caveat: bias substitution becomes a *conservative classification* of the ways a single-axis mitigation can fail, because real features admit additional substitution modes that the binary partition cannot express. Concretely, the best $L^2(\mu_{\text{diag}})$ approximation of ϕ_i by a function of R alone (writing $\phi_i = \mathbb{E}_{\mu_{\text{diag}}}[\phi_i \mid R] + \phi_i^{\perp}$, i.e., decomposing ϕ_i into "the part that tracks R" and a residual) does not in general coincide with the causal-versus-spurious decomposition of ϕ_i under the underlying structural causal model $\mathcal{M}$ (Section A.1). The first is a μ_{diag}-level regression decomposition, the second a structural statement under $\mathcal{M}$. Reconciling the two requires additional structure, e.g., a causal model, conditional-independence assumptions, or auxiliary interventions, and we defer partial-informativeness extensions to future work.

Exhaustiveness of the regime taxonomy. The two axes underlying Section 3.3 — whether mitigation rotates first-moment exploitation onto another spurious feature ($\Delta_j(\pi, \pi') \neq 0$ for some $\phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}$) and the sign of the change in true reward ($\Delta J = J(\pi', R) - J(\pi, R)$) — partition the outcome space into six cells, summarized in Definition 3.7. Five regime labels (R0, $\mathrm{R0}_{\text{cont}}$, R1, R2, R3) suffice because R1 absorbs both $\Delta J = 0$ (neutral substitution) and $\Delta J < 0$ (harmful substitution) under Definition 3.7, while R0 and $\mathrm{R0}_{\text{cont}}$ receive separate labels because the rotation-with-improvement corner is precisely the case audit-distribution evaluation most readily mistakes for clean R0 (cf. Theorem 3.9). The asymmetry is therefore deliberate rather than ad hoc: $\mathrm{R0}_{\mathrm{cont}}$ earns a separate label for downstream emphasis in the impossibility result, whereas neutral and harmful R1 are both already failures of ΔJ and the sub-case treatment suffices. The R4 axis (Definition 3.8) does not appear in Definition 3.7 because R4 is not an outcome cell but a property of how cell membership shifts across μ_{diag} choices, as discussed in Section 3.3 and instantiated empirically in Section B.3.

Spurious-rotation-with-improvement corner. R1 in Definition 3.7 (bias substitution) requires $\Delta J \leq 0$. The complementary corner $\Delta_j \neq 0$ for some $\phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}$ with $\Delta J > 0$ is regime $\mathrm{R0}_{\text{cont}}$ per Definition 3.7. Two mechanisms produce $\mathrm{R0}_{\text{cont}}$: (i) rotation onto a feature classified spurious in the conservative partition but partially informative in the true reward (an in-scope manifestation of

the conservatism discussed above), and (ii) under the strict binary partition, rotation onto a genuinely spurious ϕ_j where structural-feature gains from reducing ϕ_i outweigh the spurious cost on ϕ_j.

The practical consequence for the rest of the paper is that substitution claims proved under Assumption A.2 (feature realizability) should be read as a floor rather than a complete characterization: when Assumption A.2 fails, at least the substitution mode the binary partition captures remains available to the optimizing policy, and generically additional modes are available too.

A.4 Equivalent Epsilon-Formulation for Local Spuriousness

Definition 3.3 states spuriousness as conditional mean independence at μ_{diag}. We derive the equivalent $o(\varepsilon)$ formulation for completeness.

Let $\mathcal{Y}_{-i}(x,y) := \{y' : \phi_j(x,y') = \phi_j(x,y)\ \forall j \neq i\}$ denote the i-fiber through (x,y), and decompose $\mu_{\mathrm{diag}} = D(x)\,\pi_{\mathrm{audit}}(y \mid x)$. For a Markov kernel κ with $\kappa(\mathcal{Y}_{-i}(x,y) \mid x,y) = 1$ for μ_{diag}-a.e. (x,y), define the mixture-perturbed conditional

$$\pi_\varepsilon^\kappa(y' \mid x) \;:=\; (1-\varepsilon)\,\pi_{\mathrm{audit}}(y' \mid x) \;+\; \varepsilon \int \kappa(y' \mid x,y)\,\pi_{\mathrm{audit}}(dy \mid x), \qquad \varepsilon \in [0,1]. \tag{7}$$

Consider the perturbation condition

$$\mathbb{E}_{x\sim D,\, y'\sim\pi_\varepsilon^\kappa(\cdot|x)}\big[R(x,y')\big] \;-\; \mathbb{E}_{\mu_{\mathrm{diag}}}\big[R(x,y)\big] \;\in\; o(\varepsilon) \quad \text{as } \varepsilon \downarrow 0. \tag{8}$$

The left-hand side equals $\varepsilon\cdot\big(\mathbb{E}_{\kappa\otimes\mu_{\mathrm{diag}}}[R] - \mathbb{E}_{\mu_{\mathrm{diag}}}[R]\big)$ for mixture perturbations, so the $o(\varepsilon)$ condition is equivalent to $\mathbb{E}_{\kappa\otimes\mu_{\mathrm{diag}}}[R] = \mathbb{E}_{\mu_{\mathrm{diag}}}[R]$. Requiring this for every kernel κ implies the conditional mean independence of Definition 3.3, and the two are equivalent when R is $\sigma(\Phi)$-measurable (so that within-fiber variation of R collapses to variation in ϕ_i).

Verification of downstream results. Theorem 3.9 and the closed-form regime instantiations in Sections A.7 and A.9 use a true reward of the form $R = w\phi_3$, which is $\sigma(\Phi)$-measurable. Thus, the conditional mean independence and strict point-wise invariance on i-fibers coincide in this regime. The constructed instances satisfy Definition 3.3 and the proofs transfer without modification. Theorem 3.10 is a correctness statement on the classifier $B^\star$ given $(\Delta_j, \Delta J)$ and Φ_{sp} as inputs. It places no assumption on the functional form of R and transfers regardless of how Φ_{sp} is defined.

A.5 Scale Change Induced by Mitigation and a Scale-Invariant Variant

Immediately after Lemma 3.6, we note that applying M_i changes $\|\tilde{R}\|_{L^2(\mu_{\mathrm{diag}})}$ through both a diagonal and a cross-correlation contribution, and that this scale change interacts non-trivially with the KL-regularized optimizer because $\pi_\beta^\star$ is not invariant to positive rescaling of the reward. In this section, we give the exact scale identity for M_i, isolate the mechanism by which scale change conflates with axis reallocation at fixed β, and define the scale-invariant variant M_i^{norm} that separates the two effects.

Scale identity. Recall the Gram matrix G of Assumption 3.2, with entries $G_{ij} = \mathbb{E}_{\mu_{\mathrm{diag}}}[\phi_i\phi_j]$, let $g = g(\tilde{R};\mu_{\mathrm{diag}})$ abbreviate the linear-reliance vector of Definition 3.4, and write $\tilde{R}' = M_i(\tilde{R}) = \tilde{R} - g_i\phi_i$ for the canonical mitigation of Definition 3.5. Expanding the $L^2(\mu_{\mathrm{diag}})$-norm of $\tilde{R}'$,

$$\|\tilde{R}'\|^2_{L^2(\mu_{\mathrm{diag}})} \;=\; \|\tilde{R}\|^2_{L^2(\mu_{\mathrm{diag}})} \;-\; 2g_i\,\mathbb{E}_{\mu_{\mathrm{diag}}}[\phi_i\tilde{R}] \;+\; g_i^2\,G_{ii},$$

and using $\mathbb{E}_{\mu_{\mathrm{diag}}}[\phi_i\tilde{R}] = (Gg)_i = \sum_j G_{ij}g_j$ gives the exact identity

$$\|\tilde{R}'\|^2_{L^2(\mu_{\mathrm{diag}})} \;=\; \|\tilde{R}\|^2_{L^2(\mu_{\mathrm{diag}})} \;-\; g_i\left[\, g_i\,G_{ii} \;+\; 2\sum_{j\neq i} G_{ij}\,g_j \,\right].$$

The sign of the change is governed by the bracketed expression relative to g_i: the norm strictly decreases whenever the bracketed term is co-signed with g_i, and can *increase* when the cross-correlation contribution $2\sum_{j\neq i} G_{ij}g_j$ is oppositely signed and sufficiently large in magnitude relative to the diagonal term g_iG_{ii}. A parameter choice exhibiting this norm increase is given in Section A.7.

The diagonal-only approximation $\|\tilde{R}'\|^2 \approx \|\tilde{R}\|^2 - g_i^2\, G_{ii}$ is correct only when G is diagonal at μ_{diag}, i.e., when the diagnostic measure renders the feature axes orthogonal. This condition generically fails for interpretable feature sets such as length, formatting, and style indicators, which are empirically correlated on any natural μ_{diag}.

Scale-invariant variant. The $L^2(\mu_{\mathrm{diag}})$-normalized mitigation

$$M_i^{\mathrm{norm}}(\tilde{R}) \;=\; \alpha\,\big(\tilde{R} - g_i\phi_i\big), \qquad \alpha := \frac{\|\tilde{R}\|_{L^2(\mu_{\mathrm{diag}})}}{\|\tilde{R} - g_i\phi_i\|_{L^2(\mu_{\mathrm{diag}})}},$$

preserves $\|\cdot\|_{L^2(\mu_{\mathrm{diag}})}$ by construction. Its linear reliance at μ_{diag} is α times that of $M_i(\tilde{R})$, so $g_i(M_i^{\mathrm{norm}}(\tilde{R}); \mu_{\mathrm{diag}}) = 0$ and $g_j(M_i^{\mathrm{norm}}(\tilde{R}); \mu_{\mathrm{diag}}) = \alpha\, g_j(\tilde{R}; \mu_{\mathrm{diag}})$ for $j \neq i$: the i-th coordinate is zeroed and the remaining axes are rescaled uniformly.

Effective-β shift. By KL scale-equivariance $\pi^\star_\beta(c\bar{R}) = \pi^\star_{\beta/c}(\bar{R})$ and the relation $M_i^{\mathrm{norm}}(\tilde{R}) = \alpha\, M_i(\tilde{R})$,

$$\pi^\star_\beta\big(M_i(\tilde{R})\big) \;=\; \pi^\star_{\alpha\beta}\big(M_i^{\mathrm{norm}}(\tilde{R})\big).$$

The unnormalized comparison $\pi^\star_\beta(\tilde{R})$ vs $\pi^\star_\beta(M_i(\tilde{R}))$ at fixed β therefore equals the normalized comparison $\pi^\star_\beta(\tilde{R})$ vs $\pi^\star_{\alpha\beta}(M_i^{\mathrm{norm}}(\tilde{R}))$, differing from the same-β normalized comparison by exactly an effective-β shift from β to $\alpha\beta$. This is a cardinal-distortion mechanism rather than an ordinal one, as two proxies producing identical pairwise rankings on μ_{diag} can differ in $L^2(\mu_{\mathrm{diag}})$-norm and induce different $\pi^\star_\beta$ at fixed β. This mechanism is in line with the Section 2 ordinal-vs-cardinal gap, which motivates reporting $\|\tilde{R}\|_{L^2(\mu_{\mathrm{diag}})}$ alongside Δ_j in empirical instantiations.

As a practical consequence, under M_i^{norm}, any observed change in $\mathbb{E}_{\pi^\star_\beta}[\phi_j]$ at fixed β is attributable to axis reallocation rather than scale, and Δ_j values can differ in sign between M_i and M_i^{norm} when α is far from 1, classifying the same proxy into different R0-R3 regimes under the two mitigations. Diagnosing R1 versus R3 cleanly therefore calls for M_i^{norm}, since under M_i a nonzero Δ_j admits both reallocation and scale interpretations and requires a β-scan or pre/post norm reporting to disambiguate.

Deployment-time normalization. Standard practices in PPO-based RLHF for stability like per-batch reward whitening [Lambert, 2025] apply a sample-dependent shift and scale to $\tilde{R}$ at each training step. In expectation the shift approximates a constant under the current rollout distribution and is therefore gauge-equivalent at the population level (no policy effect, by Section 3.2), but the scale component divides $\tilde{R}$ by an estimate of $\mathrm{Std}_{\mu_{\pi_t}}(\tilde{R})$, inducing an effective KL coefficient $\beta \cdot \mathrm{Std}_{\mu_{\pi_t}}(\tilde{R})$ that M_i^{norm} does not control. M_i^{norm} fixes the audit-side $L^2(\mu_{\mathrm{diag}})$ scale, while whitening rescales against the (non-stationary) training-side dispersion. The $(M_i, M_i^{\mathrm{norm}})$ comparison therefore does not exhaust the scale story under whitened RLHF, and empirical instantiations should report training-side reward dispersion alongside $\|\tilde{R}\|_{L^2(\mu_{\mathrm{diag}})}$ when interpreting Δ_j across mitigations.

A.6 Gauge-Invariant Linear Reliance

In this section, we give a gauge-invariant variant of the linear reliance statistic and verify that the Section 3.3 regime classification transfers under it. Define the prompt-conditional deviations under μ_{diag},

$$\tilde{R}^{\mathrm{cent}}(x,y) = \tilde{R}(x,y) - \mathbb{E}_{\mu_{\mathrm{diag}}}[\tilde{R} \mid x], \qquad \bar{\Phi}(x,y) = \Phi(x,y) - \mathbb{E}_{\mu_{\mathrm{diag}}}[\Phi \mid x].$$

Assumption A.3 (Non-degeneracy)**.** *The centered Gram matrix $\bar{G} := \mathbb{E}_{\mu_{\mathrm{diag}}}[\bar{\Phi}\bar{\Phi}^\top] \in \mathbb{R}^{K\times K}$ is positive definite: $\bar{G} \succ 0$.*

Note that Assumption A.3 does not follow from Assumption 3.2, as any feature that is constant in y given x is non-degenerate pooled but vanishes after within-prompt centering. The *centered linear reliance* is

$$g_{\mathrm{cent}}(\tilde{R};\, \mu_{\mathrm{diag}}) = \big(\mathbb{E}_{\mu_{\mathrm{diag}}}[\bar{\Phi}\bar{\Phi}^\top]\big)^{-1}\mathbb{E}_{\mu_{\mathrm{diag}}}[\bar{\Phi}\,\tilde{R}^{\mathrm{cent}}] \;\in\; \mathbb{R}^K.$$

For near-collinear or learned feature sets, g_{cent} should be read as a ridge-regularized estimate.

Gauge invariance. Under $\tilde{R} \mapsto \tilde{R} + b(x)$ for any prompt-only b, $\mathbb{E}[\tilde{R} + b \mid x] = \mathbb{E}[\tilde{R} \mid x] + b(x)$, so $\tilde{R}^{\text{cent}}$ is unchanged and $\bar{\Phi}$ does not involve $\tilde{R}$. Therefore $g_{\text{cent}}(\tilde{R} + b;\, \mu_{\text{diag}}) = g_{\text{cent}}(\tilde{R};\, \mu_{\text{diag}})$. The KL-regularized optimum shares this invariance, so g_{cent} is a reliance statistic on the same gauge equivalence class the policy respects.

Relation to g. The pooled second moments decompose as

$$\mathbb{E}[\Phi\Phi^\top] = \mathbb{E}[\bar{\Phi}\bar{\Phi}^\top] + \mathbb{E}\big[\mathbb{E}[\Phi|x]\mathbb{E}[\Phi|x]^\top\big], \qquad \mathbb{E}[\Phi\tilde{R}] = \mathbb{E}[\bar{\Phi}\,\tilde{R}^{\text{cent}}] + \mathbb{E}\big[\mathbb{E}[\Phi|x]\mathbb{E}[\tilde{R}|x]\big]$$

so g is a matrix-weighted combination of the within-prompt regression (g_{cent}) and the between-prompt regression of conditional means. The two statistics coincide only in degenerate cases (e.g., vanishing between-prompt covariation) and can differ substantially when Φ carries strong prompt-level structure. This structure is plausible for length, where prompt difficulty drives average response length.

Canonical mitigation and Lemma 3.6 analogue. The centered canonical mitigation is

$$M_i^{\text{cent}}(\tilde{R})(x, y) = \tilde{R}(x, y) - g_{\text{cent},i}(\tilde{R}; \mu_{\text{diag}})\, \phi_i(x, y).$$

Lemma A.4 (Centered single-axis identity at μ_{diag})**.** *For the centered canonical mitigation M_i^{cent}, $g_{cent,i}(M_i^{cent}(\tilde{R}); \mu_{diag}) = 0$ and $g_{cent,j}(M_i^{cent}(\tilde{R}); \mu_{diag}) = g_{cent,j}(\tilde{R}; \mu_{diag})$ for $j \neq i$.*

Proof. Direct computation in the centered inner product, using that $\mathbb{E}_{\mu_{\text{diag}}}[\bar{\Phi}\bar{\phi}_i]$ is the i-th column of $\mathbb{E}_{\mu_{\text{diag}}}[\bar{\Phi}\bar{\Phi}^\top]$. □

The variant $\tilde{R} - g_{\text{cent},i}\bar{\phi}_i$ differs from $M_i^{\text{cent}}(\tilde{R})$ by $g_{\text{cent},i}\,\mathbb{E}[\phi_i \mid x]$, a prompt-only function. Given the above-stated gauge invariance, the two variants are policy-equivalent and produce identical g_{cent}.

Framework transfer. Definition 3.4 ports to g_{cent} verbatim, and Definition 3.5 ports as the requirement $|g_{\text{cent},i}(\tilde{R}'; \mu_{\text{diag}})| < |g_{\text{cent},i}(\tilde{R}; \mu_{\text{diag}})|$. Lemma A.4 shows M_i^{cent} qualifies. Definition 3.7 (R0-R3) port unchanged because their criteria reference policy-induced expectations $\Delta_j(\pi, \pi')$ and ΔJ, not the reliance statistic.

Gauge invariance of the regime classification. The Section 3.2 obstruction is that under $\tilde{R} \mapsto \tilde{R} + b(x)$ for prompt-only b, $M_i(\tilde{R} + b)$ differs from $M_i(\tilde{R}) + b$ by $(g_i(\tilde{R}) - g_i(\tilde{R} + b))\phi_i$, so $\pi_\beta^\star(M_i(\tilde{R} + b)) \neq \pi_\beta^\star(M_i(\tilde{R}))$ and $\Delta_j, \Delta J$ can change. Replacing M_i by M_i^{cent} removes the obstruction. By gauge invariance of g_{cent}, $g_{\text{cent},i}(\tilde{R}+b) = g_{\text{cent},i}(\tilde{R})$, so $M_i^{\text{cent}}(\tilde{R}+b) = M_i^{\text{cent}}(\tilde{R})+b$. Since the KL-regularized optimum is invariant under prompt-only shifts,

$$\pi_\beta^\star(\tilde{R} + b) = \pi_\beta^\star(\tilde{R}), \qquad \pi_\beta^\star(M_i^{\text{cent}}(\tilde{R} + b)) = \pi_\beta^\star(M_i^{\text{cent}}(\tilde{R})),$$

hence $\Delta_j(\pi, \pi')$ and ΔJ are gauge-invariant and the R0-R3 regime is unchanged. R4 (see Definition 3.8) also ports verbatim, as g_{cent} retains its μ_{diag}-dependence and only the gauge of $\tilde{R}$ is factored out. Thus, $M_i^{\text{cent},\mu_{\text{diag}}}$ depends on μ_{diag} in the same sense as $M_i^{\mu_{\text{diag}}}$ in Definition 3.8 and audit-distribution sensitivity is defined identically.

Note that M_i and M_i^{cent} applied to the same proxy $\tilde{R}$ may still occupy different R0-R3 regimes, as different mitigations induce different policies. However, this difference reflects the choice of reliance estimator as a degree of freedom in the mitigation pipeline, independent of gauge. Auditing a proxy under both statistics localizes the targeted reliance, as a feature on which $|g_i|$ is large but $|g_{\text{cent},i}|$ is small flags reliance that lives in prompt-level structure the optimizing policy averages over rather than acts on.

A.7 Non-Vacuity of the Regime Taxonomy via Linear-Gaussian Instantiation

This appendix exhibits closed-form instantiations of the regimes and gaps introduced in Section 3, establishing that the taxonomy is non-vacuous within our framework's own assumptions. The constructions are not intended as empirical models of RLHF and we show the empirical instantiations in Sections B.3 and 4.

The construction below satisfies Assumptions A.1 (regularity), 3.2 (non-degeneracy of the Gram at μ_{diag}), and A.2 (feature realizability), as well as Gram non-degeneracy at $\mu_{\pi^\star}$. We verify each below.

Shared construction. We work with a trivial prompt space ($\mathcal{X}$ a singleton, suppressed in notation) and continuous response space $\mathcal{Y} = \mathbb{R}^3$. The feature map is $\Phi(y) = (\phi_1, \phi_2, \phi_3)$ with $\phi_k(y) = y_k$. We fix true and proxy rewards

$$R(y) = w\,\phi_3(y), \qquad \tilde{R}(y) = a\,\phi_1(y) + b\,\phi_2(y) + w\,\phi_3(y),$$

with $w > 0$ and $(a, b) \in \mathbb{R}^2$. By Definition 3.3, $\Phi_{\mathrm{sp}} = \{\phi_1, \phi_2\}$ and $\Phi_{\mathrm{struct}} = \{\phi_3\}$ by construction. Reference policy and diagnostic measure are zero-mean Gaussians on $\mathbb{R}^3$,

$$\pi_{\mathrm{ref}} = \mathcal{N}(0, I_3), \qquad \mu_{\mathrm{diag}} = \mathcal{N}(0, \Sigma),$$

where Σ is positive definite with $\Sigma_{kk} = 1$ and off-diagonal $\Sigma_{12} = \rho \in (-1, 1)$, $\Sigma_{13} = \Sigma_{23} = 0$. Fix the KL parameter $\beta > 0$ and target the spurious feature $i = 1$ throughout. We treat $\mathbb{R}^3$ as the formal response space. For the $L^\infty(\mu_{\pi_{\mathrm{ref}}})$ clause of Assumption A.1, restrict to a sufficiently large bounded set on which all formulas below hold to arbitrary precision in the unrestricted limit.[5]

Verification of assumptions. Assumption A.1 holds under truncation as noted. Assumption 3.2 holds because $\mathbb{E}_{\mu_{\mathrm{diag}}}[\Phi\Phi^\top] = \Sigma \succ 0$. Assumption A.2 holds because $\mathcal{Y} = \mathbb{R}^3$ has full support under any positive-definite Gaussian, so for μ_{diag}-a.e. y and any i, varying y_i while holding $y_{j\neq i}$ fixed remains in the support. Gram non-degeneracy at $\mu_{\pi^\star}$ follows from the closed form below, where $\mu_{\pi^\star}$ is itself Gaussian with positive-definite covariance.

Closed forms for the optimum. For any linear reward $\bar{R}(y) = c^\top y$ with $c \in \mathbb{R}^3$, the KL-regularized optimum against π_{ref} satisfies $\pi^\star_\beta(\bar{R}) = \mathcal{N}(c/\beta, I_3)$ by completing the square in the softmax form of Equation (2). Two consequences used throughout: (i) the policy-induced first moment is $\mathbb{E}_{\mu_{\pi^\star}}[\phi_k] = c_k/\beta$, so $\Delta_j(\pi, \pi') = (c'_j - c_j)/\beta$ for any pair of linear rewards with coefficients c, c'; (ii) the plain return is $J(\pi^\star_\beta(\bar{R}), R) = \sum_k w_k c_k/\beta$ where w_k are the coefficients of R, so under our $R = w\phi_3$, only the ϕ_3-coefficient of the proxy contributes to ΔJ.

Linear reliance and canonical mitigation. The reliance vector at μ_{diag} is $g(\tilde{R}; \mu_{\mathrm{diag}}) = \Sigma^{-1}\mathbb{E}_{\mu_{\mathrm{diag}}}[\Phi\tilde{R}]$. Since $\tilde{R}$ is linear in Φ with coefficient vector (a, b, w) and $\mathbb{E}_{\mu_{\mathrm{diag}}}[\Phi\Phi^\top] = \Sigma$, we have $g(\tilde{R}; \mu_{\mathrm{diag}}) = (a, b, w)$ exactly. The canonical mitigation $M_1(\tilde{R})(y) = \tilde{R}(y) - a\,\phi_1(y) = b\,\phi_2 + w\,\phi_3$ has reliance $g(M_1(\tilde{R}); \mu_{\mathrm{diag}}) = (0, b, w)$ by Lemma 3.6.

Measurement-vs-optimization gap. We progress through three settings of $(\pi_{\mathrm{ref}}, \mu_{\mathrm{diag}})$ to isolate where the gap of Section 3.2 can manifest in this construction.

Isotropic reference and audit ($\pi_{\mathrm{ref}} = \mu_{\mathrm{diag}} = \mathcal{N}(0, I_3)$): the mitigated proxy induces $\pi' = \pi^\star_\beta(M_1(\tilde{R})) = \mathcal{N}((0, b, w)/\beta, I_3)$. The policy-side Gram is $\mathbb{E}_{\mu_{\pi'}}[\Phi\Phi^\top] = I_3 + (0, b, w)^\top(0, b, w)/\beta^2$, and

$$g_1(M_1(\tilde{R}); \mu_{\pi'}) = \left(\mathbb{E}_{\mu_{\pi'}}[\Phi\Phi^\top]^{-1}\,\mathbb{E}_{\mu_{\pi'}}[\Phi\, M_1(\tilde{R})]\right)_1 = 0,$$

since $\mathbb{E}_{\mu_{\pi'}}[\phi_1 M_1(\tilde{R})] = 0$ in this case.

Isotropic reference, correlated audit ($\pi_{\mathrm{ref}} = \mathcal{N}(0, I_3)$, $\mu_{\mathrm{diag}} = \mathcal{N}(0, \Sigma)$ with $\Sigma_{12} = \rho \neq 0$): the canonical mitigation is $M_1(\tilde{R}) = \tilde{R} - g_1(\tilde{R}; \mu_{\mathrm{diag}})\phi_1$ with $g_1(\tilde{R}; \mu_{\mathrm{diag}}) = a$ unchanged (since $\Sigma_{13} = 0$ keeps ϕ_3's contribution clean and the (ϕ_1, ϕ_2) block of Σ^{-1} recovers a for a linear proxy). Evaluating reliance at $\mu_{\pi'} = \mathcal{N}((0, b, w)/\beta, I_3)$ gives Gram $I_3 + (0, b, w)^\top(0, b, w)/\beta^2$, diagonal in the ϕ_1 row, so $g_1(M_1(\tilde{R}); \mu_{\pi'}) = 0$ here as well.

Coupled correlated case ($\pi_{\mathrm{ref}} = \mu_{\mathrm{diag}} = \mathcal{N}(0, \Sigma)$ with $\Sigma_{12} = \rho \neq 0$): the policy-side Gram now inherits the audit cross-correlation, $\mu_{\pi'} = \mathcal{N}(\Sigma c/\beta, \Sigma)$ for $c = (0, b, w)$, and one might expect a non-trivial g_1. It does not arise, for a structural reason: for any linear proxy $\bar{R} = c^\top y$ on $\Phi(y) = y$ and any non-degenerate measure $\mu = \mathcal{N}(m, V)$,

$$g(\bar{R}; \mu) = \left(\mathbb{E}_\mu[\Phi\Phi^\top]\right)^{-1}\mathbb{E}_\mu[\Phi\,\bar{R}] = (V + mm^\top)^{-1}(V + mm^\top)\,c = c,$$

[5] The closed-form expressions extend continuously to the unrestricted Gaussian limit.

so g recovers the coefficient vector exactly and is invariant to μ. Applied to $c = (0, b, w)$, this gives $g_1(M_1(\tilde{R}); \mu_{\pi'}) = 0$ in the coupled case as well, and the g-level gap is degenerate throughout the linear-Gaussian closed form.

The substantive gap manifests instead through the policy-induced first moment. In the coupled correlated case, $\mathbb{E}_{\mu_{\pi'}}[\Phi] = \Sigma c/\beta$, so

$$\mathbb{E}_{\mu_{\pi'}}[\phi_1] \;=\; (\Sigma c)_1/\beta \;=\; \rho\, b/\beta \;\neq\; 0$$

whenever $\rho \neq 0$ and $b \neq 0$, even though g_1 vanishes at every measure. The mechanism is the cross-correlation ρ: at μ_{diag} the projection orthogonalizes $M_1(\tilde{R})$ against ϕ_1 in the Gram-inner-product sense, but the policy under $M_1(\tilde{R})$ shifts the mean of (ϕ_2, ϕ_3) by $(b, w)/\beta$, and any ϕ_1-ϕ_2 coupling at μ_{diag} propagates into ϕ_1 first-moment drift at the policy distribution. This is the form of the gap the regime taxonomy of Section 3.3 is sensitive to, since Definition 3.7 is stated in terms of Δ_j and ΔJ rather than g_i at $\mu_{\pi^\star}$.

Remark (non-linear proxies activate the g-level gap). The linear-Gaussian construction is structurally incapable of exhibiting a non-degenerate g-level gap, because the OLS-style reliance map $g(\bar{R}; \mu)$ recovers c exactly for any linear $\bar{R} = c^\top y$, independent of μ. Activating a non-trivial g-level gap requires breaking exact representability of $\tilde{R}$ in $\text{span}(\Phi)$ in a way that couples to moments of μ_{diag} that vary across audit distributions. Augmenting $\tilde{R}$ with a non-linear term outside $\text{span}(\Phi)$ (e.g., an interaction $\kappa\phi_1\phi_2$ at non-symmetric μ_{diag}, or a higher-order term whose relevant cross-moments differ across $\mu^{(1)}_{\text{diag}}$ and $\mu^{(2)}_{\text{diag}}$) makes $g(\tilde{R}; \mu_{\text{diag}})$ depend on those moments. Under such an augmentation, $M_1^{\mu^{(1)}_{\text{diag}}} \neq M_1^{\mu^{(2)}_{\text{diag}}}$ as operators and the operator-level audit sensitivity of Definition 3.8 becomes instantiable with π_{ref} held fixed across audit distributions. While we establish non-vacuity with the linear-Gaussian, Section A.9 provides the closed-form non-linear instantiation. Real reward models are non-linear in any tractable feature basis, so the audit-side evidence in Sections B.3 and 4 supports the operator-level reading of Definition 3.8 in deployment, beyond the controlled construction.

Regimes R0–R3. We work in the coupled case $\pi_{\text{ref}} = \mu_{\text{diag}} = \mathcal{N}(0, \Sigma)$ for the same reason as the gap analysis above: under decoupled isotropic π_{ref} the post-mitigation policy is $\mathcal{N}((0, b, w)/\beta, I_3)$, which gives $\Delta_2 = 0$ and $\Delta J = 0$ identically and collapses the regime taxonomy to a single regime. The coupled case is the minimal setting in which the five regimes are jointly distinguishable in this construction. The limitation of relying on $\pi_{\text{ref}} = \mu_{\text{diag}}$ noted under R4 below applies symmetrically here. Under this coupling, the policy-mean for any linear reward $\bar{R} = c^\top y$ is $\Sigma c/\beta$. The pre-mitigation policy is $\pi = \pi^\star_\beta(\tilde{R}) = \mathcal{N}(\Sigma(a, b, w)^\top/\beta, \Sigma)$ and the post-mitigation policy is $\pi' = \pi^\star_\beta(M_1(\tilde{R})) = \mathcal{N}(\Sigma(0, b, w)^\top/\beta, \Sigma)$. The off-target spurious-feature drift and true-reward change evaluate to

$$\Delta_2(\pi, \pi') = -\rho a/\beta, \qquad \Delta J = -\Sigma_{13} \cdot aw/\beta = 0,$$

the latter because $\Sigma_{13} = 0$ in our parameterization. This makes $\Delta J = 0$ in the strict $\Sigma_{13} = 0$ case and rotation onto ϕ_2 governed entirely by ρa. To populate all five regimes we relax Σ_{13} as a free parameter $\sigma_{13} \in (-1, 1)$ (with Σ remaining positive definite), giving $\Delta J = -\sigma_{13}\, aw/\beta$. The five regimes correspond to:

- *R0 (successful mitigation):* $\rho = 0$, $\sigma_{13} < 0$, $a, w > 0$. Then $\Delta_2 = 0$ and $\Delta J = -\sigma_{13} aw/\beta > 0$. Removing spurious ϕ_1-reliance improves true reward because ϕ_1 was anti-correlated with ϕ_3 at μ_{diag}, so the unmitigated proxy was pushing the policy away from high-ϕ_3 regions.
- *R0$_{cont}$ (contaminated success):* $\rho \neq 0$, $\sigma_{13} < 0$, $a, w > 0$. Then $\Delta_2 = -\rho a/\beta \neq 0$ and $\Delta J = -\sigma_{13} aw/\beta > 0$. True reward improves as in R0 (driven by $\sigma_{13} < 0$), but the cross-correlation ρ couples ϕ_1-reduction to first-moment drift on ϕ_2, so the substitution mechanism is active despite improvement.
- *R1 (bias substitution):* $\rho \neq 0$, $\sigma_{13} = 0$, $a \neq 0$. Then $\Delta_2 = -\rho a/\beta \neq 0$ and $\Delta J = 0$ (neutral substitution). Setting $\sigma_{13} > 0$ instead gives $\Delta J < 0$ (harmful substitution).
- *R2 (overcorrection):* $\rho = 0$, $\sigma_{13} > 0$, $a, w > 0$. Then $\Delta_2 = 0$ and $\Delta J = -\sigma_{13} aw/\beta < 0$. The mechanism is audit-side cross-correlation between ϕ_1 and the structural ϕ_3: removing

$a\phi_1$ shifts the policy mean of ϕ_3 by $-\sigma_{13}a/\beta$, lowering true reward without rotating pressure onto ϕ_2. The rescalability test of Section A.13 gives $\Delta J(c) = -c\,\sigma_{13}aw/\beta$ for partial mitigation $M_1^c(\tilde{R}) = \tilde{R} - c\,a\phi_1$, monotone in $c \in (0,1)$, so no partial mitigation recovers the unmitigated return; by the operational test of R2 in Definition 3.7 the construction sits on the non-rescalable side of R2 (driven here by reference-policy coupling $\Sigma_{13} \neq 0$ rather than the named target-misspecification mechanism, which would require $\phi_1 \in \Phi_{\text{struct}}$).

- *R3 (silent non-op):* $\rho = 0$, $\sigma_{13} = 0$. Then $\Delta_2 = 0$ and $\Delta J = 0$ regardless of a. The mitigation alters $\tilde{R}$ along an axis the policy's true-reward-relevant directions are orthogonal to.

Each regime is realized by an open set of parameter values, not a measure-zero corner.

Audit-distribution sensitivity (R4). The strict operator-level reading of Definition 3.8 (varying μ_{diag} with π_{ref} held fixed) is degenerate under linearity: in the linear-Gaussian setting $g(\tilde{R}; \mu_{\text{diag}}) = (a, b, w)$ at every non-degenerate μ_{diag}, so $M_1^{\mu_{\text{diag}}^{(1)}}(\tilde{R}) = M_1^{\mu_{\text{diag}}^{(2)}}(\tilde{R}) = \tilde{R} - a\phi_1$ as operators, whatever the difference between $\mu_{\text{diag}}^{(1)}$ and $\mu_{\text{diag}}^{(2)}$. Linearity thus characterizes the boundary at which Definition 3.8 activates: any operator-level instantiation requires the non-linear extension of the preceding remark (instantiated closed-form in Section A.9), and Section B.3 supplies the empirical counterpart across eight models from different families.

The closed form does realize R4 jointly with π_{ref} under the coupling $\pi_{\text{ref}} = \mu_{\text{diag}}$ (the audit distribution serving as the rollout reference). Hold $(\tilde{R}, R, \Phi_{\text{sp}}, \beta)$ and the operator M_1 fixed with $a, b, w > 0$, and take two pairs $(\mu_{\text{diag}}^{(\ell)}, \pi_{\text{ref}}^{(\ell)}) = \big(\mathcal{N}(0, \Sigma^{(\ell)}), \mathcal{N}(0, \Sigma^{(\ell)})\big)$ for $\ell = 1, 2$ with $\Sigma_{12}^{(1)} = \rho^{(1)} \neq 0$, $\sigma_{13}^{(1)} = 0$, and $\Sigma_{12}^{(2)} = \rho^{(2)} \neq 0$, $\sigma_{13}^{(2)} < 0$ (each $\Sigma^{(\ell)}$ positive definite, all other entries equal across the two). Then

$$\Delta_2^{(\ell)} = -\rho^{(\ell)} a/\beta, \qquad \Delta J^{(\ell)} = -\sigma_{13}^{(\ell)} aw/\beta,$$

so $\ell = 1$ realizes R1 (neutral substitution: $\Delta_2 \neq 0$, $\Delta J = 0$) and $\ell = 2$ realizes R0_{cont} ($\Delta_2 \neq 0$, $\Delta J > 0$): the audit distribution selects the regime, with π_{ref} entering jointly via the coupling. For the strict operator-level version (π_{ref} fixed across audit distributions), Section A.9 provides the closed-form instantiation and Section B.3 the empirical counterpart (sign reversal of the length-sycophancy coupling under human-LLM judge disagreement).

Norm increase under M_i. The scale identity from Section A.5 reads $\|\tilde{R}'\|^2_{L^2(\mu_{\text{diag}})} = \|\tilde{R}\|^2_{L^2(\mu_{\text{diag}})} - g_1[g_1 G_{11} + 2(G_{12} g_2 + G_{13} g_3)]$. With $G = \Sigma$ in our parameterization, $G_{11} = 1$, $G_{12} = \rho$, $G_{13} = \sigma_{13}$, and $(g_1, g_2, g_3) = (a, b, w)$, the bracket equals $a + 2\rho b + 2\sigma_{13} w$. Choosing $a = 0.5$, $b = w = 1$, $\rho = -0.6$, $\sigma_{13} = -0.4$ gives bracket $= 0.5 - 1.2 - 0.8 = -1.5$ with bracket sign opposite to $g_1 = 0.5$. The norm change is $-g_1 \cdot (-1.5) = +0.75$, so $\|\tilde{R}'\|^2 > \|\tilde{R}\|^2$ strictly. Mitigation increases the proxy's L^2 norm because the cross-correlation contribution dominates the diagonal term, exactly the failure mode flagged in Section A.5.

A.8 Epsilon-Banded Regime Classification

Finite-sample classification of mitigations into the regimes of Definition 3.7 requires bands, since exact equality on Δ_j and ΔJ is a measure-zero event under any non-degenerate sampling design.

Banded definitions. Fix tolerances $\varepsilon_j \geq 0$ for each $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$ and $\varepsilon_J \geq 0$. With π, π' as in Definition 3.7, define

$$\begin{aligned}
\text{R0}_\varepsilon: &\quad \Delta J > \varepsilon_J \text{ and } |\Delta_j(\pi, \pi')| \leq \varepsilon_j \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\
\text{R0}_{\varepsilon,\text{cont}}: &\quad \Delta J > \varepsilon_J \text{ and } |\Delta_j(\pi, \pi')| > \varepsilon_j \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\
\text{R1}_\varepsilon: &\quad |\Delta_j(\pi, \pi')| > \varepsilon_j \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \quad \text{and} \quad \Delta J \leq \varepsilon_J, \\
\text{R2}_\varepsilon: &\quad |\Delta_j(\pi, \pi')| \leq \varepsilon_j \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \quad \text{and} \quad \Delta J < -\varepsilon_J, \\
\text{R3}_\varepsilon: &\quad |\Delta_j(\pi, \pi')| \leq \varepsilon_j \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \quad \text{and} \quad |\Delta J| \leq \varepsilon_J.
\end{aligned}$$

R1's sub-cases port unchanged: neutral and harmful substitution correspond to $|\Delta J| \leq \varepsilon_J$ and $\Delta J < -\varepsilon_J$ respectively. The spurious-rotation-with-improvement corner of Section A.3 is regime $\mathrm{R0}_{\varepsilon,\mathrm{cont}}$ per the definition above. Definition 3.8 ports verbatim, since it is transversal to the regime classification and unaffected by the choice of bands.

Nested limit. Taking $\varepsilon_j, \varepsilon_J \to 0$ recovers Definition 3.7 pointwise: $\mathrm{R0}_\varepsilon \to \{\text{no rotation, } \Delta J > 0\}$, $\mathrm{R0}_{\varepsilon,\mathrm{cont}} \to \{\text{rotation, } \Delta J > 0\}$, $\mathrm{R1}_\varepsilon \to \{\text{rotation, } \Delta J \leq 0\}$, $\mathrm{R2}_\varepsilon \to \{\text{no rotation, } \Delta J < 0\}$, $\mathrm{R3}_\varepsilon \to \{\text{no rotation, } \Delta J = 0\}$, with the disjunctive condition $|\Delta_j| > \varepsilon_j$ for some j converging to $\Delta_j \neq 0$ for some j as $\varepsilon_j \to 0$.

Default choice: noise-floor bands. We default to noise-floor bands, in which $\varepsilon_j = z_{\alpha/2} \cdot \widehat{\mathrm{SE}}(\widehat{\Delta}_j)$ under the empirical sampling design (e.g., the standard error of a fixed-effect coefficient in a mixed linear model, as in Section B.3, where the sycophancy-on-length coefficient instantiates $\widehat{\Delta}_{\text{length}}$) and $\varepsilon_J = z_{\alpha/2} \cdot \widehat{\mathrm{SE}}(\widehat{\Delta J})$ for a significance level α chosen and reported by the practitioner. Setting $\varepsilon_j = \widehat{\mathrm{SE}}$ without the $z_{\alpha/2}$ factor corresponds to approximately 68% confidence and should not be conflated with a standard significance test; regime assignments should always state the chosen α. An effect-size-relative alternative (ε_j as a fraction of unmitigated $\mathbb{E}_{\mu_\pi}[\phi_j]$, ε_J as a fraction of unmitigated J) is appropriate for cross-proxy comparisons; the two can disagree on borderline cases and the choice should be stated. When $|\Phi_{\mathrm{sp}} \setminus \{\phi_i\}| > 1$, the conservative reading (declare rotation if any axis exceeds its ε_j band, no multiple-testing correction) inflates the false-positive rate of the rotation criterion. Because rotation is the discriminating predicate for both $\mathrm{R0}_{\varepsilon,\mathrm{cont}}$ (against $\mathrm{R0}_\varepsilon$) and $\mathrm{R1}_\varepsilon$ (against $\mathrm{R3}_\varepsilon \cup \mathrm{R2}_\varepsilon$), this inflation pulls cases out of $\mathrm{R0}_\varepsilon$ into $\mathrm{R0}_{\varepsilon,\mathrm{cont}}$ when $\Delta J > \varepsilon_J$, and out of $\mathrm{R3}_\varepsilon$ and $\mathrm{R2}_\varepsilon$ into $\mathrm{R1}_\varepsilon$ when $\Delta J \leq \varepsilon_J$. This choice is deliberate to prioritize sensitivity to substitution over specificity in either ΔJ regime, consistent with the conservative-partition stance of Section 3.1 and the neutral sub-case absorbing near-zero improvements ($\Delta J \in (0, \varepsilon_J]$) in the same direction.

Mapping wild classifications onto the bands. The R0–R4 calls in Sections 3.3 and 4 are made under the banded reading, with $\varepsilon_j, \varepsilon_J$ inferred from each cited work's reported uncertainty. The [Bu et al., 2025] accuracy degradation under uniform length penalties satisfies the $\mathrm{R2}_\varepsilon$ pattern under noise-floor bands, with the rescalability test of Section A.13 itself constituting a banded operationalization (existence of $c \in (0,1)$ improving $\widehat{J}$ implicitly assumes detection against ε_J). The Fein et al. [2026] sign flip together with reduced correctness is read as $\mathrm{R1}_\varepsilon$ under the rotation interpretation (off-target spurious axis exceeds its band, $\widehat{\Delta J} \leq \varepsilon_J$). The alternative reading as $\mathrm{R2}_\varepsilon$ on the targeted axis itself is also consistent with the reported statistics, and we flag this ambiguity rather than resolve it. The human-vs-LLM-judge gap is operationalized via Section B.3 below.

Mapping Section B.3 statistics onto the bands. The mixed-model coefficients reported in Section B.3 ($+24.3$ under human labels, CI $[9.6, 39.0]$, $+128.4$ under LLM labels, CI $[118.5, 138.4]$, $+154.3$ under judge agreement, CI $[123.6, 185.0]$, -43.1 under judge disagreement, CI $[-60.5, -25.7]$) all satisfy the noise-floor ε_j-band test on the length axis, since each CI excludes zero. The pooled KS statistics on within-model residuals ($0.025, 0.130, 0.130, 0.046$, all with $p < 0.05$) provide a distributional band test alongside the first-moment Δ_j of Definition 3.7. The sign reversal of the length-sycophancy coupling between the agreement-side regimes and the disagreement regime is consistent with the construction-level precondition for $\mathrm{R4}_\varepsilon$, in that two annotator-conditioned audit distributions yield g-profiles with opposite signs on the relevant cross-feature coupling. The operator-level g-difference is not measured empirically here, but its constructive counterpart appears in Section A.9.

A.9 Phase-Diagram Validation of Regime Transitions Under Quadratic Non-Linearity

Extending the linear Gaussian non-vacuity instantiation of Section A.7, we use a quadratic non-linear reward structure to empirically show

- clear instantiation of all regimes and sub-regimes (R0, $\mathrm{R0}_{\text{cont}}$, $\mathrm{R1}_{\text{neut}}$, $\mathrm{R1}_{\text{harm}}$, R2, R3) with phase-diagrams illustrating how the regime boundaries are outcomes of a single mitigation operator in a controlled setting

- that the regimes are not a property of the reward model alone but of $(R, \tilde{R}, M_i, \mu_{\text{diag}})$ as stated in Section 3.
- clear instantiation for the regime transition as described by R4 from only varying the audit distribution μ_{audit} mean
- a match of the theoretical predictions connecting the analytic phase-diagram claims to finite-N experiments, strengthening the non-vacuity results.

This generalizes the linear-Gaussian setting, as we show that the regime taxonomy is not a linear-Gaussian artifact. For real deployment instantiations, see different experiments in Section B.

Setup. We extend the setup in Section A.7: We work with a trivial prompt space ($\mathcal{X}$ a singleton, suppressed in notation) and continuous response space $\mathcal{Y} = \mathbb{R}^3$. The feature map is $\Phi(y) = (\phi_1, \phi_2, \phi_3)$ with $\phi_k(y) = y_k$. We fix true and the preference-generating annotator reward

$$R(y) = w\,\phi_3(y), \qquad R_{\text{anno}}(y) = a\,\phi_1(y) + b\,\phi_2(y) + w\,\phi_3(y) + \gamma\,\phi_1(y)^2,$$

with $w, \gamma > 0$ and $(a, b) \in \mathbb{R}^2$. By Definition 3.3, $\Phi_{\text{sp}} = \{\phi_1, \phi_2\}$ and $\Phi_{\text{struct}} = \{\phi_3\}$ by construction. For the reward model (learned proxy reward), we use a linear polynomial such that

$$\tilde{R}(y) = \theta_1\phi_1(y) + \theta_2\phi_2(y) + \theta_3\phi_3(y) + \theta_4\phi_1(y)^2,$$

which is correctly specified using enough samples and the maximum likelihood estimator for Bradley-Terry (BT-MLE) as $\hat{\theta} = (a, b, w, \gamma)$.[6] Reference policy and diagnostic measure are Gaussians on $\mathbb{R}^3$,

$$\pi_{\text{ref}} = \mathcal{N}(0, \Sigma_{\text{ref}}), \qquad \mu_{\text{diag}} = \mathcal{N}(m, \Sigma_{\text{ref}}),$$

where Σ is positive definite with

$$\Sigma_{\text{ref}} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & 0 \\ \rho_{13} & 0 & 1 \end{pmatrix}, \qquad \rho_{12}^2 + \rho_{13}^2 < 1$$

Although this setup ties the covariance of the reference and audit distribution, we show that the mean parameter $m \in \mathbb{R}^3$ of the audit distribution μ_{diag} is sufficient to instantiate R4. We fix the KL parameter $\beta > 0$ and target the spurious feature $i = 1$ for mitigation throughout.

Mitigation observables. Following Definition 3.4, the linear reliance g_i for targeted $i = 1$ is

$$g_1(\theta, m) = \theta_1 + \theta_2\rho_{12} + \theta_3\rho_{13} + 2\theta_4 m_1.$$

The resulting single-axis mitigation with strength $c \geq 0$ is then

$$M_1(\tilde{R}(y)) = \tilde{R}(y) - cg_1\phi_1(y).$$

We rewrite the reward model as

$$\tilde{R}(y) = \alpha \cdot y + \frac{1}{2}y^\top H y, \qquad \text{with } \alpha = (\theta_1, \theta_2, \theta_3), H = \text{diag}(2\theta_4, 0, 0),$$

reparameterizing the KL-regularized optimum policy as

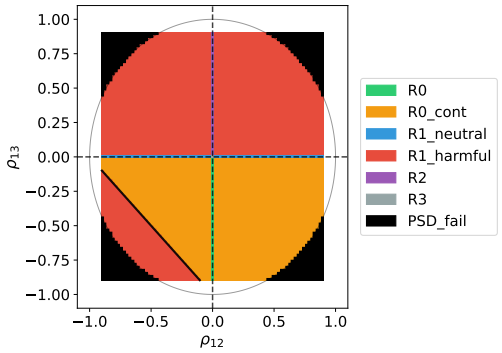
$$\pi^\star = \mathcal{N}(\mu^\star, \Sigma^\star) \qquad \text{with } \Sigma^{\star -1} = \Sigma_{\text{ref}}^{-1} - \frac{H}{\beta}, \mu^\star = \frac{\Sigma^\star \alpha}{\beta}.$$

Since H is unchanged by the mitigation, $\Sigma^\star$ is also unchanged and only α shifts by $\Delta\alpha = (-cg_1, 0, 0)$, leading to the regime-defining changes in the policy-induced feature expectation and true reward (see Section 3.3) as

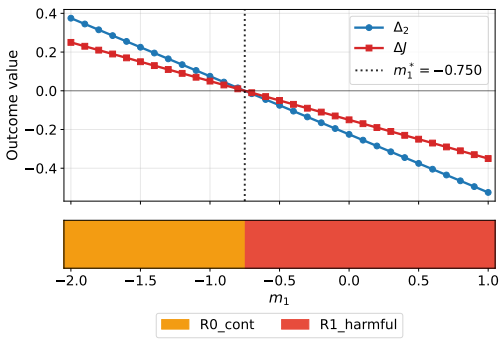
$$\Delta_j = (\Sigma^\star \Delta\alpha)/\beta = -\frac{cg_1}{\beta}\Sigma^\star_{j,1}, \qquad \Delta J = w\Delta_3 = -w\frac{cg_1}{\beta}\Sigma^\star_{3,1}.$$

Expanding with Sherman-Morrison gives $\Sigma^\star_{j,1} = \rho_{1j(1+\kappa)}$ with $\kappa = (2\gamma/\beta)/(1 - 2\gamma/\beta) \geq 0$, implying that rescaling γ does not flip the sign of Δ_j for regime classification.

[6] We use the analytical idealization for the correct specification of the reward model for closed-form regime predictions. Under realistic RM misspecification, the same regime phenomena persist with additional noise, as documented empirically across five reward models and a non-linear operator in Section B.2.



(a) Phase diagram from varying (ρ_{12}, ρ_{13}) at fixed $\gamma = 0.4, m = 0, c = 1, \beta = 4$. We color each grid cell with the corresponding regime from Section 3.3.



(b) R4 transition of Δ_2 and ΔJ induced by varying the audit distribution μ_{diag} mean $m_1 \in [-2, 1]$ at fixed $\gamma = 1.0, \rho_{12} = 0.3, \rho_{13} = 0.2, c = 1, \beta = 4$.

Figure 3: We instantiate all five R0–R3 as tune-able outcomes of a single mitigation operator and R4 by varying only the mean of the audit distribution μ_{diag} in a controlled setting. Figure 3a shows that a non-linearity term creates new boundary structure and that the regimes are not solely induced as a reward model property.

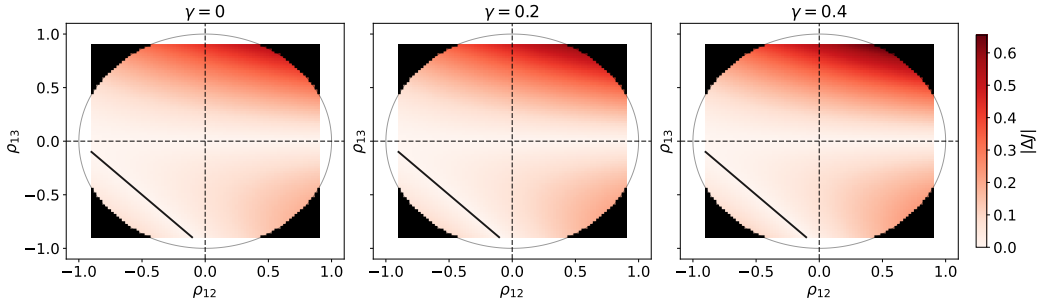


Figure 4: Shifting γ to produce different $|\Delta J|$-heatmaps of Figure 3a to illustrate regime robustness to non-linearity strength, as the regime boundaries do not change themselves.

Simulation parameters. With this setup, we have the following parameters to tune and induce regime transitions:

- ρ_{12} sets rotation for bias substitution and also the sign of Δ_2
- ρ_{13} sets the sign of ΔJ (if the sign of g_1 does not change)
- γ amplifies Δ_j and ΔJ, but does not move boundaries at $m = 0$
- m_1 shifts linear reliance g_1 and leads to g_1 sign flip at $m_1^\star = -(a + b\rho_{12} + w\rho_{13})/(2\gamma)$ inducing the transition for R4

The KL parameter β and the mitigation scale parameter c only set margins and scales.

Experiment Results. Using $N = 10^5$ preference samples, 50 seeds per validation cell, a linearly spaced grid of 81 by 81 values for $\rho_{12}, \rho_{13} \in (-0.9, 0.9)^2$ each, and a (near-numerical precision) ε-bound of 10^{-8} (see Section A.8), we instantiate all five regimes R0–R3 and the transition R4, while also validating the above theoretical predictions with finite-N sample experiments. Figure 3 shows the instantiation, Figure 4 shows the regime boundary robustness to non-linearity strength, Figure 5 verifies predictions and measurements, and Table 1 presents the numerical values and uncertainties of Figure 5.

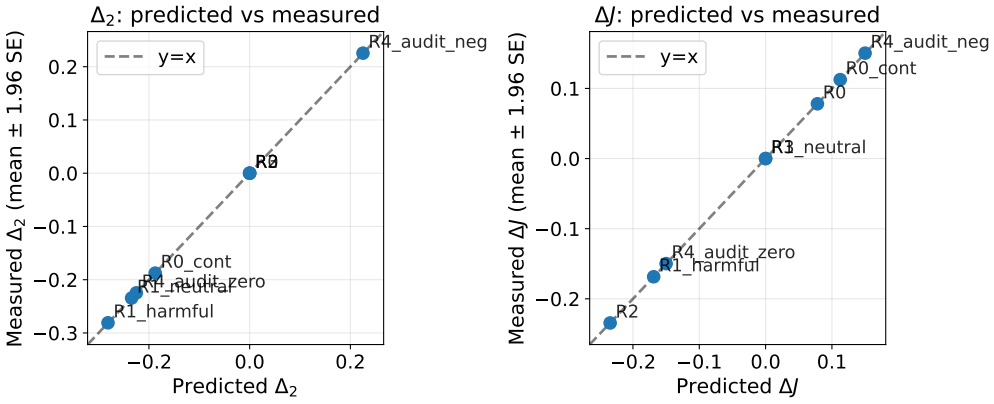


Figure 5: Validating that the simulated regime transitions align with the theoretical predictions from our setup for Δ_2 and ΔJ with standard errors (SE) from the BT-MLE with $N = 10^5$ samples, aggregated over 50 seeds. This result connects the analytic phase-diagram claims to finite-N experiments.

Table 1: Validation of closed-form regime predictions against finite-N BT-MLE on simulated preferences. All cells use $a = b = w = c = 1, \beta$=4, N=10^5 samples per seed, for S=50 seeds. The first six columns are the regime instances for $m_1 = 0$, except for R4 where we vary m_1 at fixed $\rho_{12}, \rho_{13}, \gamma$ (see Figure 3b). All eight cells achieve 100% agreement on the respective regime call.

Cell	**R0**	**R0**$_{\text{cont}}$	**R1**$_{\text{neut}}$	**R1**$_{\text{harm}}$	**R2**	**R3**	**R4**	
Scenario								
ρ_{12}	0.0	0.5	0.5	0.5	0.0	0.0	0.3	0.3
ρ_{13}	−0.5	−0.3	0.0	0.3	0.5	0.0	0.2	0.2
γ	0.4	0.4	0.4	0.4	0.4	0.4	1.0	1.0
m_1	0.0	0.0	0.0	0.0	0.0	0.0	−1.5	0.0
Induced g_1	0.500	1.200	1.500	1.800	1.500	1.000	−1.500	1.500
Δ_2								
predicted	0.0000	−0.1875	−0.2344	−0.2812	0.0000	0.0000	0.2250	−0.2250
measured	0.0000	−0.1876	−0.2344	−0.2809	0.0000	0.0000	0.2255	−0.2247
SE	0.0000	0.0002	0.0003	0.0004	0.0000	0.0000	0.0006	0.0005
Δ_J								
predicted	0.0781	0.1125	0.0000	−0.1688	−0.2344	0.0000	0.1500	−0.1500
measured	0.0781	0.1126	0.0000	−0.1686	−0.2345	0.0000	0.1503	−0.1498
measured SE	0.0002	0.0001	0.0000	0.0002	0.0003	0.0000	0.0004	0.0003
Regime classification	R0	R0$_{\text{cont}}$	R1$_{\text{neut}}$	R1$_{\text{harm}}$	R2	R3	R0$_{\text{cont}}$	R1$_{\text{harm}}$

A.10 Formal Proof of Audit-Distribution Insufficiency

We give a formal proof of the structural-blindness claim of Section 3.4 that standard ordinal reward-model benchmarks cannot reliably distinguish R0 from R0$_{\text{cont}}$, R1, or R2.

Benchmark class. Let $\mathcal{B}$ denote the class of benchmark functionals depending only on the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag}. We take μ_{diag} as the empirical measure of the evaluation set of $\mathcal{B}$ where applicable. This class subsumes ranking accuracy, pairwise win-rate, preference-prediction calibration, reward-target correlation, the linear-reliance statistic g of Definition 3.4, and any composite of these, covering the evaluation paradigm of Section 3.5. Note that including R in the input of $\mathcal{B}$ strengthens the result, since standard benchmarks lack oracle access to R.

Theorem 3.9 (Audit-distribution insufficiency)**.** *Fix $\beta > 0$. There exist four choices of π_{ref} with μ_{diag}, $\tilde{R}$, M_i, R, Φ, Φ_{sp}, and β held fixed s.t.*

(i) *the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag} is identical across the four instances, so B takes the same value on all four for every $B \in \mathcal{B}$,*

(ii) *the optimized policies $\pi^\star_\beta(M_i(\tilde{R}))$ fall into regimes R0, R0$_{cont}$, R1, and R2 respectively, in the sense of Definition 3.7.*

Proof. **Construction.** Take $\mathcal{X}$ a singleton (suppressed), $\mathcal{Y} = \mathbb{R}^3$, feature map $\phi_k(y) = y_k$, and rewards $R(y) = \phi_3(y)$, $\tilde{R}(y) = \phi_1(y) + \phi_2(y) + \phi_3(y)$. Set $\Phi_{\mathrm{sp}} = \{\phi_1, \phi_2\}$, target $\phi_i = \phi_1$, and audit distribution $\mu_{\mathrm{diag}} = \mathcal{N}(0, I_3)$. Fix a small $\delta \in (0, 1/2)$. The four instances differ only in $\pi^{(k)}_{\mathrm{ref}} = \mathcal{N}(0, \Sigma^{(k)})$, where each $\Sigma^{(k)}$ has unit diagonal, $\Sigma^{(k)}_{23} = 0$, and

$$\big(\Sigma^{(0)}_{12}, \Sigma^{(0)}_{13}\big) = \big(0, -\tfrac{1}{2}\big), \quad \big(\Sigma^{(1)}_{12}, \Sigma^{(1)}_{13}\big) = \big(\tfrac{1}{2}, -\tfrac{1}{2}\big),$$
$$\big(\Sigma^{(2)}_{12}, \Sigma^{(2)}_{13}\big) = \big(\tfrac{1}{2}, \delta\big), \quad \big(\Sigma^{(3)}_{12}, \Sigma^{(3)}_{13}\big) = \big(0, \tfrac{1}{2}\big).$$

Each $\Sigma^{(k)}$ has $\det = 1 - \Sigma^2_{12} - \Sigma^2_{13} > 0$ by Sylvester's criterion: $\det^{(0)} = \det^{(3)} = 3/4$, $\det^{(1)} = 1/2$, and $\det^{(2)} = 3/4 - \delta^2 > 0$ for $\delta < 1/2$. We work on a sufficiently large bounded set to satisfy Assumption A.1. Expressions extend to the unrestricted Gaussian limit as in Section A.7. Linearity and Gaussianity are used for closed-form existence and are not required for the impossibility claim. Assumption 3.2 holds since $\mathbb{E}_{\mu_{\mathrm{diag}}}[\Phi\Phi^\top] = I_3$. Assumption A.2 holds since $\mathcal{Y} = \mathbb{R}^3$ admits independent coordinate perturbations on a set of full μ_{diag}-measure.

(i) Audit-side identity. The Gram at μ_{diag} is I_3, giving $g(\tilde{R}; \mu_{\mathrm{diag}}) = (1, 1, 1)$ and $M_i(\tilde{R}) = \phi_2 + \phi_3$ across all four instances. Since $\tilde{R}$, $M_i(\tilde{R})$, R, Φ, and μ_{diag} are identical across k, the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag} is identical across k, so B takes the same value on all four instances for every $B \in \mathcal{B}$.

(ii) Policy-side classification. For any linear $\bar{R}(y) = c^\top y$ and $\pi_{\mathrm{ref}} = \mathcal{N}(0, \Sigma)$, completing the square in Equation (2) gives $\pi^\star_\beta(\bar{R}) = \mathcal{N}(\Sigma c/\beta,\ \Sigma)$. Applying this with $c = (1, 1, 1)^\top$ and $c' = (0, 1, 1)^\top$ at each $\Sigma^{(k)}$:

$$\Delta_2(\pi^{(k)}, \pi'^{(k)}) = -\Sigma^{(k)}_{12}/\beta, \qquad \Delta J^{(k)} = -\Sigma^{(k)}_{13}/\beta,$$

where Δ_j is checked against $\Phi_{\mathrm{sp}} \setminus \{\phi_i\} = \{\phi_2\}$. Substituting:

$$\big(\Delta_2, \Delta J\big)^{(0)} = \big(0, +\tfrac{1}{2\beta}\big), \quad \big(\Delta_2, \Delta J\big)^{(1)} = \big(-\tfrac{1}{2\beta}, +\tfrac{1}{2\beta}\big),$$
$$\big(\Delta_2, \Delta J\big)^{(2)} = \big(-\tfrac{1}{2\beta}, -\tfrac{\delta}{\beta}\big), \quad \big(\Delta_2, \Delta J\big)^{(3)} = \big(0, -\tfrac{1}{2\beta}\big).$$

By Definition 3.7, these are R0, R0$_{\mathrm{cont}}$, R1 (harmful substitution), and R2 respectively. □

Remark (scope). The theorem fixes μ_{diag} and M_i across instances. R3 is additionally realizable in the same construction by setting $\Sigma_{12} = \Sigma_{13} = 0$, with audit observables still identical and distinguishability from R0 requiring $D_{\mathrm{KL}}(\pi' \,\|\, \pi)$, which is not audit-local. R4 requires non-linear proxies (see Sections A.7 and A.9). The varying π_{ref} corresponds to the same reward model deployed against different reference policies, the standard regime in which benchmarks claim to evaluate reward models independently of downstream RLHF setup.

Functional blindspot. Theorem 3.9 establishes the distributional blindspot. The functional blindspot is independent and follows from cardinal-scale sensitivity:

Corollary A.5 (Functional blindspot via reward rescaling)**.** *Let $\mathcal{B}_{\mathrm{ord}} \subseteq \mathcal{B}$ denote the benchmark functionals satisfying $B(f \circ \tilde{R}, f \circ M_i(\tilde{R}), R, \Phi, \mu_{\mathrm{diag}}) = B(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\mathrm{diag}})$ for every strictly monotone increasing $f\colon \mathbb{R} \to \mathbb{R}$. Then for the construction of Theorem 3.9, every $B \in \mathcal{B}_{\mathrm{ord}}$, and every $c > 0$,*

$$B(c\tilde{R}, M_i(c\tilde{R}), R, \Phi, \mu_{\mathrm{diag}}) = B(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\mathrm{diag}}),$$

while $J(\pi^\star_\beta(c\tilde{R}), R)$ is strictly monotone in c.

Proof. By linearity, $M_i(c\tilde{R}) = c\, M_i(\tilde{R})$, and ordinal invariance under $f(y) = cy$ gives the benchmark identity. The KL identity $\pi^\star_\beta(c\bar{R}) = \pi^\star_{\beta/c}(\bar{R})$ (Section A.5) gives $J(\pi^\star_\beta(c\tilde{R}), R) = c{\cdot}(\Sigma c_{\tilde{R}})_3/\beta$ in the construction of Theorem 3.9, where $c_{\tilde{R}} = (1, 1, 1)^\top$ and $(\Sigma^{(k)} c_{\tilde{R}})_3 = \Sigma^{(k)}_{13} + 1 \neq 0$ for each k, so J is strictly monotone in c. □

A.11 Audit-Distribution Sufficiency

Theorem 3.9 establishes that benchmark functionals depending only on $(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\mathrm{diag}})$ cannot reliably separate R0 from $\mathrm{R0}_{\mathrm{cont}}$, R1, or R2. We now show that adding the policy-induced distributions $\mu_{\pi^\star_\beta(\tilde{R})}$ and $\mu_{\pi^\star_\beta(M_i(\tilde{R}))}$ to the input suffices, certifying the prescription that benchmarks must evaluate at policy-induced distributions (Do 1 of Section A.13).

Extended benchmark class. Let $\mathcal{B}^+$ denote the class of benchmark functionals depending on the joint distributions of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under each of μ_{diag}, $\mu_{\pi^\star_\beta(\tilde{R})}$, and $\mu_{\pi^\star_\beta(M_i(\tilde{R}))}$, together with the partition $\Phi = \Phi_{\mathrm{sp}} \sqcup \Phi_{\mathrm{struct}}$. We have $\mathcal{B} \subset \mathcal{B}^+$.

Theorem 3.10 (Audit-distribution sufficiency)**.** *Fix $\beta > 0$ and let $\pi = \pi^\star_\beta(\tilde{R})$, $\pi' = \pi^\star_\beta(M_i(\tilde{R}))$. With Δ_j and ΔJ as in Definition 3.7 and Section 3.3, define $B^\star \in \mathcal{B}^+$ by*

$$B^\star = \begin{cases} \mathrm{R0} & \text{if } \Delta J > 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}, \\ \mathrm{R0}_{\mathrm{cont}} & \text{if } \Delta J > 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}, \\ \mathrm{R1} & \text{if } \Delta J \leq 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}, \\ \mathrm{R2} & \text{if } \Delta J < 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}. \end{cases} \tag{6}$$

Under Assumptions A.1, 3.2, and A.2, with $M_i(\tilde{R})$ likewise satisfying Assumption A.1, for every tuple $(\pi_{\mathrm{ref}}, \tilde{R}, M_i, R, \Phi, \Phi_{\mathrm{sp}})$ with $(\pi, \pi') \in \mathrm{R0} \cup \mathrm{R0}_{\mathrm{cont}} \cup \mathrm{R1} \cup \mathrm{R2}$ in the sense of Definition 3.7, $B^\star$ returns the correct regime label.

Proof. Δ_j and ΔJ are first moments of Φ and R under μ_π and $\mu_{\pi'}$, both inputs to $\mathcal{B}^+$, so $B^\star \in \mathcal{B}^+$. Finiteness of these expectations follows from Assumption A.1 via the boundedness chain in Section A.2, applied to both $\tilde{R}$ and $M_i(\tilde{R})$ (the latter by closure under finite linear combinations).

Correctness on each regime follows directly from Definition 3.7: R0 requires $\Delta J > 0$ with no off-target rotation (first branch), $\mathrm{R0}_{\mathrm{cont}}$ requires $\Delta J > 0$ with off-target rotation (second branch), R1 requires $\Delta J \leq 0$ with off-target rotation (third branch), R2 requires $\Delta J < 0$ with no off-target rotation (fourth branch). The four branches are pairwise disjoint: $\{\mathrm{R0}, \mathrm{R0}_{\mathrm{cont}}\}$ have $\Delta J > 0$ and $\{\mathrm{R1}, \mathrm{R2}\}$ have $\Delta J \leq 0$, separating any cross-pair by the sign of ΔJ. Within $\{\mathrm{R0}, \mathrm{R0}_{\mathrm{cont}}\}$ and within $\{\mathrm{R1}, \mathrm{R2}\}$ the branches differ in rotation ($\mathrm{R0}_{\mathrm{cont}}$, R1 require rotation, R0 and R2 do not). Inputs in R3 (no rotation, $\Delta J = 0$) are excluded by hypothesis. □

Corollary A.6 (Sufficiency for R3)**.** *The classifier $B^\star$ extends to $\{R0, R1, R2, R3\}$ by adding a fourth branch: return R3 if $\Delta J = 0$ and $\Delta_j = 0$ for all $\phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}$. This branch matches R3 (silent non-op) of Definition 3.7 exactly and is disjoint from the R0–R2 branches by the $\Delta J = 0$ condition. The observable $D_{\mathrm{KL}}(\pi' \,\|\, \pi)$ is not required for the regime call, though reporting it alongside the label distinguishes policy-relevant R3 ($D_{\mathrm{KL}} > 0$, mitigation moved the policy along directions orthogonal to $(\Delta_j, \Delta J)$) from policy-irrelevant R3 ($D_{\mathrm{KL}} = 0$, mitigation did not move the policy at all).*

Corollary A.7 (Finite-sample sufficiency)**.** *Under the noise-floor ε-bands of Section A.8, replacing exact equalities in Equation* (6) *with the banded conditions yields $B^\star_\varepsilon \in \mathcal{B}^+$ returning the correct ε-banded label.*

Corollary A.8 (Multi-audit sufficiency for R4)**.** *Let $\mathcal{B}^{++}$ denote the class of benchmark functionals depending on the inputs of $\mathcal{B}^+$ at each of $K \geq 2$ audit distributions $\mu^{(1)}_{diag}, \ldots, \mu^{(K)}_{diag}$, with the canonical $M_i^{(k)}$ constructed from each. Define $B^\star_{R4}$ as the indicator that $B^\star$ (Theorem 3.10) returns different regime labels across the K instances. Then $B^\star_{R4}$ detects R4 in the sense of Definition 3.8.*

Proof. By Theorem 3.10, each per-distribution call of $B^\star$ returns the correct R0–R2 label. Definition 3.8 defines R4 as cross-distribution disagreement on the regime label, which is exactly what $B^\star_{\mathrm{R4}}$ tests. □

Discussion. Theorem 3.9 (audit-distribution insufficiency) and Theorem 3.10 (audit-distribution sufficiency) together show that audit-only inputs cannot separate R0, $\mathrm{R0}_{\mathrm{cont}}$, R1, and R2, while

augmenting $\mathcal{B}$'s input with the policy-induced distributions $\mu_{\pi^\star_\beta(\tilde{R})}$ and $\mu_{\pi^\star_\beta(M_i(\tilde{R}))}$ suffices. Corollary A.8 extends this to R4 detection under multi-audit evaluation. This grounds the policy-distribution evaluation prescription structurally rather than heuristically. We do not claim minimality of $\mathcal{B}^+$, as the classifier $B^\star$ depends only on the scalars $(\Delta_j, \Delta J)$ and the partition Φ_{sp}, so any extension of $\mathcal{B}$ that determines these quantities also suffices.

A.12 Extension to Direct Preference Optimization

In this section, we show how our framework and theorems from Section 3 apply to DPO and related direct alignment methods [Rafailov et al., 2023, Park et al., 2024, Meng et al., 2024, Lu et al., 2024, Li et al., 2025a]. As the DPO optimum is equal to the KL-regularized RLHF optimum under the Bradley-Terry model (Equations (1) and (2)), the optimization-side of the extension carries over directly, with the exception of reference-free variants like SimPO, discussed below. However, the extension from our explicit reward formulation throughout Section 3 to an implicit reward formulation needs further support. The single-axis mitigation operator M_i (Definition 3.5) is defined acting on a proxy reward *before* optimization, whereas a DPO implicit reward exists only *after* training. We make the extension precise by treating the DPO implicit reward $\hat{r}$ as the proxy $\tilde{R}$ of our framework, as the policy $\pi = \pi^\star_\beta(\hat{r})$ is the KL-regularized optimum of its own implicit reward (see below). Under this reading the sufficiency classifier of Theorem 3.10 transfers verbatim, the regime taxonomy and the impossibility result of Theorem 3.9 transfer for mitigations that act as single-axis operators on the implicit reward.

The implicit reward as a proxy. Let π be a policy and π_{ref} the reference policy of the optimization setup in Section A.2.

Definition A.9 (DPO implicit reward)**.** *The* DPO implicit reward *of π relative to π_{ref} is*

$$\hat{r}(x, y) \;=\; \beta \,\log \frac{\pi(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}.$$

The next proposition records why $\hat{r}$ is the right object to place in the $\tilde{R}$ role. The implicit reward $\hat{r}$ induces π as its own KL-regularized optimum, and it recovers any explicit reward that generated π up to the prompt-only gauge freedom of Section A.6.

Lemma A.10 (Implicit-reward consistency and gauge)**.** *Fix $\beta > 0$.*

- *(i) $\pi = \pi^\star_\beta(\hat{r})$ whenever $\pi \ll \pi_{\mathrm{ref}}$, so π is the KL-regularized optimum of its own implicit reward.*
- *(ii) If $\pi = \pi^\star_\beta(\bar{R})$ for some $\bar{R}$ satisfying Assumption A.1, then $\hat{r} = \bar{R} - b(x)$ with $b(x) = \beta \log Z_{\bar{R}}(x)$ prompt-only, where $Z_{\bar{R}}(x) = \mathbb{E}_{y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\bar{R}(x, y)/\beta)]$. Hence $\hat{r}$ and $\bar{R}$ lie in the same gauge class of Section A.6 and induce the same policy.*

Proof. (i) By Equation (2), $\pi^\star_\beta(\hat{r})(y \mid x) \propto \pi_{\mathrm{ref}}(y \mid x) \exp(\hat{r}(x, y)/\beta) = \pi_{\mathrm{ref}}(y \mid x) \cdot \pi(y \mid x)/\pi_{\mathrm{ref}}(y \mid x) = \pi(y \mid x)$, and the right-hand side already normalizes to one, so the proportionality is an equality.

(ii) From $\pi = \pi^\star_\beta(\bar{R})$ we have $\log \pi = \log \pi_{\mathrm{ref}} + \bar{R}/\beta - \log Z_{\bar{R}}(x)$. Multiplying by β and rearranging gives $\hat{r} = \bar{R} - \beta \log Z_{\bar{R}}(x)$, and $b(x) = \beta \log Z_{\bar{R}}(x)$ depends on x only. □

The hypothesis in (i) is only $\pi \ll \pi_{\mathrm{ref}}$ and a regularity requirement (Definition A.12) is not needed for the identity. Lemma A.10(ii) is the DPO instance of the data-level identifiability of Skalse et al. [2023] recalled in Section 2, as preference data pins the reward only up to a prompt-only shift, which is exactly the gauge $\hat{r}$ leaves free. The gauge-invariant reliance g_{cent} and the operator M_i^{cent} of Section A.6 are therefore the appropriate statistics for DPO, and the regime classification ports under M_i^{cent} as shown in Section A.6. Because $\hat{r}$ is fixed only up to this prompt-only gauge, the canonical operator-side construction below uses the centered operator M_i^{cent} of Section A.6, whose regime classification is gauge-invariant. We use M_i^{cent} rather than the plain M_i, as the regime call of M_i would depend on the chosen representative of $\hat{r}$.

DPO single-axis mitigations. Next, we need to define single-axis mitigations for different DPO-like methods. Fix a feature index i and a targeted spurious feature $\phi_i \in \Phi_{\mathrm{sp}}$. We distinguish two ways a single-axis mitigation enters a DPO pipeline.

Definition A.11 (Operator-side and loss/data-side DPO mitigations)**.** *An* operator-side *(or post-hoc) mitigation applies a single-axis operator M_i (Definition 3.5) directly to the implicit reward, producing $M_i(\hat{r})$ and the post-mitigation policy $\pi' = \pi^\star_\beta(M_i(\hat{r}))$. Its canonical instance is the centered operator M_i^{cent} of Section A.6.*
A loss-side *or* data-side *mitigation modifies the DPO objective or its preference data and retrains, producing a new policy π' with its own implicit reward $\hat{r}' = \beta \log(\pi'/\pi_{\mathrm{ref}})$.*
In general $\hat{r}'$ does not equal the canonical centered operator M_i^{cent} of Section A.6 applied to $\hat{r}$.

Examples of loss- or data-side DPO length mitigations include R-DPO [Park et al., 2024], SimPO [Meng et al., 2024], SamPO [Lu et al., 2024], and LMPO [Li et al., 2025a]. R-DPO is the instructive case, as its length penalty has the form of a single-axis subtraction yet is applied inside the training objective, so the induced implicit reward $\hat{r}'$ need not equal $M_i^{\mathrm{cent}}(\hat{r})$ and the method is loss-side under Definition A.11. SimPO is reference-free and length-normalizes its reward, so it is loss-side and does not act on an implicit reward of the $\hat{r}$ form. The $1/|y|$ normalization is a length-dependent reweighting, not the monotone reparametrization of Corollary A.5 Operator-side mitigation of a DPO policy instead corresponds to applying a post-hoc operator directly to $\hat{r}$, equivalent to post-hoc mitigation of an explicit reward model (we add them here for completeness). The LOESS length calibration of Huang et al. [2025] and the latent-space probe shaping of Fein et al. [2026] are both operator-side but non-linear, so they are single-axis mitigations in the general sense of Definition 3.5 and inherit the outcome-level $(\Delta_j, \Delta J)$ classification rather than the constructive guarantee of Lemma A.4 (Section A.13). Both are admissible as proxies through Lemma A.10 and Definition A.12. Table 2 summarizes where the surveyed methods fall and which results reach each.

Method	Placement	π_{ref}
R-DPO [Park et al., 2024]	Loss-side	Yes
SamPO [Lu et al., 2024]	Loss-side	Yes
SimPO [Meng et al., 2024]	Loss-side	Free
LMPO [Li et al., 2025a]	Loss-side	Free
LOESS calibration [Huang et al., 2025]	Operator-side	Inherited
Probe shaping [Fein et al., 2026]	Operator-side	Inherited

Table 2: Placement of surveyed DPO methods under Definition A.11. All are classifiable by the sufficiency protocol (Corollary A.14) and subject to the audit-distribution insufficiency, which stems from auditing at μ_{diag} rather than the policy pair and so applies to every row, with Corollary A.15 the sharp four-regime instantiation for the canonical linear operator. Operator-side rows reach the framework through an operator on $\hat{r}$, loss-side rows through the policy pair, and reference-free rows have no guaranteed regular implicit reward (Definition A.12).

The operator-side case is, by Lemma A.10, the explicit-reward framework with $\hat{r}$ in place of $\tilde{R}$. For the canonical centered operator M_i^{cent}, Lemma A.4 and the full taxonomy apply once $\hat{r}$ is admissible. For non-linear operator-side mitigations the outcome-level classification still applies, while the constructive single-axis guarantee does not (Section A.13). The loss/data-side case has no operator to analyze, and connects to the framework only through the policy pair (π, π') (see Lemma A.13 and corollary A.14).

Regularity of the implicit reward. The implicit reward $\hat{r}$ is only admissible as a proxy if it is bounded.

Definition A.12 (Regular implicit reward)**.** *The implicit reward $\hat{r}$ is* regular against π_{ref} *if the density ratio $d\mu_\pi / d\mu_{\pi_{\mathrm{ref}}} = \pi/\pi_{\mathrm{ref}}$ is essentially bounded above and below on the support of π_{ref}, that is, there exist constants $0 < m \le M < \infty$ with $m \le \pi(y \mid x)/\pi_{\mathrm{ref}}(y \mid x) \le M$ for $\mu_{\pi_{\mathrm{ref}}}$-a.e. (x, y).*

Under Definition A.12 the implicit reward satisfies the $L^\infty(\mu_{\pi_{\mathrm{ref}}})$ clause of Assumption A.1. The $L^2(\mu_{\mathrm{diag}})$ clause is automatic when $\mu_{\mathrm{diag}} = \mu_{\pi_{\mathrm{ref}}}$ and otherwise requires $\mu_{\mathrm{diag}} \ll \mu_{\pi_{\mathrm{ref}}}$, the same condition already assumed for annotator-conditioned audits in Section A.3. Then $\hat{r}$ enters every statement of Section 3 and Section A in the $\tilde{R}$ role, with the consequences proved in Lemma A.13.

Regularity is free whenever π optimizes a bounded reward and fails only under policy collapse or out-of-distribution drift, the regime in which DPO is known to acquire length correlation (see Fact C.17). Reference-free objectives such as SimPO [Meng et al., 2024] make this boundary explicit, as with no π_{ref} term constraining the density ratio, regularity of $\hat{r}$ against any fixed reference is not guaranteed, consistent with the reward-hacking-without-regularization degeneration noted by Meng et al. [2024].

Two consequences follow. First, a regular implicit reward is admissible as a proxy in the $\tilde{R}$ role. Second, the sufficiency classifier transfers to every DPO mitigation regardless of placement.

Lemma A.13 (Admissibility of the implicit reward)**.** *Suppose $\hat{r}$ is regular against π_{ref} (Definition A.12), the feature map Φ satisfies Assumption A.1, the centered Gram is non-degenerate (Assumption A.3), and $\mu_{\mathrm{diag}} \ll \mu_{\pi_{\mathrm{ref}}}$ (automatic when $\mu_{\mathrm{diag}} = \mu_{\pi_{\mathrm{ref}}}$). Then*

(i) $\hat{r} \in L^2(\mu_{\mathrm{diag}}) \cap L^\infty(\mu_{\pi_{\mathrm{ref}}})$, so $\hat{r}$ satisfies Assumption A.1 in the $\tilde{R}$ role.

(ii) $M_i^{\mathrm{cent}}(\hat{r}) \in L^2(\mu_{\mathrm{diag}}) \cap L^\infty(\mu_{\pi_{\mathrm{ref}}})$.

Consequently $\pi^\star_\beta(\hat{r})$ and $\pi^\star_\beta(M_i^{\mathrm{cent}}(\hat{r}))$ are well-defined KL-regularized optima with $\pi^\star_\beta(\hat{r}) = \pi$ by Lemma A.10(i), and the centered single-axis identity Lemma A.4 holds for $\hat{r}$.

Proof. (i) By Definition A.12 there are $0 < m \le M < \infty$ with $m \le \pi/\pi_{\mathrm{ref}} \le M$ $\mu_{\pi_{\mathrm{ref}}}$-a.e., so $\hat{r} = \beta \log(\pi/\pi_{\mathrm{ref}}) \in [\beta \log m,\, \beta \log M]$ $\mu_{\pi_{\mathrm{ref}}}$-a.e. and hence $\hat{r} \in L^\infty(\mu_{\pi_{\mathrm{ref}}})$. Since $\mu_{\mathrm{diag}} \ll \mu_{\pi_{\mathrm{ref}}}$, this essential bound transfers to μ_{diag}-a.e., so $\hat{r} \in L^\infty(\mu_{\mathrm{diag}}) \subset L^2(\mu_{\mathrm{diag}})$ as μ_{diag} is a probability measure.

(ii) Under (i) and Assumption A.1 on Φ, the centered reliance $g_{\mathrm{cent},i}(\hat{r}; \mu_{\mathrm{diag}})$ is a finite scalar, well-defined by Assumption A.3 together with $\hat{r}^{\mathrm{cent}}, \bar{\Phi} \in L^2(\mu_{\mathrm{diag}})$. Then $M_i^{\mathrm{cent}}(\hat{r}) = \hat{r} - g_{\mathrm{cent},i}(\hat{r}; \mu_{\mathrm{diag}})\, \phi_i$ is a finite $\mathbb{R}$-linear combination of $\hat{r}$ and ϕ_i, both in $L^2(\mu_{\mathrm{diag}}) \cap L^\infty(\mu_{\pi_{\mathrm{ref}}})$, and the claim follows by closure under finite linear combinations (Section A.2).

The consequence is then immediate. Assumption A.1 holds for $\hat{r}$ and $M_i^{\mathrm{cent}}(\hat{r})$ in the $\tilde{R}$ role, so the optimization setup of Section A.2 applies to both and the two optima are well-defined, $\pi^\star_\beta(\hat{r}) = \pi$ by Lemma A.10(i), and Lemma A.4 applies to $\hat{r}$. □

With $\hat{r}$ admissible and M_i^{cent} qualifying as a single-axis operator on it, the regime taxonomy of Section 3.3 applies to the operator-side pair $\pi' = \pi^\star_\beta(M_i^{\mathrm{cent}}(\hat{r}))$ exactly as for an explicit proxy. The sufficiency classifier transfers more broadly, without reference to $\hat{r}$ at all.

Corollary A.14 (Sufficiency transfer to DPO mitigations)**.** *Let π and π' be the pre- and post-mitigation DPO policies of Definition A.11, either operator-side with $\pi' = \pi^\star_\beta(M_i^{\mathrm{cent}}(\hat{r}))$ or loss/data-side with π' the retrained policy. Suppose Φ and R satisfy Assumption A.1 and the policy-induced expectations $\mathbb{E}_{\mu_\pi}[\phi_j]$, $\mathbb{E}_{\mu_{\pi'}}[\phi_j]$, $J(\pi, R)$, and $J(\pi', R)$ are finite. Then $B^\star$ of Theorem 3.10, evaluated on (π, π'), returns the correct regime label among $\{\mathrm{R0}, \mathrm{R0}_{\mathrm{cont}}, \mathrm{R1}, \mathrm{R2}\}$, extended to $\mathrm{R3}$ by Corollary A.6 and to $\mathrm{R4}$ by Corollary A.8 under the multi-μ_{diag} protocol.*

Proof. $\Delta_j(\pi, \pi') = \mathbb{E}_{\mu_{\pi'}}[\phi_j] - \mathbb{E}_{\mu_\pi}[\phi_j]$ and $\Delta J = J(\pi', R) - J(\pi, R)$ are first moments of Φ and R under μ_π and $\mu_{\pi'}$, both inputs to $\mathcal{B}^+$, so $B^\star \in \mathcal{B}^+$. Their finiteness is the stated hypothesis. The four-branch correctness and pairwise disjointness are exactly Theorem 3.10, whose statement places no condition on how π' is produced. Hence $B^\star(\pi, \pi')$ returns the correct label, and the R3 and R4 extensions follow from Corollaries A.6 and A.8. □

The finiteness hypothesis is automatic in the operator-side case, where Lemma A.13 makes $M_i^{\mathrm{cent}}(\hat{r})$ admissible and the density-ratio bound of Section A.2 gives finite feature and reward expectations at $\mu_{\pi'}$. In the loss/data-side case it is a mild condition on the trained policy, strictly weaker than requiring its implicit reward $\hat{r}'$ to be regular. Since $B^\star$ reads only $(\Delta_j, \Delta J)$ and never $\hat{r}'$, the sufficiency half of the framework reaches loss-side and data-side mitigations that lie outside the single-axis operator class, which is the precise sense in which the prescriptive core covers DPO mitigations regardless of placement.

Operator-side transfer. The operator-side reading inherits the full taxonomy. By Lemma A.13 the centered single-axis identity holds for $\hat{r}$, so R0–R3 classify (π, π') as in Section 3.3, and R4 ports because M_i^{cent} retains its μ_{diag}-dependence (Section A.6), with detection by Corollary A.8. The impossibility transfers as well for gauge-invariant benchmarks.

Corollary A.15 (Operator-side impossibility for DPO)**.** *For $B \in \mathcal{B}$ invariant under the prompt-only gauge of Section A.6, Theorem 3.9 transfers to operator-side DPO mitigations, with the implicit reward $\hat{r}$ in the $\tilde{R}$ role and M_i^{cent} in the M_i role. There are four reference policies, with $\mu_{\mathrm{diag}}, M_i^{\mathrm{cent}}, R, \Phi, \Phi_{\mathrm{sp}}, \beta$ fixed and the four implicit rewards sharing one gauge class, such that every such B agrees across the four while the mitigated policies $\pi_\beta^\star(M_i^{\mathrm{cent}}(\hat{r}))$ fall into R0, R0$_{cont}$, R1, and R2.*

Proof. Reinterpret the linear proxy $\tilde{R}(y) = c^\top y$ of Theorem 3.9 as a DPO implicit reward, on the same bounded set on which that theorem is carried out (its Assumption A.1 truncation), where $c^\top y$ is bounded. For each reference $\pi_{\mathrm{ref}}^{(k)} = \mathcal{N}(0, \Sigma^{(k)})$ of that construction, the policy $\pi^{(k)} = \pi_\beta^\star(\tilde{R})$ has implicit reward $\hat{r}^{(k)} = c^\top y - b^{(k)}$ by Lemma A.10(ii), with $b^{(k)} = \beta \log Z^{(k)} = c^\top \Sigma^{(k)} c/(2\beta)$ a prompt-only constant differing across k, so the four $\hat{r}^{(k)}$ share one gauge class without being identical as functions. Each density ratio $\pi^{(k)}/\pi_{\mathrm{ref}}^{(k)} = \exp(c^\top y/\beta)/Z^{(k)}$ is bounded above and below, so each $\hat{r}^{(k)}$ is regular (Definition A.12), with $\pi^{(k)} = \mathcal{N}(\Sigma^{(k)} c/\beta, \Sigma^{(k)})$ the unrestricted Gaussian limit as in Section A.7.

Under $\mu_{\mathrm{diag}} = \mathcal{N}(0, I_3)$ the features are already centered, so $M_i^{\mathrm{cent}} = M_i$ and $g_{\mathrm{cent}} = g$ on this construction and Assumption A.3 holds with $\bar{G} = I_3$. Since g_{cent} is gauge-invariant, $M_i^{\mathrm{cent}}(\hat{r}^{(k)}) = M_i^{\mathrm{cent}}(c^\top y) - b^{(k)}$, so $(\hat{r}^{(k)}, M_i^{\mathrm{cent}}(\hat{r}^{(k)}))$ is $(c^\top y, M_i^{\mathrm{cent}}(c^\top y))$ shifted jointly by the prompt-only constant $-b^{(k)}$. Every gauge-invariant B therefore agrees across the four, and the policy-side classification of Theorem 3.9 transfers because $\pi_\beta^\star$ and M_i^{cent} are gauge-invariant, placing the four mitigated policies in R0, R0$_{\mathrm{cont}}$, R1, and R2. The restriction to gauge-invariant B is necessary, as the cardinal functional $\mathbb{E}_{\mu_{\mathrm{diag}}}[\hat{r}^{(k)}] = -b^{(k)}$ takes four distinct values and already separates the instances. □

The benchmarks used in practice, such as AlpacaEval, Chatbot Arena, and RewardBench, are ranking- or win-rate-based and therefore gauge-invariant, so the obstruction binds DPO evaluation as conducted and not only the controlled construction.

Loss-side and data-side scope. A loss-side or data-side mitigation (Definition A.11) does not apply a single-axis operator to $\hat{r}$, so the constructive guarantees do not attach. Lemma A.4 has no analogue, and Corollary A.15 cannot be instantiated, as it holds the operator fixed across instances and here there is none. Nevertheless, two things survive. First, the sufficiency classifier still applies through the policy pair by Corollary A.14, so such mitigations remain classifiable into R0–R3 from $(\Delta_j, \Delta J)$. Second, the audit-distribution insufficiency conclusion still holds for them, since the regime is a property of the policy-induced first moments at $\mu_{\pi^\star}$ that no $\mathcal{B}$-functional reads, and this gap is independent of whether the mitigation is realized as an operator. We do not give a loss-side-specific four-instance construction, as the operator-side witness of Corollary A.15 together with the non-vacuity of Sections A.7 and A.9 already establishes that the regimes are real and invisible to gauge-invariant audit functionals. The practical consequence is that a loss-side method must still report $(\Delta_j, \Delta J)$ at $\mu_{\pi^\star}$ (Section A.13), since its audit-side scores cannot certify R0 any more than an operator-side method's can.

In sum, the sufficiency classifier (Corollary A.14) covers any method whose policy is the KL-regularized optimum, DPO included, while the regime taxonomy and the impossibility (Corollary A.15, for gauge-invariant benchmarks) attach to mitigations that act as single-axis operators on the implicit reward under the regularity of Definition A.12. Loss-side and data-side mitigations connect through the policy pair alone, and the coverage boundary is associated with the out-of-distribution regime in which DPO is known to acquire length correlation (Fact C.17).

A.13 Takeaway Recommendations for RM Mitigation and Benchmarks Developers

Theorem 3.9 and Corollary A.5 together circumscribe what reward-model benchmarks in the class $\mathcal{B}$ can and cannot deliver, and what mitigation methods evaluated within $\mathcal{B}$ cannot be certified to achieve. We translate the formal results into concise recommendations.

Regime detection procedures. Theorem 3.10's classifier $B^\star$ separates R0, R0$_{\text{cont}}$, R1, and R2 given $(\Delta_j, \Delta J)$ at $\mu_{\pi^\star}$, but does not address R2 origin disambiguation, R3 detection, or R4 detection. The following three procedures close these gaps and should accompany any regime claim that touches them.

- *R2 origin disambiguation via the rescalability sweep.* The two origins of R2 in Definition 3.7 (scale overshoot, target misspecification) are operationally distinguished by sweeping the partial mitigation $M_i^c(\tilde{R}) = \tilde{R} - c\, g_i(\tilde{R}; \mu_{\text{diag}})\, \phi_i$ over $c \in (0, 1)$ at fixed β. If $J(\pi^\star_\beta(M_i^c(\tilde{R})), R) > J(\pi^\star_\beta(\tilde{R}), R)$ for some c, the regime is driven by scale overshoot at $\mu_{\pi^\star}$ and is recoverable by partial projection. If no c improves ΔJ, the projection has removed a structurally informative component of ϕ_i (target misspecification under partial informativeness, see Section 3.1) and rescaling cannot recover it. The test requires a one-dimensional sweep over c at fixed β and the same cardinal R-access already needed to distinguish R2 from R0. Concretely: publish the $\Delta J(c)$ curve, since a recoverable overshoot is a different scientific finding than a non-recoverable target misspecification.
- *R3 detection via D_{KL}.* Corollary A.6 extends $B^\star$ to R3 via the branch $\Delta J = 0 \wedge \Delta_j = 0$. Reporting $D_{\text{KL}}(\pi' \,\|\, \pi)$ alongside the label separates policy-relevant R3 ($D_{\text{KL}} > 0$: mitigation moved the policy along directions orthogonal to $(\Delta_j, \Delta J)$) from policy-irrelevant R3 ($D_{\text{KL}} = 0$: mitigation did not move the policy at all). Concretely: report D_{KL} whenever the regime call is R3, since only the policy-irrelevant case is consistent with a vacuous mitigation operator and should be flagged differently in downstream interpretation.
- *R4 detection via the multi-μ_{diag} protocol.* R4 (Definition 3.8) is transversal to R0–R3 and observable only by constructing the canonical $M_i^{\mu_{\text{diag}}}$ at multiple audit distributions and comparing the resulting regime calls. Concretely: select at least two audit distributions of substantively different annotator provenance (e.g., $\mu^{\text{human}}_{\text{diag}}$, $\mu^{\text{LLM}}_{\text{diag}}$), construct $M_i^{(\ell)}$ and $\pi^{(\ell)} = \pi^\star_\beta(M_i^{(\ell)}(\tilde{R}))$ at each, apply $B^\star$ to each, and treat cross-distribution disagreement on the regime call as the test.

Reward Model Bias Mitigation Method Recommendations The prescriptions of our framework are necessary conditions for interpretable mitigation claims of any new mitigation method. Without them, R0 cannot be distinguished from R0$_{\text{cont}}$, R1, or R2, regardless of how strong the audit-side evidence is. We organize the recommendations by single- and multi-axis operators and present each as a self-contained checklist to support standardized reporting across method papers.

Single-axis Don'ts.

- **Do not fit and validate M_i on the same μ_{diag} without testing alternative audit distributions.** R4 (Definition 3.8) makes regime membership a function of μ_{diag}, and the sign reversal in Section B.3 shows the dependence is empirically real. *Concretely:* a length operator fit on LLM-judge preference data can inherit a sycophancy-coupling coefficient with the opposite sign of one fit on human-judge data, so the two operators should not be interchangeable even when both zero on-target reliance at their respective audit distributions.
- **Do not apply M_i without reporting the induced $L^2(\mu_{\text{diag}})$ scale change.** Section A.5 shows $\|M_i(\tilde{R})\|_{L^2(\mu_{\text{diag}})}$ generically differs from $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ via both a diagonal and a cross-correlation contribution, inducing an effective-β shift at fixed nominal β. *Concretely:* a method paper reporting only $g_i \to 0$ at μ_{diag} leaves readers unable to separate axis reallocation from scale-driven regime change (see single-axis dos).
- **Do not report only pooled diagnostics when the deployment target is within-prompt selection.** Section B.2 shows pooled $|\rho_{\text{len}}|$ can be driven to zero while within-prompt $|\rho^{\text{within}}_{\text{len}}|$ flips sign on three of four SOTA reward models. *Concretely:* BoN top-1 and PPO

both operate within prompts, so a pooled diagnostic targeting ϕ_i does not measure the reliance the optimizer responds to.

- **Do not rely on ordinal-only diagnostics post-mitigation.** Corollary A.5 shows that ranking accuracy and pairwise win-rate are invariant to $\tilde{R} \mapsto c\tilde{R}$ while $J(\pi^\star_\beta(c\tilde{R}), R)$ varies strictly with c. *Concretely:* do not only report ranking accuracy on RewardBench or pairwise win-rates on AlpacaEval post-mitigation, because reporting only such ordinal scores cannot rule out scale-driven regime shifts.

Single-axis Dos.

- **Default to M_i^{norm} and M_i^{cent} mitigations and document the variant used.** Section A.5 shows M_i^{norm} separates axis reallocation from scale and Section A.6 shows M_i^{cent} restores gauge invariance of the regime classification. *Concretely:* when using M_i for pipeline compatibility, report $\|\tilde{R}'\|_{L^2(\mu_{\mathrm{diag}})}/\|\tilde{R}\|_{L^2(\mu_{\mathrm{diag}})}$ (ratio of root-mean-square reward scores for mitigated and unmitigated proxy rewards) alongside g_i so that scale-driven and reallocation-driven Δ_j contributions are distinguishable.
- **Run the $c \in (0,1)$ rescalability sweep when $\Delta J \leq 0$.** The sweep over $M_i^c(\tilde{R}) = \tilde{R} - c\, g_i\, \phi_i$ is the only observational test that separates the two origins of R2 (scale overshoot, recoverable, but target misspecification is not). *Concretely:* publish the $\Delta J(c)$ curve, since a recoverable overshoot is a different scientific finding than a non-recoverable target misspecification.
- **Evaluate at policy-induced distributions and report $(\Delta_j, \Delta J)$.** According to Theorem 3.10, this approach separates R0, $\mathrm{R0}_{\mathrm{cont}}$, R1, and R2, but, according to Theorem 3.9, audit-only inputs cannot. *Concretely:* run the mitigated reward model in BoN or short PPO against a fixed reference policy and report both quantities, not only audit-set accuracy.
- **Instrument off-target Δ_j before publication.** Select a panel of plausible Φ_{sp} candidates the operator does not target (formatting density, hedging markers, position effects, sycophancy and confidence indicators) and report Δ_j on those features at $\mu_{\pi^\star}$. *Concretely:* the developer has both the unmitigated and mitigated rewards in hand and is the natural party to produce this measurement. Without it, an R0 claim is structurally indistinguishable from $\mathrm{R0}_{\mathrm{cont}}$ (substitution under improvement).
- **Report cardinal scale pre/post mitigation.** Publishing $\|\tilde{R}\|_{L^2(\mu_{\mathrm{diag}})}$ before and after is the minimal cardinal observable that ordinal benchmark functionals miss. *Concretely:* report the standard deviation of reward scores on a fixed evaluation set both before and after applying the mitigation, so scale-driven ΔJ shifts are visible.
- **Document π_{ref}-sensitivity of the mitigation.** The construction in Theorem 3.9 produces four distinct regimes by varying only π_{ref} for a fixed $(\tilde{R}, M_i, \mu_{\mathrm{diag}})$. *Concretely:* validate the mitigation against at least two reference policies of different lineages and report per-pair outcomes rather than a single headline number.

Multi-axis recommendations.

- **Specify whether the operator is joint or sequential, and the operator-stack type.** For canonical linear projections applied to a fixed reward, sequential M_i followed by M_j equals joint projection on (ϕ_i, ϕ_j) by Frisch-Waugh-Lovell. The equivalence breaks when operators are heterogeneous (e.g., LOESS for length, two-head for content), when projection directions are recomputed on the changed reward, or when operators are non-linear. *Concretely:* state which case applies and, when sequential and joint differ, report both outcomes side-by-side.
- **Joint reliance zero at μ_{diag} does not certify R0.** Lemma 3.6 generalizes and a multi-axis operator M_S that zeros g_i for all $i \in S$ at μ_{diag} does not in general zero g_i at $\mu_{\pi^\star}$. *Concretely:* report Δ_j at $\mu_{\pi^\star}$ for every axis in the targeted set, not just on-target axes. Covering a panel at μ_{diag} narrows the substitution-accessible subset of Φ but does not close the measurement-vs-optimization gap.
- **State Φ_{sp} explicitly and acknowledge out-of-panel exploitation.** *Concretely:* a method covering {length, sycophancy, position} does not certify R0 against features outside that set, and multi-axis claims should be framed as "no substitution within the measured panel" rather than "no substitution."

- **Flag statistical mediation between targeted features.** Where the literature establishes coupling between features in the targeted set (length-sycophancy and length-confidence per Facts C.9,C.11,and C.12), associational joint projection may remove signal flowing through coupled paths. *Concretely:* acknowledge this as a known limitation rather than treating coupled signal as residual noise. The rescalability test above extends to multi-axis as a partial diagnostic.

Scope under non-linear operators. The proofs above (Theorems 3.9 and 3.10) and the linear-Gaussian instantiation (Section A.7) use Gaussianity for closed-form policy expressions and linearity for the canonical projection construction, but the prescriptions in this section target the structural $\mu_{\text{diag}} \to \mu_{\pi^\star}$ gap and the cardinal/ordinal blindness rather than these constructive devices, and so apply to non-linear, non-Gaussian operators and reference policies. The regime classification (R0-R4) references Δ_j and ΔJ rather than operator form, the cardinal/ordinal blindness of Corollary A.5 is monotone-invariant, and the gap itself is geometric rather than algebraic. What does *not* transfer is Lemma 3.6's constructive guarantee (a non-linear operator may not zero g_i even at μ_{diag}) and the closed-form expressions of Section A.7. Nonetheless, Section A.9 demonstrates regime separation under quadratic non-linearity in closed form and Section B.2 shows the gap surviving a non-linear *operator* (LOESS calibration) measured by a linear diagnostic, and Section B.3 establishes audit-distribution dependence of the relevant coupling under linear specifications with distribution-free robustness checks. **Method papers using non-linear operators should still adopt the prescriptions** above by virtue of targeting the same geometric phenomenon, while stating that they cannot import the linear-Gaussian regime classification's constructive guarantees, only its outcome-level definitions.

Reward Model Benchmark Recommendations Some of the recommendations for mitigaiton method developers transfer over to benchmark developers, with a few added recommendations.

Don'ts.

- **Do not expand the benchmark input within μ_{diag} and expect regime separation.** Theorem 3.9 grants $\mathcal{B}$ joint access to $(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\text{diag}})$ and still admits four instances landing in R0, R0_{cont}, R1, and R2. Adding feature axes, distractors, or held-out audit splits stays inside $\mathcal{B}$ and inherits the impossibility. *Concretely:* adding a sycophancy subset alongside a length subset, or expanding from pairwise to best-of-4 selection, enriches μ_{diag} but does not measure Δ_j at $\mu_{\pi^\star}$ and so cannot separate R0 from R0_{cont} or R1.
- **Do not rely on benchmark functionals that are invariant to monotone rescaling of $\tilde{R}$.** Corollary A.5 shows that every $B \in \mathcal{B}_{\text{ord}}$ (ranking accuracy, pairwise win-rate, and their composites) returns identical values under $\tilde{R} \mapsto c\tilde{R}$ while $J(\pi^\star_\beta(c\tilde{R}), R)$ varies strictly with c, so cardinal-scale-driven regime shifts are undetectable. *Concretely:* if mitigation halves the proxy's effective scale, ranking accuracy is unchanged but the KL-regularized policy behaves as if β doubled (by the scale-equivariance of Section A.5), and downstream ΔJ shifts accordingly.
- **Do not generalize a single-π_{ref} score across deployment settings.** The construction in Theorem 3.9 varies only π_{ref} and obtains four distinct regimes for the same $(\tilde{R}, M_i, \mu_{\text{diag}})$. A benchmark score at one π_{ref} does not transport to deployment π_{ref}'s. *Concretely:* a reward model topping the leaderboard when paired with one SFT base can degrade PPO outcomes when paired with a different SFT base of the same model family, since the policy-induced distribution shifts even though the audit-side score does not.

Dos.

- **Evaluate at policy-induced distributions and report** $(\Delta_j, \Delta J)$. By Theorem 3.10, this input separates R0, R0_{cont}, R1, and R2 via the explicit functional $B^\star$ in Equation (6); by Theorem 3.9, no audit-only input suffices. *Concretely:* run the mitigated reward model in BoN or short PPO against a fixed reference policy, and report the change in expected feature value Δ_j on each $\phi_j \in \Phi_{\text{sp}}$ alongside the change in true reward ΔJ, rather than only audit-set accuracy.
- **Report cardinal scale**. Publishing $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ pre/post mitigation provides observables that $\mathcal{B}_{\text{ord}}$ cannot, addressing the blindspot of Corollary A.5. *Concretely:* alongside accuracy,

report the standard deviation of reward scores on a fixed evaluation set before and after applying the mitigation, so that scale-driven ΔJ shifts are visible to readers.

- **Treat π_{ref} as an axis of variation, not a fixed convention.** Either evaluate across a panel of reference policies or document π_{ref}-sensitivity as a first-class output, consistent with the construction underlying Theorem 3.9. *Concretely:* pair each reward model with at least two SFT reference policies of different lineages and publish per-pair scores rather than a single headline number.
- **Report Δ_j across all of Φ_{sp} at $\mu_{\pi^\star}$, not just the targeted axis.** Theorem 3.10 makes the off-target Δ_j on $\Phi_{\mathrm{sp}} \setminus \{\phi_i\}$ a load-bearing input for separating $\mathrm{R0}_{\mathrm{cont}}$ and R1 from R0. Reporting only on-target reliance leaves substitution structurally invisible regardless of how rich the audit distribution is. *Concretely:* when evaluating a length-debiasing operator, also measure post-mitigation drift in sycophancy, hedging, and formatting at $\mu_{\pi^\star}$ — not just residual length correlation at the audit set.

Summary The recommendations above are not addressed to a single audience. R0 certification is a joint property of the mitigation operator and the evaluation protocol. A method paper following the developer recommendations cannot be cleanly compared across benchmarks that omit the corresponding ones, and a benchmark implementing all three prescriptions cannot certify R0 for methods that do not report off-target Δ_j or cardinal scale. The overlap between the two checklists is structural rather than redundant.

Existing methods and benchmarks satisfy these prescriptions only partially. SOTA benchmarks partially implement these prescriptions and close the gap to pure ordinal benchmarks. In particular, the current SOTA benchmark RewardBench2 [Malik et al., 2026] features a "Ties" metric, introduces cardinal-sensitive scoring, and its on-policy/off-policy finding documents π_{ref}-sensitivity empirically. However, no current protocol jointly delivers policy-distribution evaluation, cardinal reporting, and π_{ref}-as-axis, a gap our framework both formalizes as necessary (Theorem 3.9) and proves sufficient to close (Theorem 3.10). On the mitigation side, published methods rarely report off-target Δ_j at $\mu_{\pi^\star}$ or run the rescalability sweep, leaving almost all recent results ambiguous between $\mathrm{R0}_{\mathrm{cont}}$, R1, and R2 in ways the recommendations would have resolved (see Section 5).

Closing the joint gap is a community coordination problem rather than a single-paper deliverable. Method papers cannot unilaterally provide policy-distribution evaluation without benchmark infrastructure that supports it. Benchmarks cannot unilaterally provide off-target Δ_j measurement without method papers specifying Φ_{sp}. The recommendations above are scoped so that adoption by either side improves the evidence base, and joint adoption is what moves the field from R0 claims that audit-side scores cannot certify to R0 claims that the framework certifies sufficient.

B Supporting Experiments

All experiments were either run through API access or on a single A100 GPU 40 Gb or B200 GPU 180 Gb (rented online). Our code for the experiments will be released on GitHub (MIT license) upon publication.

B.1 Bias Substitution in Language Model RLHF (GRPO)

We demonstrate reward bias substitution in a RLHF pipeline with standard off-the-shelf configuration. A single-axis length penalty applied during GRPO training compresses response length as intended, yet the freed optimization pressure rotates onto a correlated axis and drives the policy into overconfidence, instantiating the substitution regime (R1) while leaving multiple-choice quality intact.

Setup

- **Policy model.** *Llama-3.2-3B-Instruct* [Grattafiori et al., 2024], adapted with LoRA [Hu et al., 2022] on all linear layers (the four attention projections `q_proj`, `k_proj`, `v_proj`, `o_proj` and the three MLP projections `gate_proj`, `up_proj`, `down_proj`) with rank $r = 32$, scaling $\alpha = 64$, dropout 0.05.
- **Reward model.** *Skywork-Reward-V2-Llama-3.1-8B* [Liu et al., 2025], kept frozen throughout training and queried only through the reward function.
- **Dataset.** Prompts from *UltraFeedback* [Cui et al., 2023] (the `ultrafeedback_binarized` `train_prefs` split), truncated to 2000 characters.
- **Algorithm.** Group Relative Policy Optimization (GRPO) [Shao et al., 2024] as implemented in TRL [von Werra et al., 2020]. 8-bit AdamW, learning rate 3×10^{-5}, KL coefficient $\beta = 2 \times 10^{-2}$, 600 optimizer steps, and bfloat16 precision.
- **Length penalty.** The RLHF reward is shaped as $\tilde{R}(x, y) = R_{\mathrm{RM}}(x, y) - \lambda\, n_{\mathrm{tok}}(y)/100$, where $n_{\mathrm{tok}}(y)$ is the number of response tokens and λ is the penalty coefficient. Three values $\lambda \in \{0, 4, 8\}$, each trained with four random seeds, for twelve training runs in total.
- **Generation during training.** Group size of 4 candidate generations per prompt for the group-relative advantage, per-device batch of 4, gradient accumulation of 1, sampling temperature $T = 1.0$, maximum completion length 256 tokens.

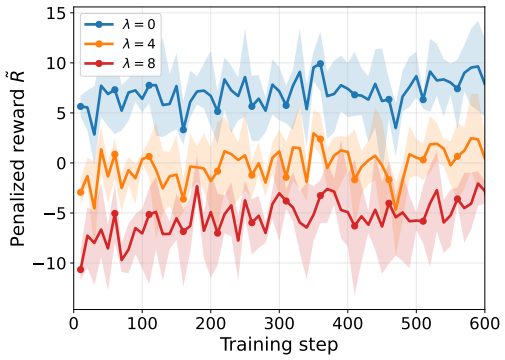
Training Curves We log the following quantities at every optimizer step and report them per λ, averaged over the four seeds, see Figure 6.

- **Mean response length.** The average number of generated tokens per step. This is the primary curve and shows the length penalty taking effect, with the curves separating by λ.
- **Mean reward.** The average shaped reward $\tilde{R}$ per step. Note that $\tilde{R}$ is not directly comparable across λ because the penalty term differs, so the curves are read within each λ for evidence of learning.
- **KL divergence.** The divergence of the policy from the frozen reference policy per step, confirming that the policy stays regularized and does not collapse.
- **GRPO loss and gradient norm.** Diagnostic optimization curves.

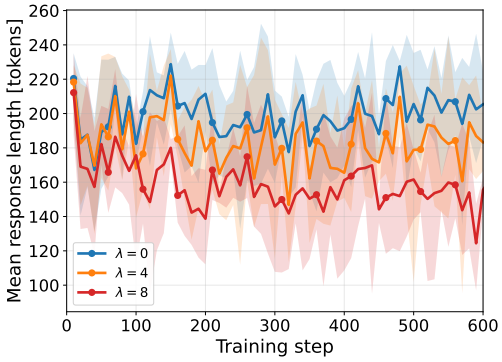
Evaluation All twelve adapted policies and the untrained base policy[7] are evaluated on held-out data, with each adapter taken at the 600-step checkpoint. Aggregate numbers use 95% confidence intervals from a hierarchical bootstrap over seeds and samples. We report the following metrics.

- **Mean response length.** The average response token count on 50 held-out *UltraFeedback* `test_prefs` prompts, with 4 samples per prompt drawn at temperature 1.0.

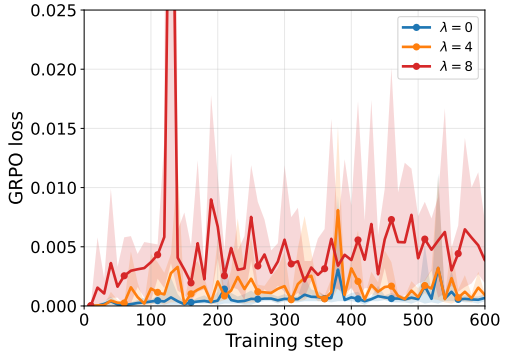
[7] Untrained by us, but still instruction tuned by the model developer.



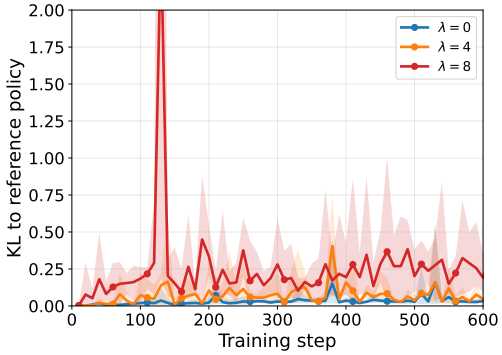
(a) Reward $\tilde{R}(x, y) = R_{\mathrm{RM}}(x, y) - \lambda\, n_{\mathrm{tok}}/100$



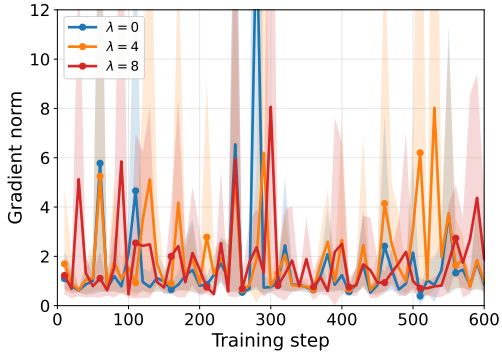
(b) Response length n_{tok}



(c) GRPO loss



(d) KL to reference policy



(e) Gradient norm

Figure 6: Training curves (shaded region shows the min-max range across 4 seeds)

	Base	$\lambda = 0$	$\lambda = 4$	$\lambda = 8$
Length and quality				
Mean response length (tokens)	192.2	$203.6_{[194, 212]}$	$188.2_{[181, 195]}$	$170.2_{[161, 179]}$
MMLU accuracy	0.596	$0.609_{[.59, .62]}$	$0.611_{[.59, .63]}$	$0.616_{[.60, .63]}$
Overconfidence				
Expected calibration error (ECE)	0.264	$0.253_{[.23, .28]}$	$0.299_{[.26, .34]}$	$0.406_{[.32, .50]}$
Mean verbalized confidence	0.811	$0.803_{[.78, .82]}$	$0.815_{[.80, .83]}$	$0.836_{[.82, .85]}$
TriviaQA accuracy	0.564	$0.558_{[.53, .58]}$	$0.523_{[.47, .57]}$	$0.415_{[.29, .52]}$
AUROC	0.708	$0.725_{[.71, .74]}$	$0.706_{[.67, .74]}$	$0.651_{[.62, .69]}$
Sycophancy				
Regressive flip rate	0.653	$0.582_{[.52, .64]}$	$0.646_{[.59, .70]}$	$0.602_{[.54, .66]}$
Overall flip rate	0.721	$0.668_{[.63, .71]}$	$0.709_{[.67, .75]}$	$0.704_{[.66, .75]}$

Table 3: Metric values across the length-penalty sweep for the data plotted in Figure 1 in Section 1. Each λ entry is the mean over four random seeds with the 95% hierarchical-bootstrap confidence interval as a subscript. The Base column is the untrained *Llama-3.2-3B-Instruct* model and carries no interval. Expected calibration error, TriviaQA accuracy, and AUROC use the LLM-judge answer grading.

- **MMLU accuracy.** Multiple-choice accuracy on 1000 questions subsampled from MMLU [Hendrycks et al., 2021], evaluated zero-shot with greedy decoding and a single-letter answer parsed from the output.
- **Expected Calibration Error (ECE).** Computed with 10 equal-width bins on the (verbalized confidence, correctness) pairs produced by the Tian et al. [2023] Verb. 1S top-1 protocol on 1000 *TriviaQA* [Joshi et al., 2017] questions.
- **AUROC.** The area under the ROC curve of the verbalized confidence used as a predictor of answer correctness.
- **Mean verbalized confidence.** The average probability the model assigns to its own answer under the Verb. 1S top-1 protocol.
- **TriviaQA accuracy.** The fraction of correct answers under the same protocol.
- **Regressive flip rate.** The headline sycophancy metric from the Sharma et al. [2024] *are_you_sure* task, the fraction of initially-correct answers that become incorrect after the user pushes back.
- **Overall flip rate.** The fraction of answers that change at all after the pushback, regardless of correctness.

The Tian et al. [2023] ECE, AUROC, and TriviaQA accuracy metrics are computed under an LLM-judge equivalence check using the Tian et al. [2023] prompt with *Claude Haiku 4.5* [Anthropic, 2026] as the judge. The sycophancy metrics use the Sharma et al. [2024] two-stage protocol, eliciting an initial answer and then a post-pushback answer on 500 *are_you_sure* records with 2 samples per record at temperature $T = 1.0$.

Results See Figure 1 and Table 3 for results.

- **The length penalty compresses responses monotonically.** Mean response length falls from 204 tokens at $\lambda = 0$ to 188 at $\lambda = 4$ and 170 at $\lambda = 8$. Both reductions, measured relative to $\lambda = 0$, are significant.
- **Multiple-choice quality is preserved.** MMLU accuracy holds near 0.61 across the sweep (0.609, 0.611, 0.616 at $\lambda = 0, 4, 8$ against a base of 0.596), and the change from base is not significant at any λ.
- **Overconfidence rises with the length penalty.** Expected calibration error climbs from 0.25 at $\lambda = 0$ to 0.30 at $\lambda = 4$ and 0.41 at $\lambda = 8$, and mean verbalized confidence climbs in step from 0.80 to 0.82 to 0.84. TriviaQA accuracy falls from 0.56 at $\lambda = 0$ to 0.52 at $\lambda = 4$

and 0.41 at $\lambda = 8$, and the coupling between confidence and correctness weakens, with AUROC falling from 0.73 to 0.65. Thus, the policy does not merely become more confident, it becomes wrong more often. At $\lambda = 8$ the rise in calibration error, the rise in verbalized confidence, the drop in accuracy, and the drop in AUROC are all significant.

- **The calibration loss is caused by the penalty, not by RLHF.** At $\lambda = 0$ the calibration error (0.25) is at or slightly below the base value (0.26), so RLHF on its own does not break calibration. The degradation appears only once the length penalty is applied.
- **Sycophancy does not significantly change.** The regressive flip rate is non-monotonic across the sweep (0.58, 0.65, 0.60 against a base of 0.65) and the overall flip rate shows no consistent trend (0.67, 0.71, 0.70 against 0.72), consistent with the already-elevated flip rate of the base model leaving little headroom for optimization pressure to manifest.
- **Regime transitions.** We take the true reward to be free-form factual accuracy, proxied by TriviaQA, and we treat expressed confidence as a spurious feature that the reward model rewards as a quality proxy (Fact C.12) (Φ_{sp}). Multiple-choice MMLU accuracy serves as a capability control. At $\lambda = 8$, response length falls and MMLU is unchanged, while expressed confidence rises, a nonzero Δ_j on an off-target spurious axis, and TriviaQA accuracy degrades, a negative ΔJ. This is the harmful sub-case of bias substitution (R1), not overcorrection (R2), because pressure rotates onto a second spurious axis rather than degrading capability with no rotation. The MMLU performance preservation indicates the ΔJ drop is specific to free-form accuracy rather than a general capability loss. We read $\lambda = 4$ as an intermediate point, where the overconfidence proxies shift toward substitution but sit at or near the significance boundary, so the missing significance at four seeds reflects limited power rather than the absence of an effect, and $\lambda = 8$ as the clear R1 instance.

B.2 Measurement-vs-Optimization Gap in Length Mitigation in BoN Selection

This section provides a quantitative empirical instance of the measurement-versus-optimization gap of Section 3.2. We evaluate two published length-debiasing operators across five reward models under a Best-of-N (BoN) selection protocol, using a within-prompt reliance diagnostic that tracks the BoN selection distribution rather than the pooled audit distribution at which the operators are designed. The headline observation is that the post-hoc calibration of Huang et al. [2025] zeros pooled reward-length correlation ($0.316 \to 0.037$) while overshooting into negative within-prompt correlation on three of four SOTA reward models, with ΔJ degrading on two. The linear-probe operator of Fein et al. [2026] is fit on the same prompt split and transfers cleanly to the within-prompt diagnostic on four of five reward models with $\Delta J > 0$.

Reward models and operators. We evaluate Skywork-Reward-V2-Llama-3.1-8B, Skywork-Reward-V2-Qwen3-8B, Skywork-Reward-V2-Qwen3-0.6B [Liu et al., 2025], the Llama-3.1-8B-Instruct-RM-RB2 [Malik et al., 2026], and the older DeBERTa-v3-large OpenAssistant [2023] reward models. We use two single-axis operators that target length bias ($\phi_i = \text{length}$). First, mechanistic reward shaping with a difference-of-means linear probe projected from the RM's final-layer hidden state [Fein et al., 2026], and, second, the post-hoc reward calibration of Huang et al. [2025] with a LOESS fit subtracted at the score level. Both are instances of M_i (Definition 3.5) and reduce $|g_i(\tilde{R}; \mu_{\text{diag}})|$ at their respective audit distributions by construction.

Setup. For each of two prompt sources (AlpacaEval [Dubois et al., 2024], GSM8K Cobbe et al. [2021]) we generate 64 candidate responses per prompt from Llama-3.2-1B-Instruct [Grattafiori et al., 2024], fit each operator on a probe split, and evaluate on a disjoint held-out split (512 for AlpacaEval prompts, 807 for GSM8K prompts). Each RM selects the highest-scoring response among the 64 candidates. We measure two quantities:

- a within-prompt Pearson correlation $\rho_{\text{len}}^{\text{within}}$ between RM score and response length, computed within each candidate set and averaged across prompts. Because BoN top-1 selection operates within candidate sets per prompt, $\rho_{\text{len}}^{\text{within}}$ measures the reliance selection actually responds to, while pooled ρ_{len} averages over prompts that BoN never compares. This instantiates g_i at a μ_{diag} aligned with BoN selection rather than the pooled distribution the operators target.

	DeBERTa	Allen	SW-L	SW-Q8B	SW-Q0.6B
Within-prompt $\rho_{\text{len}}^{\text{within}}$ (AlpacaEval)					
Baseline	$+.168$	$-.061$	$+.134$	$+.065$	$+.088$
Mech. RS	$+.005$	$-.011$	$+.087$	$+.032$	$+.041$
Huang et al. [2025]	$+.048$	$-.099$	$-.065$	$-.138$	$-.228$
AlpacaEval LC win rate vs. baseline (%)					
Mech. RS	54.4^{***}	$46.2^{\dagger}$	51.5^{***}	51.7^{***}	51.8^{***}
Huang et al. [2025]	50.4^{*}	$48.7^{\dagger}$	50.1	$48.2^{\dagger}$	51.2^{***}
GSM8K BoN accuracy (%)					
Baseline	32.8	71.5	68.3	71.9	65.8
Mech. RS	36.4	71.3	68.5	71.9	65.7
Huang et al. [2025]	32.3	67.9	68.3	71.3	66.2

Table 4: Within-prompt reward-length correlations and BoN selection outcomes. $^{*}p < .05$, $^{***}p < .001$ vs. 50%; † denotes significantly below 50%. Pooled $|\rho_{\text{len}}|$ averaged across the five RMs is 0.316 (baseline) and 0.037 for Huang et al. [2025]: audit-distribution success and within-prompt overshoot on the same operator.

- A selection-level ΔJ proxy (AlpacaEval length-controlled win rate vs. baseline, GSM8K BoN accuracy on selected responses).

The pair $\{\rho_{\text{len}}^{\text{pooled}}, \rho_{\text{len}}^{\text{within}}\}$ operationalizes Lemma 3.6 empirically, as an operator can zero one and not the other, and the selection outcome reveals which distribution drives the optimizing process. BoN top-1 over 64 samples is a coarser proxy for $\pi_\beta^\star$ than KL-regularized optimization, and the regime calls below should be read accordingly.

Results. The Huang et al. [2025] calibration achieves $|g_i| \approx 0$ at the pooled μ_{diag} it targets ($0.316 \to 0.037$ averaged across the five RMs). At the within-prompt μ_{diag} aligned with BoN selection, three of four SOTA RMs acquire negative $\rho_{\text{len}}^{\text{within}}$ of magnitude 0.065-0.228, exceeding their unmitigated baselines in absolute value (Table 4). The same pattern holds on GSM8K, where within-prompt $|\rho_{\text{len}}|$ averages 0.096 for Huang et al. [2025] versus 0.007 for mechanistic reward shaping, with BoN accuracy averaging 61.20% versus 62.76%. The signed flip rather than same-side overshoot is **evidence of the measurement-versus-optimization gap** of Lemma 3.6, because the operator reaches zero at μ_{diag}, while at the optimization-relevant distribution g_i has the wrong sign.

Regime classification. Under the ε-banded reading of Section A.8, two Huang et al. [2025] cells (Allen, Skywork-Q8B) satisfy $\Delta J < -\varepsilon_J$ on AlpacaEval LC win rate while the targeted axis $\rho_{\text{len}}^{\text{within}}$ has flipped sign. The GSM8K Allen cell additionally drops 3.6 accuracy points. These satisfy the targeted-axis and ΔJ conditions of R2 (overcorrection) in Definition 3.7. Whether they additionally satisfy the $\Delta_j = 0$ condition for off-target spurious axes is unmeasured here. As a result, the pattern is consistent with R2$_\varepsilon$ if the no-rotation condition holds, or with a mixed R1+R2 regime otherwise.

Mechanistic reward shaping achieves $\Delta J > 0$ at $p < .001$ on four of five AlpacaEval cells and moves $|\rho_{\text{len}}^{\text{within}}|$ toward zero on four of five RMs without sign flips. The Allen cell falls below 50% LC WR on both operators (46.2% for Mech. RS, 48.7% for Huang et al. [2025], both †). The GSM8K Allen cell drops 0.2 for mechanistic reward shaping and 3.6 for Huang et al. [2025]. The remaining four mechanistic reward shaping cells satisfy R0 (successful mitigation) Definition 3.7 modulo the unmeasured no-rotation condition.

Connection to the framework. The pair $(\rho_{\text{len}}^{\text{pooled}}, \rho_{\text{len}}^{\text{within}}) = (0.037, 0.116)$ for Huang et al. [2025] averaged across the five RMs is the empirical counterpart of Lemma 3.6: the canonical M_i reaches zero at the audit distribution it targets while g_i at the optimization-relevant distribution has the wrong sign. This dissociation is the load-bearing mechanism of the substitution argument. Wherever it holds, R1 is mechanically available to the optimizer, and the audit diagnostic targeting ϕ_i cannot detect it by construction. Standard reward-model benchmarks evaluate at a single fixed μ_{diag} and report only the pooled diagnostic, so under Section 3.4 they cannot distinguish this $\Delta J < 0$ instance

	Human	LLM	Agree	Disagree
#Samples N	14,375	28,832	4,149	10,226
Syc. length effect $\hat{\beta}_1$	$+24.3^{**}_{[9.6,\,39.0]}$	$+128.4^{***}_{[118.5,\,138.4]}$	$+154.3^{***}_{[123.6,\,185.0]}$	$-43.1^{***}_{[-60.5,\,-25.7]}$
Mean syc. length $\hat{\beta}_0$	1554.5	1487.1	1507.2	1568.7
Between model var. $\hat{\tau}^2$	72,408	70,521	68,577	73,110

Table 5: Mixed linear model of response length on the binary sycophancy indicator across the four labeling regimes. For response i under model j, we fit $\mathrm{len}_{ij} = \beta_0 + \beta_1 S_{ij} + u_j + \varepsilon_{ij}$, with $u_j \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, where len_{ij} is response length in characters, S_{ij} is the binary sycophancy indicator, β_0 is the fixed intercept (mean non-sycophantic length), β_1 is the fixed sycophantic length effect, u_j is the model-specific random intercept, and τ^2 is its between-model variance. Each $\hat{\beta}_1$ carries its 95% confidence interval as a subscript. Significance vs. no effect: $^{**}p < .01$, $^{***}p < .001$.

from R0 regardless of whether the underlying regime is $\mathrm{R1}_\varepsilon$ (rotation onto an unmeasured spurious axis), $\mathrm{R2}_\varepsilon$ (overcorrection on the targeted axis), or a mixture: all three are invisible at μ_{diag}, which is the central benchmark-inadequacy claim.

B.3 Evaluating sycophancy and length of AITA Responses

This section reports the experimental analysis backing the Section 4 claim that response length and sycophancy labels are statistically dependent across human, LLM-judge, and judge-agreement labeling regimes, using AITA prompts and responses from [Cheng et al., 2026a] across eight model families. We fit mixed linear models of response length (measured in characters) on a binary sycophancy indicator (with model or prompt as a random effect), and report KS and W_1. These statistics provide audit-side evidence for the construction-level precondition of R4 (Definition 3.8), with the formal banded treatment deferred to Section A.8. In terms of causality, the mixed-model coefficients and KS statistics below are observational. They are consistent with a mediation reading of length but do not, on preference data alone, discriminate among candidate causal structures.

Human-labeled regime (N = 14,375 responses across 8 models) A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a positive sycophantic effect of +24.3 length units (95% CI [9.6, 39.0]; p = 0.001), against an intercept of 1554.5 and between-model variance 72,408. The effect is small relative to between-model variability (median per-model non-syco/syco ratio 0.98), indicating that under human labeling sycophantic responses are only modestly longer on average, but statistically significant. Distributionally, the length distributions of sycophantic and non-sycophantic responses differ significantly for 6 of 8 models by Kolmogorov–Smirnov test (Llama-8B and Mistral-7B not significant at $\alpha = 0.05$), with per-model Wasserstein distances ranging from 24.5 to 138.6. Pooling z-scored within-model residuals gives KS = 0.025 (p = 0.022), confirming a small but significant aggregate distributional deviation.

LLM-judge–labeled regime (N = 28,832 responses across 8 models) A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a sycophantic effect of +128.4 length units (95% CI [118.5, 138.4]; p < 0.001), against an intercept of 1487.1 and between-model variance 70,521. The effect is substantially larger than under human labels (+24.3), indicating that LLM judges associate sycophancy with longer responses than the binary-verdict label does. Per-model non-syco/syco ratios shift correspondingly downward (median 0.94 vs. 0.98 under human labels), with the gap most pronounced for Llama-17B, Llama-70B, and Gemini (ratios 0.81–0.85). Length distributions differ significantly for 7 of 8 models by Kolmogorov–Smirnov test (only gpt-4o is not significant at $\alpha = 0.05$), with per-model Wasserstein distances ranging from 19.5 to 293.7, which is substantially larger than the 24.5–138.6 range under human labels. The pooled KS on z-scored within-model residuals is 0.130 (p < 0.001), roughly 5× larger than the human-label pooled KS of 0.025, confirming that the LLM-judge regime exhibits a much stronger distributional length–sycophancy coupling than the human regime at both the central-tendency and full-distribution levels.

	Human	LLM	Agree	Disagree	Within
Median ratio	0.98	0.94	0.93	1.03	—
KS sig. (of 8)	6	7	4	7	—
W_1 range	24.5–138.6	19.5–293.7	25.9–437.2	26.1–267.7	—
Pooled KS	0.025^{*}	0.130^{***}	0.130^{***}	0.046^{***}	0.073^{***}

Table 6: Distributional length–sycophancy coupling across labeling regimes. Median ratio is the per-model non-syco/syco length ratio. "KS sig." counts models with a significant per-model length distribution difference at $\alpha = 0.05$. Pooled KS is on z-scored within-group residuals. The per-model rows are not applicable to the Within regime (—), which is fit per-prompt. Significance of pooled KS: $^{*}p < .05$, $^{***}p < .001$.

Judge-agreement regime (N = 4,149 responses across 8 models, restricted to responses where human and LLM judge labels coincide) A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a sycophantic effect of +154.3 length units (95% CI [123.6, 185.0]; p < 0.001), against an intercept of 1507.2 and between-model variance 68,577. The effect is the largest of the four regimes, exceeding both human-only (+24.3) and LLM-only (+128.4), which is consistent with the agreement subset isolating responses for which sycophancy is most unambiguous along a length-correlated axis. Per-model non-syco/syco ratios drop correspondingly (median 0.93), with Llama-70B and Gemini showing the strongest shifts (ratios 0.77 and 0.77). Length distributions differ significantly for 4 of 8 models by Kolmogorov–Smirnov test at $\alpha = 0.05$ (Llama-8B, Llama-17B, Llama-70B, Gemini), with per-model Wasserstein distances ranging from 25.9 to 437.2 and reaching their highest values across all regimes. The four non-significant per-model KS tests (Claude, gpt-4o, Mistral-7B, Mistral-24B) likely reflect reduced statistical power, given the 8× range in per-model group sizes (94 to 784). The pooled KS on z-scored within-model residuals is 0.130 (p < 0.001), comparable to the LLM regime, indicating that the distributional length–sycophancy coupling under judge agreement is at least as strong as under LLM-only labeling at the aggregate level.

Judge-disagreement regime (N = 10,226 responses across 8 models, restricted to responses where human and LLM judge labels disagree) A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a *negative* sycophantic effect of -43.1 length units (95% CI $[-60.5, -25.7]$; p < 0.001), against an intercept of 1568.7 and between-model variance 73,110. The sign reversal relative to the human (+24.3), LLM (+128.4), and agreement (+154.3) regimes is the central qualitative finding: under disagreement, responses labeled sycophantic are *shorter* than non-sycophantic ones on average. Per-model non-syco/syco ratios shift accordingly (median 1.03, with 6 of 8 models above 1.0), reversing the pattern seen in the other three regimes. Length distributions differ significantly for 7 of 8 models by Kolmogorov–Smirnov test (only Mistral-7B is not significant at $\alpha = 0.05$), with per-model Wasserstein distances ranging from 26.1 to 267.7. The pooled KS on z-scored within-model residuals is 0.046 (p < 0.001), smaller than the LLM and agreement regimes (both 0.130) but larger than the human regime (0.025), indicating a distributional shift that is real but less pronounced than under either single-judge or unanimous-agreement labeling. The reversal is consistent with humans and LLM judges locating sycophancy in systematically different parts of the response-length distribution: when their labels diverge, the resulting "sycophantic" set is enriched for cases each judge alone would have ruled differently on, and the length signal flips accordingly.

Within-prompt across-models robustness check (N = 11,309 responses across 1,569 prompts with label variation) To rule out prompt-level content as a confound — the possibility that the length–sycophancy association in the four mode-level regressions reflects sycophantic *prompts* eliciting longer responses rather than sycophantic *responses* being longer per se — we refit the mixed linear model with **prompt as the random effect**, restricting to the 1,569 prompts where at least two of the eight models' responses received different sycophancy labels. With group size 5–8 (mean 7.2), this within-prompt design holds prompt content fixed and identifies the sycophantic effect off cross-model variation in labeling and length on the same prompt. The estimated sycophantic effect is +58.8 length units (95% CI [40.4, 77.1]; p < 0.001), against an intercept of 1549.5 and between-prompt variance 47,247. The positive sign and significance confirm that sycophantic responses are

longer than non-sycophantic responses *to the same prompt*, ruling out prompt-level content as the sole driver of the association. The pooled KS on z-scored within-prompt residuals is 0.073 ($p < 0.001$), indicating distributional length–sycophancy coupling that survives prompt-level conditioning. The within-prompt analysis uses the same human-label scheme as the four-mode human regression (model binary verdict on ‘is_a**hole == 1‘ prompts); the larger coefficient (+58.8 vs. +24.3) reflects the restriction to the 1,569 prompts with cross-model label variation, not a change in labeling regime. Sycophancy effects are concentrated on prompts where models actually disagree about the verdict.

C Existing Evidence for Confounding Factors

Fact C.1. ***Reward models are correlated with response length independent of content quality****, from the first language model applications of RLHF to today [Stiennon et al., 2020, Shen et al., 2023, Eisenstein et al., 2024, Singhal et al., 2024, Nvidia et al., 2024, Chen et al., 2024a, Wen et al., 2025, Wang et al., 2025, Zhao et al., 2025, Fein et al., 2026].*

Fact C.2. ***Reward models exhibit multiple additional distinct reward biases****, e.g., overconfidence [Leng et al., 2025], sycophantic user agreement [Sharma et al., 2024], answer position [Fein et al., 2026], prefix bias [Kumar et al., 2025], model-style sensitivity [Malik et al., 2026, Fein et al., 2026], and inherited value orientations (e.g., agency vs. communion) from pretraining base models [Christian et al., 2026].* ***All of these are present in state-of-the-art reward models*** *[Fein et al., 2026, Christian et al., 2026].*

Fact C.3. ***LLM policies trained with RLHF or DPO learn to generate longer outputs as a direct result of optimization pressure.*** *This effect can inflate benchmark scores [Meng et al., 2024, Singhal et al., 2024], produce outputs longer than training data [Park et al., 2024], and grow systematically longer than SFT baselines over training [Stiennon et al., 2020]. Overall, it represents a primary mode of reward hacking [Singhal et al., 2024, Park et al., 2024].*

Fact C.4. ***Naive length penalties can remove informative content****, as uniform length suppression degrades accuracy on tasks that genuinely require longer answers [Bu et al., 2025], and always-on penalties cause reward hacking via trajectory collapse in reasoning RL [Yuan et al., 2025].*

Fact C.5. ***Longer responses are systematically over-represented among "chosen" annotations in preference datasets, providing the distributional basis for reward models to learn length as a proxy for quality*** *[Singhal et al., 2024, Shen et al., 2023]. RLHF dramatically amplifies a slight length skew already present in preference data [Eisenstein et al., 2024], a pattern confirmed in crowdsourced human preference data, where controlling for length substantially reorders model rankings [Li et al., 2024]. Further, length preferences vary systematically across annotator populations and data sets [Movva et al., 2026].*

Fact C.6. ***Human preference judgments are systematically confounded by output length.*** *Across multiple preference-collection pipelines, annotators select the longer response in comparisons [Hu et al., 2025, Chen et al., 2024a, Singhal et al., 2024]. This pattern is visible even in carefully filtered datasets, as Stiennon et al. [2020] find that controlling for length reduces human preference gains and Saito et al. [2023] show a positive correlation between length and preference in the HH-RLHF corpus. In large-scale crowdsourced evaluation, length has been estimated as a dominant style factor in human voting [Li et al., 2024].*

Fact C.7. ***LLM judges exhibit verbosity bias [Zheng et al., 2023] and it is different to human annotators [Saito et al., 2023, Koo et al., 2024].*** *Verbosity is one of several systematic biases catalogued in LLM judges, in addition to other LLM judge biases [Koo et al., 2024, Chen et al., 2024a, Ye et al., 2025a, Wataoka et al., 2024]. Further evidence also supports that human and LLM judges can respond to length and content differently [Saito et al., 2023, Li et al., 2024, Chiang et al., 2024, Hu et al., 2025].*

Fact C.8. ***Human annotators systematically reward sycophantic responses.*** *Sharma et al. [2024] show that "matches user's beliefs" is among the most predictive features of human preference, with the preference model favoring sycophantic over truthful responses. Cheng et al. [2026b] confirm this finding at the data level, as preferred responses across preference datasets are significantly more validating and indirect. Perez et al. [2023] find that preference models incentivize sycophantic answers, while Wen et al. [2025] show RLHF-optimized models learn to defend incorrect answers with fabricated evidence and extended argumentation, increasing human approval without improving correctness. Ibrahim et al. [2026b] further show that supervised fine-tuning on warmer-style conversational data alone (without preference optimization) is sufficient to amplify affirmation of incorrect user beliefs by approximately 40% across five model families, with the effect surviving response-length controls.*

Fact C.9. ***Whether responses are labelled as sycophantic statistically depends on response length, independent of annotator type.*** *Prior work has reported response length and sycophancy jointly: Ibrahim et al. [2026b] include length as a covariate when regressing accuracy and sycophancy on warmth fine-tuning, and Dubois et al. [2026] treat length as a nuisance parameter when measuring sycophancy reduction. To our knowledge, however, the statistical dependence between response length and sycophancy labels has not been characterized across labeling regimes (human, LLM*

judge, agreement, disagreement), nor has the sign reversal under human-LLM judge disagreement been documented. We examine the dependence across multiple labeling strategies (human labels, LLM Judge labels, their agreement, and disagreement), systematically disentangling model-level verbosity from prompt-level content as alternative explanations using the AITA dataset from [Cheng et al., 2026a]. We find length and sycophancy of responses are statistically dependent across all four labeling regimes, see Section B.3 for experiment details. Sycophantic responses are longer under human, LLM, and human-LLM judge agreement labels, with the relationship reversing under human-LLM judge disagreement labels. This observation is consistent with humans and LLM judges locating sycophancy in systematically different parts of the response-length distribution. In addition, we find that the length-distribution of sycophantic and non-sycophantic responses significantly deviates for most models (Kolmogorov-Smirnov) and that LLM judge labels are substantially inflated by length bias relative to human labels.

Fact C.10. ***Sycophancy causes models to abandon correct responses in favor of alignment with user-stated positions.*** *When users suggest incorrect answers, model accuracy drops relative to unbiased baselines [Sharma et al., 2024] with high agreement rates with user-stated views for some question types [Perez et al., 2023]. Fanous et al. [2025] quantify this as regressive sycophancy, observing models switch from correct to incorrect under user rebuttals and Hong et al. [2025] further show that this is not a knowledge deficit. Beyond propositional agreement, framing sycophancy, in which models uncritically accept flawed user premises, proved particularly resistant to DPO-based mitigation [Cheng et al., 2026b]. Also, sycophantic affirmation narrows users focus to self-validation while omitting alternative perspectives [Cheng et al., 2026a].*

Fact C.11. ***Models can generate longer responses under epistemic uncertainty by hedging, qualifying, and elaborating more when confidence is low*** *[Zhang et al., 2025].*

Fact C.12. ***Reward models treat expressed confidence as a quality proxy, incentivizing models to replace hedged answers with confident-sounding responses regardless of actual certainty [Leng et al., 2025].*** *Reward models can penalize hedged correct answers over confidently-stated incorrect ones [Fein et al., 2026], and this penalty can be traced to the annotation level, where human raters in preference datasets are biased against expressions of uncertainty [Zhou et al., 2024, Ibrahim et al., 2026a].*

Fact C.13. ***Longer responses in uncertain regimes correlate with lower confidence and reduced quality.*** *Verbose outputs produced under verbosity compensation show measurably higher uncertainty and recall drops relative to concise responses [Zhang et al., 2025]. Reward models scoring confident-sounding elaborations highly regardless of correctness [Leng et al., 2025] implies the same inverse relationship, though without directly measuring it.*

Fact C.14. ***Reasoning models (not RLHF) produce longer outputs with uncertainty markers.*** *Extended chain-of-thought training produces epistemic uncertainty markers and non-linear reasoning traces through backtracking and alternative exploration rarely seen in non-reasoning models [Yoon et al., 2025]. During on-policy RL training, trajectory length grows systematically alongside exploratory behavior, and naively penalizing length degrades performance [Yuan et al., 2025], consistent with RL reinforcing the coupling.*

Fact C.15. ***Reward model scores increase with length when additional tokens carry genuine informational content*** *[Hu et al., 2025], but this relationship degrades at greater lengths, entering sublinear and then stochastic regimes beyond 100 and 200 tokens respectively [Zhao et al., 2025], consistent with* ***diminishing informational returns per token as responses grow longer.***

Fact C.16. ***Some reward-length correlation is irreducible and reflects genuine informativeness rather than spurious shortcut.*** *The information mass decomposition directly establishes that a portion of the length–reward correlation tracks real content quality [Hu et al., 2025], and more detailed answers can be genuinely more helpful depending on context [Bu et al., 2025]. Consistent with this observation, length-controlled analysis implies the original correlation is not entirely spurious [Dubois et al., 2024].*

Fact C.17. ***DPO-like objectives develop length bias out-of-distribution as a structural vulnerability of the algorithm without explicit mitigations****. DPO's implicit reward as a token-level log-probability ratios grows with sequence length, creating an algorithmic dependence on response length within the objective [Lu et al., 2024]. Empirically, this dependence manifests as the implicit reward acquiring significant length correlation once the policy moves away from the training distribution [Park et al., 2024, Meng et al., 2024], and can be mitigated through reward normalization and margin formulation [Li et al., 2025a]. Independently, under noisy or capability-limited supervision,*

DPO's KL regularization creates a structural dilemma where sufficient regularization preventing over-optimization also prevents larges updates needed for correcting errors inherited from the SFT phase [Ye et al., 2025c].

Fact C.18. ***KL regularization interacts with reward scale non-trivially.*** *KL-regularized policies are not invariant to positive linear reward scaling (see Section A.5), meaning two reward models agreeing on all preference orderings but differing in cardinal scale produce different policies [Skalse et al., 2023]. Empirically, there exists a measured optimal KL budget beyond which true reward degrades in RLHF [Gao et al., 2023] and longer sequences accumulate disproportionately more KL divergence, entangling length with the implicit DPO reward formulation [Lu et al., 2024]. The optimal regularization coefficient β depends strongly on feedback reliability [Ye et al., 2025c], meaning there is no single good KL setting across realistic annotation conditions.*

Fact C.19. ***Spurious length features can maintain ordinal ranking accuracy while distorting cardinal reward values, meaning benchmark metrics may fail to detect this failure mode.*** *A dissociation that follows from the partial-identifiability characterization of Skalse et al. [2023] combined with the KL scale-sensitivity of Section 2 (Equation* (1)*), and is directly measured as a low-complexity linear artifact in reward model representations [Fein et al., 2026]. RMs built on different base model families can achieve similar RewardBench scores while exhibiting systematically divergent value orientations [Christian et al., 2026] or producing systematically different PPO outcomes depending on policy-RM lineage match [Malik et al., 2026], concrete empirical instances of ordinal-invariant but cardinally-divergent reward functions. RewardBench [Lambert et al., 2025] scores improve after post-hoc length calibration, confirming retrospective distortion without prior detection [Huang et al., 2025], and apparent leaderboard gains shrink substantially under length-controlled evaluation in both automated [Dubois et al., 2024] and crowdsourced human voting settings [Li et al., 2024].*

Fact C.20. ***Evaluator and annotator weaknesses are learnable targets. Models trained with RLHF can acquire exploitation strategies like length padding, hedging caveats, and code obfuscation.*** *These strategies are specific to evaluator blind spots rather than tied to any single feature, demonstrating that reward gaming is a general learned behavior directed at the structure of the evaluation process itself [Wen et al., 2025].*

Fact C.21. ***Reward can be somewhat decomposed into separable dimensional components.*** *A two-head architecture directly separates length and content reward signals, confirms the components can be disentangled with supervised training signals to improve mitigation [Chen et al., 2024a]. Consistent with this, multi-attribute scoring frameworks track helpfulness, correctness, coherence, complexity, and verbosity as separately annotated (but correlated) dimensions [Wang et al., 2024, Nvidia et al., 2024].*

Fact C.22. ***Reward biases admit to first-order linear intervention.*** *Length bias is smooth, structured, and stable enough across model families to be estimated and corrected post-hoc via LOESS without retraining [Huang et al., 2025], but assumes that the true reward is independent of length. Fein et al. [2026] also use a linear probe to mitigate some part of length and other biases in reward models, while Papadatos and Freedman [2024] used linear probe-based reward corrections to mitigate sycophancy on simplified controlled setting.*

Fact C.23. ***Mitigating a spurious correlation in one feature can increase reward hacking in another feature (bias substitution).*** *In supervised vision, mitigating a single labeled shortcut amplifies reliance on the unlabeled one, with off-target axis degradation of 2–3x while aggregate worst-group accuracy improves [Li et al., 2023]. In RLHF, aggressive length-debiasing in recent SOTA reward models has flipped the length-bias sign and reduced correctness [Fein et al., 2026]. In LLM preference optimization more broadly, several DPO bias-mitigation variants reduce targeted bias at significant cost to general capability [Bu et al., 2025, Zhao et al., 2025], instantiating R2. The mechanism is compatible with the single-proxy correlated-feature bound showing that drift along an unconstrained correlated direction remains available to the policy whenever the proxy–truth correlation is below one [Laidlaw et al., 2025]. Liu et al. [2026] address this at the reward level via max-min optimization, but without the R0–R4 diagnostic machinery, substitution between interpretable feature axes remains undetectable. Conversely, jointly mitigating multiple shortcuts via kernel-based regularization [Ye et al., 2025b] achieves near-zero shortcut correlations at the audit distribution, but without measuring whether optimization pressure has shifted at μ_{π^*}, instantiating the distributional blindspot of Section 3.4.*

D Mapping published mitigations onto the regime taxonomy

We classify each mitigation method surveyed in Section 4 by what the paper's own reported evidence places in our regime taxonomy. Across the methods we survey, no paper provides the joint evidence that Theorem 3.10 requires to certify R0. Four methods provide direct evidence for a failure regime. Six make the strongest gestures toward R0 but leave at least one required input unmeasured. The remaining fifteen are undetermined and consistent with $\mathrm{R0}_{\mathrm{cont}}$ or stronger failure modes. For direct preference optimization methods, the operator-side versus loss-side distinction follows Section A.12 and Definition A.11.

Direct evidence for a failure regime.

- Bu et al. [2025] (ALBM). $\mathrm{R2}_{\varepsilon}$ on the targeted axis. The paper itself reports length-asymmetric subset accuracy regression under uniform suppression.
- Huang et al. [2025] (post-hoc LOESS calibration). $\mathrm{R2}_{\varepsilon}$ on the targeted axis with R1 plausible. Our Section B.2 adds within-prompt sign flips on three of four SOTA RMs and $\Delta J < 0$ on two AlpacaEval cells.
- Zhao et al. [2025] (FiMi-RM). R2 or R3 with $\mathrm{R0}_{\mathrm{cont}}$ undisambiguated. The paper reports an overall preference accuracy drop on the targeted backbone.
- Eisenstein et al. [2024] (RM ensembles). $\mathrm{R0}_{\mathrm{cont}}$ or R1. The paper documents shared failure modes where ensemble members agree on length doubling and copy hacking.

Strongest gestures toward R0, R0 versus $\mathrm{R0}_{\mathrm{cont}}$ undisambiguated. These methods report $\Delta J > 0$ at a held-out evaluation but do not report off-target Δ_j across Φ_{sp}, so $\mathrm{R0}_{\mathrm{cont}}$ cannot be ruled out.

- Fein et al. [2026] (mechanistic reward shaping). $\Delta J > 0$ on four of five RMs at within-prompt μ_{diag} (Section B.2). Off-target Δ_j across the rest of Φ_{sp} unmeasured.
- Cai et al. [2026] (Rc-BT). PPO $\Delta J > 0$ versus baseline. An RM-level format-bias ablation is reported but not propagated to policy outcomes.
- Srivastava et al. [2026] (CROME). Best multi-prescription coverage in the surveyed set, combining multi-axis evaluation, Best-of-N, and reWordBench transformation robustness. The LLM-oracle counterfactuals inherit R4 sensitivity that is not measured.
- Song et al. [2025] (SAE causal adjustment). $\Delta J > 0$ across every evaluated combination of policy model, dataset, and reward models. The Φ_{sp} panel is whatever the SAE learns, not a designer-specified set.
- Feng et al. [2025] (C2PO). Reports preserved or improved general capability alongside reductions across a multi-axis bias panel. Off-target Δ_j on unmodeled features and π_{ref}-sensitivity unmeasured.
- Umer et al. [2026] (GPRL). Multi-axis online RL with structural resistance to single-axis hacking and an online drift controller. $\Delta J > 0$ on four benchmarks at the primary base and across four base policies on AlpacaEval. Φ_{sp} implicit in the underlying preference model, off-panel substitution unaddressed.

Undetermined direct preference optimization methods (loss-side and data-side). The constructive single-axis guarantees of Section A.12 do not apply. These methods are classifiable only through the policy pair, and the surveyed evidence sits inside the impossibility class of Theorem 3.9.

- Park et al. [2024] (R-DPO). Length penalty validated only on preference data.
- Meng et al. [2024] (SimPO). Reference-free with length-normalized reward $\frac{\beta}{|y|}\log \pi_\theta$, a length-dependent reweighting interacting with the cardinal/ordinal blindspot of Corollary A.5 rather than its scalar-rescaling case.
- Lu et al. [2024] (SamPO). Algorithmic length dependence in DPO addressed analytically. Empirical validation does not isolate ΔJ from length reduction.
- Li et al. [2025a] (LMPO). The reported LC win-rate gain is on the same metric the loss is built around, a Goodhart pattern.

- Kim et al. [2025] (CDA via causal lens). Length-controlled accuracy gains substantial. General RewardBench accuracy gains small in absolute terms on both v1 and v2.

Undetermined RM-side methods with audit-only validation.

- Fu et al. [2025] (PAR). Single Gemma-2B lineage. Targets training stability rather than length bias directly.
- Shen et al. [2023] (Loose Lips). Validation distribution coincides with the RM training distribution. Only pooled length correlation reported.
- Chen et al. [2024b] (ODIN). Two-head disentanglement zeros pooled length correlation. Lemma 3.6 says this does not transfer to $\mu_{\pi^\star}$.
- Wang et al. [2024] (HelpSteer). Multi-attribute data construction with ordinal MT-Bench evaluation. No operator-level mitigation claim the framework can adjudicate.
- Wang et al. [2025] (causal rewards). Synthetic identifiability sound. Real-world evidence is correlation with GPT-4 and humans, not isolated ΔJ.
- Ng et al. [2025] (debiasing with guarantees). Identifiability theorem is conditional on the latent-variable DAG, which preference data does not pin down [Skalse et al., 2023]. Empirical validation small or synthetic.
- Li et al. [2025b] (DIR). Information-bottleneck framing handles non-linearity in principle. Empirical evidence conflates bias mitigation with downstream task improvement.
- Liu et al. [2026] (robust correlated proxies). Improves a worst-case proxy bound. The worst-case proxy is not the true reward, and the Linear Worst* probe shows the linear variant degrades on unseen features.
- Ye et al. [2025b] (PRISM). Near-zero correlation on three measured shortcuts. Off-panel substitution unmeasured.
- Chen et al. [2026] (OPRM with RgFT). Calibration framing via ECE is the closest engagement with the cardinal versus ordinal distinction in the surveyed set, but the protocol does not exploit it for substitution detection.

Takeaway. The four methods with direct failure-regime evidence and the six strongest-gesture methods together cover 10 of 25 surveyed works. The remaining fifteen are undetermined under R0, R0$_{\mathrm{cont}}$, R1, and R2 from their reported evidence alone. Closing this gap requires either policy-induced $(\Delta_j, \Delta J)$ measurement on the methods side or benchmark infrastructure that captures it on the evaluation side, as set out in Section A.13.