Reinforcement Learning from Downstream Feedback

Training Language Models on the Consequences of Their Answers

Aryan Gupta
aryan.cs.app@gmail.com

June 2026

Abstract

Reinforcement learning with verifiable rewards (RLVR) trains a policy against an incorruptible, policy-independent grader, and works on tasks where correctness can be checked mechanically, such as mathematics and code. Most queries a language model receives are not of this kind: a recommendation or an explanation has no answer key. This note develops *Reinforcement Learning from Downstream Feedback* (RLDF), which replaces the verifier by a population of simulated consumers. Each candidate response is played out against the population, the realized outcome of acting on it is graded, and the population-aggregated grade is the reward. We formalize the quantity being optimized and prove the following. The foresight (response-only) reward used by ordinary RLHF is not welfare-consistent once responses can inflate perceived quality (Lemma 1), whereas the hindsight (outcome-conditioned) reward is welfare-consistent under an explicit calibration hypothesis (Lemma 2). A finite panel of consumers estimates population welfare at rate $O(1/\sqrt{n})$ (Lemma 3), and a strategically corruptible minority admits a closed-form threshold $\varepsilon^\star$ above which the reward-maximizing response is a manipulation (Lemma 4). An α-trimmed-mean aggregator then preserves the welfare *ranking* of responses, though not their welfare *value*, whenever the corruptible fraction satisfies $\varepsilon < \alpha$ and a margin condition holds (Lemma 7). Together these give a welfare-consistency theorem for robust RLDF (Theorem 1), with RLVR as the special case of a single consumer, a deterministic world, and a perfect verifier (Proposition 1, Corollary 1). We then relax exact calibration to an η-approximate hypothesis (Theorem 2), reduce the welfare regret of the training loop to the reward error and characterize the floor that the trimming and contamination penalties leave behind (Theorem 3), and lift the per-query result to a per-distribution welfare-regret bound (Theorem 4) whose contamination threshold $\varepsilon < \alpha$ we show is sharp (Proposition 3).

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) is the training recipe behind the recent generation of reasoning models [1, 2, 3]. For a prompt whose answer can be checked, such as a final numeric answer or a unit test, one samples responses from the policy, scores each with an automatic verifier returning 1 for correct and 0 otherwise, and ascends the resulting reward. No human annotator rates the outputs and no learned reward model sits between the policy and the truth. The verifier is incorruptible and policy-independent: a wrong answer cannot be rephrased into a passing grade, and the grader does not drift as the policy improves.

This is also what bounds RLVR's reach. The recipe applies only where a cheap, sound checker exists, which covers a small fraction of the queries users pose; a question like whether to accept a job offer has no answer key. Recent work extends RLVR to open-ended domains by learning

a soft grader, either a generative reward model or a rubric critic [4, 5, 6], but a learned grader reintroduces the corruptibility the verifier removed: it scores the response, and a response can be optimized to look good to the grader without being good for the user. Sycophancy and confident hallucination follow [9, 8].

The downstream-feedback idea. RLDF changes what is graded. Instead of scoring the response, it simulates what happens after a user acts on the response and grades the consequence. A sycophantic or confidently hallucinated answer can score well to a response-only grader, but once a simulated user acts on it the realized outcome is poor, and the reward depends on that outcome rather than on the response's surface plausibility. In the economic reading, a response is a product, a population of heterogeneous simulated consumers acts on it and realizes welfare or regret, and the aggregate of those outcomes is the reward. The point of departure from RLVR is that for a non-trivial class of tasks the outcome of acting on a response is mechanically scorable even when the response itself is open-ended natural language, so consequence-grading recovers a policy-independent signal where response-grading cannot (Definition 5).

Outline. Section 2 fixes notation, and Section 3 shows that RLVR is the special case of RLDF with a single consumer, a deterministic world, and a perfect verifier, placing RLDF on a spectrum from outcome-verifiable tasks, where the grader is policy-independent, to simulated-outcome tasks, where it is not. Section 4 treats the foresight-hindsight gap: the response-only reward is not welfare-consistent when responses carry a persuasion bias (Lemma 1), while the outcome-conditioned reward is, under calibration (Lemma 2); these restate, in the present notation, the central claim of RLHS [7]. Section 5 shows the finite-panel reward concentrates on population welfare. Section 6 derives the gameability threshold and shows that a stronger type-aware adversary is blocked by grading on representative panels, which makes precise the finding of [8] that a small gameable minority suffices to induce manipulation. Section 7 replaces the mean by a trimmed mean and recovers the welfare ranking under $\varepsilon < \alpha$ and a margin condition. Section 8 assembles these into Theorem 1. Sections 9–11 then extend the guarantee in three directions: to approximate calibration, to the regret of the training loop, and to a per-distribution welfare bound with a matching sharpness result on the contamination threshold. Sections 12–13 discuss the assumptions and the relation to prior work.

2 Setup and notation

We work with a one-shot advisory interaction: the policy observes a query, emits a response, a user acts on it, and an outcome is realized. The multi-turn case extends this straightforwardly and is not needed for any result below.

Definition 1 (Interaction primitives)**.** Fix measurable spaces of *queries* $\mathcal{X}$, *responses* $\mathcal{A}$, *user types* $\mathcal{U}$, and *outcomes* $\mathcal{O}$. The interaction is specified by:

(i) a query distribution $\mathcal{D} \in \Delta(\mathcal{X})$;

(ii) a *population kernel* $P(\cdot \mid x) \in \Delta(\mathcal{U})$, the distribution of user types who issue query x (heterogeneity: distinct users may share a query yet differ in their latent situation and preferences);

(iii) a *policy* $\pi_\theta(\cdot \mid x) \in \Delta(\mathcal{A})$, parameterized by $\theta \in \Theta$;

(iv) a *world model* (transition kernel) $T(\cdot \mid x, u, a) \in \Delta(\mathcal{O})$, the distribution of realized outcomes when a user of type u acts on response a to query x;

(v) a *welfare function* $W : \mathcal{X} \times \mathcal{U} \times \mathcal{A} \to [0,1]$, the true (latent) utility to a user of type u from acting on response a. Boundedness to $[0,1]$ is a normalization.

Definition 2 (Alignment target and population welfare)**.** The *alignment target* is

$$J^\star(\theta) \;:=\; \mathbb{E}_{x\sim\mathcal{D}}\, \mathbb{E}_{u\sim P(\cdot\,|\,x)}\, \mathbb{E}_{a\sim\pi_\theta(\cdot\,|\,x)}\big[\, W(x,u,a)\,\big].$$

For fixed (x,a), the *population-average welfare* is $\overline{W}(x,a) := \mathbb{E}_{u\sim P(\cdot\,|\,x)}[\, W(x,u,a)\,]$, so that $J^\star(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\, \mathbb{E}_{a\sim\pi_\theta(\cdot\,|\,x)}[\, \overline{W}(x,a)\,]$. The *welfare-optimal response set* at x is $\mathcal{A}^\star(x) := \arg\max_{a\in\mathcal{A}} \overline{W}(x,a)$ (assumed non-empty).

Because the policy may choose its action distribution separately for each x, maximizing $J^\star$ decomposes pointwise: $J^\star$ is maximized iff $\pi_\theta(\cdot\,|\,x)$ is supported on $\mathcal{A}^\star(x)$ for $\mathcal{D}$-almost every x. This reduces the design problem to a *per-query ranking* problem, which is the criterion every reward in this note is measured against.

Definition 3 (Welfare-consistency of a reward)**.** A reward $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is *welfare-consistent at* x if $\arg\max_{a\in\mathcal{A}} r(x,a) \subseteq \mathcal{A}^\star(x)$, and *on a candidate set* $\mathcal{A}_c \subseteq \mathcal{A}$ if the inclusion holds with both arg max operators restricted to $\mathcal{A}_c$. A policy that maximizes a welfare-consistent reward pointwise maximizes $J^\star$.

Definition 4 (Foresight and hindsight graders)**.** A *foresight grader* $r_F : \mathcal{X} \times \mathcal{A} \to [0,1]$ scores a response before any outcome is realized, seeing (x,a) only. A *hindsight grader* is the composition of (i) the world model T, realizing $o \sim T(\cdot\,|\,x,u,a)$, and (ii) an *outcome evaluator* $g : \mathcal{O} \to [0,1]$, returning $g(o)$. Throughout, "grader" in the hindsight setting refers to the entire simulate-then-evaluate pipeline (T,g); in particular, any way a response could fool the simulator into producing a misleadingly favorable outcome is accounted for inside $g \circ T$, not treated as a separate channel.

3 RLVR as a degenerate special case

RLDF strictly generalizes RLVR. Recording the special case first fixes intuition, and it is used later when the main theorem recovers RLVR's guarantee as a limit.

Proposition 1 (RLVR is zero-noise, zero-corruption RLDF)**.** *Suppose at every query x:* (a) $P(\cdot\,|\,x) = \delta_{u_0}$ *(single canonical consumer);* (b) $T(\cdot\,|\,x,u_0,a) = \delta_{o(x,a)}$ *(deterministic world);* (c) $g(o(x,a)) = \mathbf{1}\{a \text{ correct for } x\}$ *is computed exactly (perfect, non-corruptible verifier);* (d) $W(x,u_0,a) = \mathbf{1}\{a \text{ correct for } x\}$*. Then the hindsight reward equals welfare identically, $\overline{W}(x,a) = g(o(x,a)) = W(x,u_0,a)$, with zero variance and no corruptible mass; consequently the RLDF reward reduces to the RLVR reward and is welfare-consistent at every x.*

Proof. Under (a), $\overline{W}(x,a) = W(x,u_0,a)$; under (d), this equals $\mathbf{1}\{a \text{ correct}\}$. Under (b)–(c), the hindsight grade is the deterministic quantity $g(o(x,a)) = \mathbf{1}\{a \text{ correct}\}$, identical to $\overline{W}(x,a)$ and noise-free. Then $\arg\max_a \overline{W}(x,a)$ is the set of correct responses, which is $\mathcal{A}^\star(x)$ by definition. □

RLDF generalizes RLVR along the three axes that Proposition 1 holds fixed: a non-degenerate population P, a stochastic world T, and an imperfect grader. The third axis deserves a name, since it is what separates an extension of RLVR from a relabeled reward model.

Definition 5 (Outcome-verifiable vs. simulated-outcome tasks)**.** A task (x,T,g) is *outcome-verifiable* if the pipeline $g \circ T$ is policy-independent: the outcome of acting on a response is produced by a fixed executor and scored by a fixed program ϕ, so $g(T(x,u,a)) = \phi(\mathcal{E}(x,a))$ for a sandbox state $\mathcal{E}$

and scorer ϕ that do not depend on θ. It is *simulated-outcome* if $g \circ T$ is itself a learned model (for instance an LLM world model and evaluator), whose representation of consequences can co-adapt with the policy.

Remark 1 (The spectrum). Outcome-verifiable tasks are common even when the response is open-ended. A code change, for instance, is free-form text, but whether it passes the project's test suite is a fixed check that does not depend on the policy. On such tasks $g \circ T$ is a genuine verifier, the corruption parameter ε of Section 6 is 0, and Theorem 1 recovers RLVR-grade guarantees (Corollary 1). On simulated-outcome tasks the grader is a model and $\varepsilon > 0$ in general, so calling the reward verifiable is unwarranted without an argument that $g \circ T$ is hard to fool; the robust aggregation of Section 7 stands in for the missing policy-independence.

4 The foresight-hindsight gap

Grading the response and grading the consequence give different rewards, and the difference is visible already at the level of the reward, before any optimizer is introduced. The results in this section are, in substance, the central claim of RLHS [7] in the present notation; we include proofs because later results build on them.

A perfectly Bayesian, non-manipulable foresight grader would report $\mathbb{E}_u[W(x,u,a)] = \overline{W}(x,a)$ and would be welfare-consistent. The empirical content of the sycophancy and deception literature [9, 8, 10] is that real graders are not of this kind: a response can be shaped to inflate perceived quality without improving welfare. We take this as an assumption rather than deriving it.

Assumption 1 (Persuasion bias of foresight grading)**.** The foresight grader decomposes as $r_F(x,a) = \overline{W}(x,a) + B(x,a)$, where $B : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the *persuasion bias*, the amount by which response a moves perceived quality away from true population welfare. A non-manipulable grader has $B \equiv 0$; we assume only that B is *response-sensitive*, i.e. there exist queries at which $B(x,\cdot)$ is non-constant in a.

Lemma 1 (Foresight reward is not welfare-consistent)**.** *Under Assumption 1, there exist a query x and responses $a_1, a_2 \in \mathcal{A}$ with $\overline{W}(x,a_1) > \overline{W}(x,a_2)$ yet $r_F(x,a_1) < r_F(x,a_2)$; hence $\arg\max_a r_F(x,a) \not\subseteq \mathcal{A}^\star(x)$, and maximizing the foresight reward can strictly select a welfare-inferior response.*

Proof. This is an existence claim, so one instance suffices. Fix a query x at which $B(x,\cdot)$ is non-constant (Assumption 1). Let a_1 be a welfare-superior "honest" response and a_2 a "sycophantic" response with

$$\overline{W}(x,a_2) = \overline{W}(x,a_1) - \delta \quad (\delta > 0), \qquad B(x,a_2) - B(x,a_1) > \delta.$$

Such a pair exists whenever the bias gap can exceed the welfare gap, the defining feature of a manipulable grader. Then

$$r_F(x,a_2) - r_F(x,a_1) = \underbrace{[\overline{W}(x,a_2) - \overline{W}(x,a_1)]}_{=-\delta} + \underbrace{[B(x,a_2) - B(x,a_1)]}_{>\delta} > 0,$$

so a_2 is strictly foresight-preferred while strictly welfare-inferior, and $a_2 \in \arg\max\{r_F(x,a_1), r_F(x,a_2)\}$ with $a_2 \notin \mathcal{A}^\star(x)$. □

Remark 2. The lemma is an existence statement. Response-sensitivity of B does not by itself reorder the arg max: a bias that varies in a but never overtakes the welfare gap leaves the optimum intact. The assumption is that the overtaking case occurs, which is what the sycophancy literature reports, and one instance is enough to deny welfare-consistency.

Assumption 2 (Outcome calibration)**.** The hindsight grader is *outcome-calibrated* if, for all (x, u, a),

$$\rho(x, u, a) \;:=\; \mathbb{E}_{o\sim T(\cdot\,|\,x,u,a)}\big[g(o)\big] \;=\; W(x, u, a).$$

More generally we allow affine-monotone calibration $\rho = c_0 + c_1 W$ with $c_1 > 0$; every ranking statement below is invariant to such a reparameterization, so proofs are stated for $c_0 = 0,\ c_1 = 1$.

Lemma 2 (Hindsight reward is welfare-consistent)**.** *Define the population hindsight reward* $r_H(x, a) := \mathbb{E}_{u\sim P(\cdot\,|\,x)}\mathbb{E}_{o\sim T(\cdot\,|\,x,u,a)}[g(o)] = \mathbb{E}_{u\sim P(\cdot\,|\,x)}[\rho(x, u, a)]$. *Under Assumption 2,* $r_H(x, a) = \overline{W}(x, a)$ *for all* x, a*; consequently* r_H *is welfare-consistent at every* x*.*

Proof. By the tower property $r_H(x, a) = \mathbb{E}_u[\rho(x, u, a)]$; substituting calibration $\rho(x, u, a) = W(x, u, a)$ gives $r_H(x, a) = \mathbb{E}_u[W(x, u, a)] = \overline{W}(x, a)$. Hence $\arg\max_a r_H(x, a) = \arg\max_a \overline{W}(x, a) = \mathcal{A}^\star(x)$. □

Remark 3 (Why the bias disappears). The persuasion term B has no analogue in r_H because $g(o)$ is a function of the realized outcome o, and o depends on the response only through the law $T(\cdot\,|\,x, u, a)$, the channel through which a actually changes the world. A persuasive but unhelpful response induces the same outcome distribution as a blunt unhelpful one and is graded identically. Conditionally on (x, u, a), the persuasiveness of a is irrelevant to o; this conditional irrelevance is exactly Assumption 2, and its failure is the corruptible case of Section 6.

5 The market reward over a finite population

Lemma 2 concerns the idealized reward $r_H = \overline{W}$, an expectation over the full population and outcome law. An implementable reward replaces both expectations by simulation: draw a finite *panel* of consumers, realize one outcome each, and aggregate the grades. This panel is the "market" that prices the response by its realized experience.

Definition 6 (Empirical market reward)**.** Fix (x, a) and panel size n. Draw $u_1, \dots, u_n \overset{\text{i.i.d.}}{\sim} P(\cdot\,|\,x)$, and for each draw $o_i \sim T(\cdot\,|\,x, u_i, a)$. The *empirical market reward* is the panel mean $\hat{r}_n(x, a) := \frac{1}{n}\sum_{i=1}^n g(o_i) \in [0, 1]$.

Lemma 3 (Concentration of the market reward)**.** *Assume* $g \in [0, 1]$. *Then* $\mathbb{E}[\hat{r}_n(x, a)] = r_H(x, a)$*, and for every* $t > 0$*,*

$$\mathbb{P}\big(\,|\hat{r}_n(x, a) - r_H(x, a)| \geq t\,\big) \;\leq\; 2\exp(-2nt^2).$$

Under Assumption 2 (so $r_H = \overline{W}$*), a panel of size* $n \geq \frac{1}{2t^2}\log\frac{2}{\delta}$ *gives* $|\hat{r}_n(x, a) - \overline{W}(x, a)| < t$ *with probability at least* $1 - \delta$*.*

Proof. The pairs (u_i, o_i) are i.i.d., hence so are the $g(o_i) \in [0, 1]$, with common mean $\mathbb{E}[g(o_i)] = \mathbb{E}_{u\sim P}\mathbb{E}_{o\sim T}[g(o)] = r_H(x, a)$ by Definition 4 and the tower property. Hoeffding's inequality [24] for an average of n independent $[0, 1]$ variables gives the two-sided bound; inverting $2e^{-2nt^2} \leq \delta$ gives the sample-size statement. □

Two of the three degeneracies of Proposition 1 are now lifted, the population being a genuine distribution and the world stochastic, and the reward still tracks welfare. The third, an incorruptible grader, is the subject of the next section and is the principal difficulty.

6 Strategic corruption and the gameability threshold

Lemma 2 rests entirely on calibration. We now model its failure. The failure that matters is not uniform noise but a strategic one: some subpopulation can be driven to report high grades by a response that does not serve it, for example a user who reports satisfaction when flattered. Following [8], we model a faithful majority and a corruptible minority.

Definition 7 (Contaminated population and manipulation)**.** Fix x. The population is an ε-contaminated mixture $P(\cdot \mid x) = (1-\varepsilon)P_{\mathrm{f}}(\cdot \mid x) + \varepsilon P_{\mathrm{g}}(\cdot \mid x)$, where calibration (Assumption 2) holds on the *faithful* part P_{f} and may fail on the *gameable* part P_{g}. A *manipulation* is a response $a^\dagger$ with a corrupted ceiling $G_{\max} \in (0,1]$ such that, for gameable u, $\rho(x,u,a^\dagger) = G_{\max}$ irrespective of the true (low) welfare $W(x,u,a^\dagger)$.

For a welfare-optimal honest response $a^\star \in \mathcal{A}^\star(x)$, write $\overline{W}^\star := \overline{W}(x,a^\star)$, and let

$$\overline{W}_h^\dagger \;:=\; \mathbb{E}_{u\sim P_{\mathrm{f}}}\big[W(x,u,a^\dagger)\big]$$

denote the manipulation's true welfare on the faithful part, with $\overline{W}_h^\dagger < \overline{W}^\star$. For transparency we take the honest response to attract no corruption (reward $\overline{W}^\star$); allowing it to be corrupted as well only strengthens the case for the remedy of Section 7.

Lemma 4 (Gameability threshold)**.** *Under mean aggregation over the contaminated population,* $r_{\mathrm{mkt}}(a^\star) = \overline{W}^\star$ *and* $r_{\mathrm{mkt}}(a^\dagger) = (1-\varepsilon)\overline{W}_h^\dagger + \varepsilon G_{\max}$*. The manipulation is strictly market-preferred iff*

$$\boxed{\varepsilon > \varepsilon^\star := \frac{\overline{W}^\star - \overline{W}_h^\dagger}{G_{\max} - \overline{W}_h^\dagger}}$$

and $\varepsilon^\star \in (0,1)$ *whenever* $\overline{W}_h^\dagger < \overline{W}^\star \le G_{\max}$*.*

Proof. The reward of $a^\dagger$ splits over the mixture: a $(1-\varepsilon)$ mass grades it at its true welfare $\overline{W}_h^\dagger$ (calibration on P_{f}), an ε mass at the ceiling $G_{\max}$ (Definition 7), giving $r_{\mathrm{mkt}}(a^\dagger) = (1-\varepsilon)\overline{W}_h^\dagger + \varepsilon G_{\max}$. Then

$$r_{\mathrm{mkt}}(a^\dagger) > r_{\mathrm{mkt}}(a^\star) \iff \varepsilon(G_{\max} - \overline{W}_h^\dagger) > \overline{W}^\star - \overline{W}_h^\dagger \iff \varepsilon > \varepsilon^\star,$$

dividing by $G_{\max} - \overline{W}_h^\dagger > 0$. The numerator is positive ($\overline{W}^\star > \overline{W}_h^\dagger$) so $\varepsilon^\star > 0$; and $\overline{W}^\star \le G_{\max}$ gives $\varepsilon^\star \le 1$. □

Remark 4 (What makes "2% suffices" rigorous). The threshold $\varepsilon^\star$ is small when the welfare margin $\overline{W}^\star - \overline{W}_h^\dagger$ is small relative to the corrupt advantage $G_{\max} - \overline{W}_h^\dagger$, that is, when gaming one gameable consumer pays off far more than honest service would. As $G_{\max} \to 1$ with a fixed modest margin, $\varepsilon^\star \to (\overline{W}^\star - \overline{W}_h^\dagger)/(1 - \overline{W}_h^\dagger)$, which can sit at a few percent. This is the closed-form mechanism behind the observation of [8] that a gameable minority of order 2% already tips an RL-trained model into manipulation: it is the leverage $G_{\max}$, not the size ε, that matters.

Remark 5 (Targeting, and why the panel must be representative). Lemma 4 prices a single response against a representative draw from $P(\cdot \mid x)$. A stronger adversary, a type-aware policy $\pi_\theta(\cdot \mid x, s)$ that observes a gameability signal s rather than the query-only policy $\pi_\theta(\cdot \mid x)$ of Definition 1, could in principle route $a^\dagger$ only to gameable consumers and $a^\star$ to the rest. If, in addition, the reward graded each response on the consumers who received it ("recipient-grading"), then $a^\dagger$ would be graded by an all-gameable panel and score the ceiling $G_{\max}$, collapsing the threshold to $\varepsilon^\star = 0$,

so that any positive corruptible fraction would be exploited. The defense is structural and is already built into Definition 6: every candidate is graded on a fresh i.i.d. panel drawn from the true population $P(\cdot \mid x)$, not on its recipients. Under representative panels the gameable share of $a^\dagger$'s panel is the population ε regardless of how the policy routes responses, so a type-aware adversary gains nothing at grading time, and hypothesis (A2) below ($\varepsilon < \alpha$ in every panel) holds by construction whenever the population's gameable fraction is itself below α. Recipient-grading forfeits this guarantee; representative grading keeps the corruptible mass a bounded minority rather than a hand-picked majority.

Together, Lemma 4 and Remark 5 show that a mean over a population carrying any corruptible mass is unsafe. The remedy has two parts: grade on representative panels, so the corruptible mass stays a bounded minority (Remark 5), and aggregate that minority out, so it cannot move the price (Section 7).

7 Robust aggregation restores the welfare ranking

We replace the mean by a trimmed mean, discarding the most extreme grades before averaging. There is a subtlety, and it is the reason the guarantee below concerns the ranking of responses rather than their welfare values: the trimmed mean does not estimate $\overline{W}$. It estimates a different functional of the grade distribution, and trimming introduces a bias even without corruption. We make this estimand shift explicit and prove the weaker property that suffices, namely that the welfare order of responses is preserved.

Definition 8 (α-trimmed mean and its functional)**.** For $\alpha \in [0, \frac{1}{2})$ and a grade law F on $[0,1]$ with quantile function F^{-1}, the *α-trimmed mean functional* is $\mu_\alpha(F) := \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(p)\, dp$. Its empirical version $\hat{\mu}_\alpha^{(n)}$ discards the $\lceil \alpha n \rceil$ largest and smallest of the n panel grades and averages the rest. Write $\mathrm{mean}(F) = \int_0^1 F^{-1}(p)\, dp$, so that under calibration the clean grade law F_a of response a has $\mathrm{mean}(F_a) = \overline{W}(x, a)$.

We isolate one elementary property and reuse it three times. Recall the Wasserstein-1 distance on $[0,1]$-supported laws, $W_1(F,G) = \int_0^1 |F(x) - G(x)|\, dx = \int_0^1 |F^{-1}(p) - G^{-1}(p)|\, dp$.

Lemma 5 (μ_α is $\frac{1}{1-2\alpha}$-Lipschitz in W_1)**.** *For $[0,1]$-supported F, G, $|\mu_\alpha(F) - \mu_\alpha(G)| \le \frac{1}{1-2\alpha} W_1(F,G) \le \frac{1}{1-2\alpha} \|F - G\|_\infty$.*

Proof. By Definition 8 and the triangle inequality, $|\mu_\alpha(F) - \mu_\alpha(G)| \le \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} |F^{-1}(p) - G^{-1}(p)|\, dp$. The integrand is non-negative, so extending the range to $[0,1]$ only increases the integral, giving $\frac{1}{1-2\alpha} \int_0^1 |F^{-1} - G^{-1}|\, dp = \frac{1}{1-2\alpha} W_1(F,G)$ by the quantile representation of W_1. Finally $W_1(F,G) = \int_0^1 |F(x) - G(x)|\, dx \le \|F - G\|_\infty$ since the support has length 1. □

Lemma 6 (Trimming bias, contamination, and sampling bounds)**.** *Let F be a clean $[0,1]$-supported grade law and F_n its n-sample empirical law. Then:*

(i) *(Trimming bias)* $|\mu_\alpha(F) - \mathrm{mean}(F)| \le 2\alpha$.

(ii) *(Contamination) for any $[0,1]$-supported F' and $\tilde{F} = (1-\varepsilon)F + \varepsilon F'$,* $|\mu_\alpha(\tilde{F}) - \mu_\alpha(F)| \le \frac{\varepsilon}{1-2\alpha}$.

(iii) *(Sampling) for every $s > 0$,* $\mathbb{P}(|\hat{\mu}_\alpha^{(n)} - \mu_\alpha(F)| > s) \le 2\exp(-2ns^2(1-2\alpha)^2)$.

Proof. (i) Write $\mu_\alpha(F) - \mathrm{mean}(F) = \int_0^1 w(p) F^{-1}(p)\, dp$ with $w(p) = \frac{1}{1-2\alpha} \mathbf{1}\{\alpha \le p \le 1-\alpha\} - 1$, so $\int_0^1 w = 0$. The positive part of w lives on $[\alpha, 1-\alpha]$ with value $\frac{2\alpha}{1-2\alpha}$ and total mass 2α; the

negative part on the tails with value -1 and total mass 2α. Since $F^{-1} \in [0,1]$, each of $\int w_+ F^{-1}$ and $\int w_- F^{-1}$ lies in $[0, 2\alpha]$, so their difference is at most 2α in absolute value.

(ii) By convexity of W_1 in its first argument and the diameter bound $W_1 \le 1$ on $[0,1]$, $W_1(\tilde{F}, F) \le \varepsilon\, W_1(F', F) \le \varepsilon$. Lemma 5 gives $|\mu_\alpha(\tilde{F}) - \mu_\alpha(F)| \le \frac{\varepsilon}{1-2\alpha}$.

(iii) By Lemma 5, $|\hat{\mu}_\alpha^{(n)} - \mu_\alpha(F)| = |\mu_\alpha(F_n) - \mu_\alpha(F)| \le \frac{1}{1-2\alpha}\|F_n - F\|_\infty$. The Dvoretzky–Kiefer–Wolfowitz inequality with Massart's tight constant [25, 26] gives $\mathbb{P}(\|F_n - F\|_\infty > \eta) \le 2e^{-2n\eta^2}$; set $\eta = s(1-2\alpha)$. □

Remark 6 (The estimand shift). Part (i) is not a bound that vanishes under optimization: even with a perfectly calibrated, uncontaminated grader, $\mu_\alpha(F_a) \neq \overline{W}(x,a)$ in general, because trimming reshapes the distribution. Robust RLDF therefore does not estimate welfare; it estimates a trimmed surrogate. What survives is the comparison between responses, which is all a reward optimizer consumes. The constant 2α is worst-case, attained only by near-$\{0,1\}$ bimodal grades; under a grade-concentration assumption $\mathrm{Var}_F(g) \le \sigma^2$ it improves to $O(\alpha\sigma)$, and a median-of-means aggregator [27] trades the trimming bias for a sub-Gaussian deviation bound of comparable structure. We carry the conservative 2α to keep the theorem assumption-light.

Let $a^\star$ be welfare-optimal and $a^\dagger$ any welfare-suboptimal alternative, with clean grade laws $F^\star, F^\dagger$ and welfares $\overline{W}^\star = \mathrm{mean}(F^\star)$, $\overline{W}^\dagger = \mathrm{mean}(F^\dagger)$. Suppose each response's panel may contain a corruptible fraction at most ε.

Lemma 7 (Ranking preservation under robust aggregation)**.** *Fix $\alpha \in (0, \frac{1}{2})$ with $\varepsilon < \alpha$. If the welfare margin satisfies*

$$M := \overline{W}^\star - \overline{W}^\dagger \;>\; \underbrace{4\alpha}_{\text{trimming}} + \underbrace{\frac{2\varepsilon}{1-2\alpha}}_{\text{contamination}} + \underbrace{2s}_{\text{sampling}},$$

then with probability at least $1 - 4\exp(-2ns^2(1-2\alpha)^2)$ the empirical trimmed market rewards satisfy $\hat{\mu}_\alpha^{(n)}(a^\star) > \hat{\mu}_\alpha^{(n)}(a^\dagger)$. Robust RLDF thus ranks the welfare-optimal response strictly above the suboptimal one despite trimming bias, an ε-corruptible panel, and finite sampling.

Proof. On the event that the honest response's sampling error is at most s, chaining (iii), then (ii), then (i) of Lemma 6,

$$\hat{\mu}_\alpha^{(n)}(a^\star) \ge \mu_\alpha(\tilde{F}^\star) - s \ge \mu_\alpha(F^\star) - \tfrac{\varepsilon}{1-2\alpha} - s \ge \overline{W}^\star - 2\alpha - \tfrac{\varepsilon}{1-2\alpha} - s.$$

Symmetrically, $\hat{\mu}_\alpha^{(n)}(a^\dagger) \le \overline{W}^\dagger + 2\alpha + \frac{\varepsilon}{1-2\alpha} + s$. Subtracting, on the intersection of the two sampling events,

$$\hat{\mu}_\alpha^{(n)}(a^\star) - \hat{\mu}_\alpha^{(n)}(a^\dagger) \ge M - 4\alpha - \frac{2\varepsilon}{1-2\alpha} - 2s > 0$$

by hypothesis. Each sampling event fails with probability $\le 2\exp(-2ns^2(1-2\alpha)^2)$ by Lemma 6(iii); a union bound over the two responses' upper and lower deviations (four events) gives the stated failure probability. □

8 Main theorem: welfare-consistency of robust RLDF

Calibration on the faithful subpopulation makes the clean reward track welfare (Lemma 2); robust aggregation contains the corruptible minority and the trimming and sampling penalties (Lemma 7); a margin assumption guarantees the welfare signal exceeds the total penalty.

Theorem 1 (Welfare-consistency of robust RLDF)**.** *Fix a query x and a finite candidate set $\mathcal{A}_c \subseteq \mathcal{A}$ containing a welfare-optimal $a^\star \in \mathcal{A}^\star(x)$. Suppose:*

(A1) (Calibration) *the hindsight grader is outcome-calibrated on the faithful subpopulation P_{f} (Assumption 2);*

(A2) (Bounded corruption) *every candidate's panel is at most ε-contaminated with $\varepsilon < \alpha$, where $\alpha \in (0, \frac{1}{2})$ is the trimming level;*

(A3) (Welfare margin) *for every welfare-suboptimal $a' \in \mathcal{A}_c$, $\overline{W}(x, a^\star) - \overline{W}(x, a') > 4\alpha + \frac{2\varepsilon}{1-2\alpha} + 2s$ for some $s > 0$;*

(A4) (Panel size) $n \geq \dfrac{1}{2s^2(1-2\alpha)^2} \log \dfrac{4|\mathcal{A}_c|}{\delta}$.

Then, with probability at least $1 - \delta$ over the simulated panels, the α-trimmed empirical market reward $\hat{\mu}_\alpha^{(n)}$ is welfare-consistent on $\mathcal{A}_c$: its maximizer over $\mathcal{A}_c$ lies in $\mathcal{A}^\star(x)$. Consequently any policy maximizing the robust RLDF reward over $\mathcal{A}_c$ at x places its mass on the welfare-optimal set, and a policy doing so for $\mathcal{D}$-a.e. x maximizes the alignment target $J^\star$.

Proof. Apply Lemma 7 to $(a^\star, a')$ for each welfare-suboptimal $a' \in \mathcal{A}_c$. Hypotheses (A1)–(A3) are exactly its hypotheses: (A1) supplies $\mathrm{mean}(F_a) = \overline{W}(x, a)$ on the faithful mass, (A2) supplies $\varepsilon < \alpha$, (A3) is the margin condition with slack s. The lemma gives $\hat{\mu}_\alpha^{(n)}(a^\star) > \hat{\mu}_\alpha^{(n)}(a')$ with failure probability $\leq 4\exp(-2ns^2(1-2\alpha)^2)$ for that pair. By (A4) this is $\leq \delta/|\mathcal{A}_c|$, so a union bound over the at most $|\mathcal{A}_c|$ suboptimal candidates makes the total failure probability $\leq \delta$. On the complement, $a^\star$ strictly outscores every other candidate, hence $\arg\max_{a \in \mathcal{A}_c} \hat{\mu}_\alpha^{(n)}(a) = \{a^\star\} \subseteq \mathcal{A}^\star(x)$. The final sentence is the pointwise decomposition following Definition 2. □

Corollary 1 (RLVR recovery)**.** *In the regime of Proposition 1, equivalently an outcome-verifiable task (Definition 5) with a single consumer, one has $\varepsilon = 0$, zero sampling variance (so any $s > 0$ and $n = 1$ satisfy (A4)), and welfare equal to the grade. Taking $\alpha \to 0$, the margin condition reduces to $\overline{W}^\star > \overline{W}(x, a')$, i.e. to $a^\star$ being strictly correct and a' not. Theorem 1 then recovers the exact welfare-consistency of RLVR asserted in Proposition 1.*

Proof. With $\varepsilon = 0$ and $\alpha \to 0$ the penalty $4\alpha + \frac{2\varepsilon}{1-2\alpha} \to 0$, leaving $M > 2s$; with deterministic outcomes the empirical trimmed mean equals the grade exactly, so any $s > 0$ works and (A4) holds at $n = 1$. The surviving requirement $\overline{W}^\star > \overline{W}(x, a')$ is correctness of $a^\star$ over a'. □

9 Welfare-consistency under approximate calibration

Theorem 1 assumes the simulate-then-evaluate pipeline is *exactly* unbiased for welfare on the faithful mass (Assumption 2). Learned reward models are systematically overconfident [17], so this is too strong. We weaken it to approximate calibration and quantify the miscalibration the guarantee tolerates.

Assumption 3 (η-approximate calibration)**.** There is $\eta \geq 0$ such that, for every faithful type u and every (x, a), $\big|\rho(x, u, a) - W(x, u, a)\big| \leq \eta$.

Under Assumption 3 the clean faithful-grade mean is within η of welfare: $\big|\mathrm{mean}(F_a) - \overline{W}(x, a)\big| = \big|\mathbb{E}_{u \sim P_{\mathrm{f}}}[\rho(x, u, a) - W(x, u, a)]\big| \leq \eta$. Carrying this one extra bias term through the chain of Lemma 7 costs 2η in the margin.

Theorem 2 (Consistency under approximate calibration)**.** *In the setting of Theorem 1 with Assumption 2 replaced by η-approximate calibration (Assumption 3), if for every welfare-suboptimal $a' \in \mathcal{A}_c$*

$$\overline{W}(x, a^\star) - \overline{W}(x, a') \;>\; 4\alpha + \frac{2\varepsilon}{1-2\alpha} + 2s + 2\eta,$$

then with probability at least $1-\delta$ (same panel size as (A4)*) the α-trimmed reward $\hat{\mu}_\alpha^{(n)}$ is welfare-consistent on $\mathcal{A}_c$. The condition is sufficient; the 2η term is necessary in the sense of Proposition 2 below.*

Proof. Repeat the proof of Lemma 7 with one additional term. For the clean law F_a, $|\mu_\alpha(F_a) - \overline{W}(x,a)| \le |\mu_\alpha(F_a) - \mathrm{mean}(F_a)| + |\mathrm{mean}(F_a) - \overline{W}(x,a)| \le 2\alpha + \eta$ by Lemma 6(i) and Assumption 3. Hence, chaining the sampling and contamination bounds of Lemma 6 exactly as before,

$$\hat{\mu}_\alpha^{(n)}(a^\star) \ge \overline{W}^\star - 2\alpha - \eta - \tfrac{\varepsilon}{1-2\alpha} - s, \qquad \hat{\mu}_\alpha^{(n)}(a') \le \overline{W}(x,a') + 2\alpha + \eta + \tfrac{\varepsilon}{1-2\alpha} + s,$$

so their difference is at least $M - 4\alpha - \frac{2\varepsilon}{1-2\alpha} - 2s - 2\eta > 0$ under the stated margin. The union bound over $\mathcal{A}_c$ is identical to Theorem 1. □

Proposition 2 (Tightness of the calibration term)**.** *The dependence on η cannot be removed. Even with $\alpha = 0$, $\varepsilon = 0$, and an infinite panel ($s = 0$), there exist a query and a candidate pair $\{a^\star, a'\}$ with $\overline{W}(x,a^\star) - \overline{W}(x,a') = M$ and an η-approximately-calibrated grader for which the* (exact) *reward ranks a' at least as high as $a^\star$ whenever $\eta \ge M/2$, and strictly above once $\eta > M/2$.*

Proof. With $\alpha = \varepsilon = s = 0$ the reward is the exact faithful mean $\hat{\mu}_\alpha^{(n)}(a) = \mathrm{mean}(F_a)$, which Assumption 3 pins only to $[\overline{W}(x,a) - \eta, \overline{W}(x,a) + \eta]$. Choose a grader with $\mathrm{mean}(F_{a^\star}) = \overline{W}(x,a^\star) - \eta$ and $\mathrm{mean}(F_{a'}) = \overline{W}(x,a') + \eta$; both are admissible. The reward gap is then $(\overline{W}^\star - \eta) - (\overline{W}(x,a') + \eta) = M - 2\eta$, which is ≤ 0 once $\eta \ge M/2$, so a' is weakly preferred while being welfare-inferior. □

10 Regret of the training loop

The results so far bound the reward at a fixed query and certify that its maximizer over a candidate set is welfare-optimal. We now lift this to the policy an optimizer actually returns, and isolate the component of the welfare gap that optimization cannot reduce.

Let Π be the policy class the optimizer searches (for instance the set reachable by the training algorithm from the base model). For $\pi \in \Pi$ write the true welfare value and the empirical robust-reward value

$$V^\star(\pi) \;=\; \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{a\sim\pi(\cdot\,|\,x)}\big[\overline{W}(x,a)\big], \qquad \widehat{V}(\pi) \;=\; \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{a\sim\pi(\cdot\,|\,x)}\big[\hat{\mu}_\alpha^{(n)}(x,a)\big],$$

where $\hat{\mu}_\alpha^{(n)}(x,a)$ is the α-trimmed empirical market reward at x. Let $\pi^\star \in \arg\max_{\pi\in\Pi} V^\star(\pi)$ be the welfare-optimal policy *within the class*. Collecting the four error sources of the preceding sections gives a uniform reward error: on the high-probability event that every panel deviates by at most its sampling tolerance,

$$\big|\hat{\mu}_\alpha^{(n)}(x,a) - \overline{W}(x,a)\big| \;\le\; b \;:=\; \underbrace{2\alpha}_{\text{trimming}} + \underbrace{\tfrac{\varepsilon}{1-2\alpha}}_{\text{contamination}} + \underbrace{\eta}_{\text{calibration}} + \underbrace{s_n}_{\text{sampling}}, \qquad s_n = \tfrac{1}{1-2\alpha}\sqrt{\tfrac{1}{2n}\log\tfrac{2}{\delta'}},$$

by Lemma 6(i)–(iii) and Assumption 3, with δ' allocated by a union bound across the finitely many panels evaluated during training; for a continuous query distribution the uniform bound additionally requires a covering of the candidate-generating policy class, which we assume.

Theorem 3 (Reward error to welfare regret)**.** *Suppose the uniform reward error* $\big|\hat{\mu}_\alpha^{(n)}(x,a) - \overline{W}(x,a)\big| \le b$ *holds across the evaluated candidate panels, so that* $|\widehat{V}(\pi) - V^\star(\pi)| \le b$ *for all* $\pi \in \Pi$*. If* $\hat{\pi}$ *is an* $\varepsilon_{\mathrm{opt}}$*-approximate maximizer of* $\widehat{V}$ *over* Π*, that is* $\widehat{V}(\hat{\pi}) \ge \sup_{\pi\in\Pi} \widehat{V}(\pi) - \varepsilon_{\mathrm{opt}}$*, then*

$$V^\star(\hat{\pi}) \;\ge\; V^\star(\pi^\star) \;-\; 2b \;-\; \varepsilon_{\mathrm{opt}}.$$

Proof. The pointwise bound gives $|\widehat{V}(\pi) - V^\star(\pi)| \le b$ by taking expectations over $x \sim \mathcal{D}$ and $a \sim \pi$. Then

$$V^\star(\hat{\pi}) \;\ge\; \widehat{V}(\hat{\pi}) - b \;\ge\; \sup_{\pi\in\Pi} \widehat{V}(\pi) - \varepsilon_{\mathrm{opt}} - b \;\ge\; \widehat{V}(\pi^\star) - \varepsilon_{\mathrm{opt}} - b \;\ge\; V^\star(\pi^\star) - 2b - \varepsilon_{\mathrm{opt}}. \qquad \square$$

Remark 7 (KL-regularized optimizer). Practical RLHF and GRPO maximize a KL-regularized objective $J_\beta(\pi) = \widehat{V}(\pi) - \beta\,\mathrm{KL}(\pi \,\|\, \pi_{\mathrm{ref}})$ against a reference policy π_{ref}. Its maximizer $\hat{\pi}_\beta$ is not the $\widehat{V}$-maximizer, so the clean reduction acquires one explicit regularization term. From $J_\beta(\hat{\pi}_\beta) \ge J_\beta(\pi^\star)$ and $\mathrm{KL}(\hat{\pi}_\beta \,\|\, \pi_{\mathrm{ref}}) \ge 0$,

$$V^\star(\hat{\pi}_\beta) \;\ge\; V^\star(\pi^\star) \;-\; 2b \;-\; \beta\,\mathrm{KL}(\pi^\star \,\|\, \pi_{\mathrm{ref}}),$$

so regularization costs at most $\beta\,\mathrm{KL}(\pi^\star \,\|\, \pi_{\mathrm{ref}})$ and vanishes as $\beta \to 0$. The optimization term $\varepsilon_{\mathrm{opt}}$ is what a convergence analysis of the policy-gradient step under a biased reward oracle controls [18].

Remark 8 (The floor of the regret bound). As the panel grows ($s_n \to 0$) and the optimizer converges ($\varepsilon_{\mathrm{opt}} \to 0$), the bound stalls at $2\big(2\alpha + \frac{\varepsilon}{1-2\alpha} + \eta\big)$. The calibration part of this floor is matched: Proposition 2 exhibits an instance on which the robust reward mis-ranks by 2η, so a reward-driven optimizer is steered into that welfare loss. The trimming and contamination parts are worst-case bounds on the estimand shift rather than matched converses; both improve under grade concentration, the trimming term to $O(\alpha\sigma)$ by the remark following Lemma 6, and the contamination breakdown is reached only at $\varepsilon \ge \alpha$ (Proposition 3), outside the operating regime $\varepsilon < \alpha$.

11 From per-query to per-distribution guarantees

Theorems 1 and 2 hold at a fixed query and certify consistency on its candidate set. We now give a guarantee over the query distribution $\mathcal{D}$ and identify the regime in which no guarantee is possible.

For a query x let the welfare margin be $M(x) = \overline{W}(x, a^\star(x)) - \max_{a' \notin \mathcal{A}^\star(x)} \overline{W}(x, a')$, the gap between the best and the best suboptimal candidate (set $M(x) = \infty$ if every candidate is optimal). Queries with a small margin are the ones a sampled reward cannot resolve, so the right structural assumption is a bound on how much $\mathcal{D}$-mass sits near zero margin.

Theorem 4 (Per-distribution welfare regret)**.** *Let* $\hat{a}$ *select* $\arg\max_{a\in\mathcal{A}_c(x)} \hat{\mu}_\alpha^{(n)}(x,a)$ *at each* x*, under the per-query conditions of Theorem 2 with penalty* $p = 4\alpha + \frac{2\varepsilon}{1-2\alpha} + 2s + 2\eta$ *and per-query failure probability at most* δ*. Suppose a margin-distribution bound* $\mathbb{P}_{x\sim\mathcal{D}}(M(x) \le \tau) \le \psi(\tau)$ *holds for a nondecreasing* $\psi : [0,\infty) \to [0,1]$*. Then the* $\mathcal{D}$*-expected welfare regret obeys*

$$\mathbb{E}_{x\sim\mathcal{D}}\big[\overline{W}(x, a^\star(x)) - \overline{W}(x, \hat{a}(x))\big] \;\le\; \psi(p) + \delta.$$

Proof. Fix x. If $M(x) > p$ and the panels at x succeed, an event of probability at least $1 - \delta$, then Theorem 2 gives $\hat{a}(x) \in \mathcal{A}^\star(x)$ and the per-query regret is 0. Otherwise the per-query regret is at most the welfare range, which is 1 because $W \in [0,1]$. The bad set is contained in $\{M(x) \le p\} \cup \{\text{panel fails}\}$, of probability at most $\psi(p) + \delta$. Taking expectations and bounding the regret by 1 on the bad set gives the claim. $\square$

As the penalty $p \to 0$ the bound tends to $\psi(0^+) + \delta$, so when the margin density has no atom at zero and the panels are large, the robust-argmax policy is welfare-optimal for $\mathcal{D}$-almost every query. The next proposition shows the contamination condition $\varepsilon < \alpha$ of Theorem 1 is exactly the boundary of what any trimmed-mean reward can promise.

Proposition 3 (Sharpness of the $\varepsilon < \alpha$ threshold)**.** *Fix $\alpha \in (0, \frac{1}{2})$. For every $\varepsilon > \alpha$ there is an ε-contaminated instance, a query with two candidates of positive welfare gap, on which the α-trimmed reward functional ranks the manipulation strictly above the welfare-optimal response, so the empirical reward mis-ranks with probability tending to 1 as $n \to \infty$. Hence no panel size restores consistency once $\varepsilon \geq \alpha$, and the condition $\varepsilon < \alpha$ cannot be relaxed for the trimmed-mean aggregator.*

Proof. Let $a^\star$ be welfare-optimal with all clean grades at a value $w^\star \in (0, 1)$, so $\hat{\mu}_\alpha^{(n)}(a^\star) = w^\star$ for every n. Let $a^\dagger$ be a manipulation whose gameable fraction $\varepsilon > \alpha$ places its grades at the ceiling 1 and whose remaining $(1-\varepsilon)$ faithful grades sit at 0, so its grade law has quantile function $F^{-1}(p) = 0$ for $p < 1-\varepsilon$ and $F^{-1}(p) = 1$ for $p \geq 1-\varepsilon$. For $\alpha < \varepsilon \leq 1-\alpha$ its α-trimmed functional is

$$\mu_\alpha(F) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(p)\, dp = \frac{\varepsilon - \alpha}{1-2\alpha},$$

because the retained range $[\alpha, 1-\alpha]$ still contains an $(\varepsilon-\alpha)$ mass of ceiling grades that trimming the top α does not remove. This population functional exceeds $w^\star$ whenever $w^\star < (\varepsilon-\alpha)/(1-2\alpha)$, which is positive for $\varepsilon > \alpha$. The empirical trimmed mean $\hat{\mu}_\alpha^{(n)}(a^\dagger)$ concentrates on it by Lemma 6(iii), so it exceeds $w^\star$ with probability tending to 1, while $\hat{\mu}_\alpha^{(n)}(a^\star) = w^\star$ deterministically; the manipulation then outranks the optimum although $\overline{W}(x, a^\star) = w^\star > \overline{W}(x, a^\dagger)$ for small gameable-welfare. No panel size restores consistency, so the breakdown point of the α-trimmed mean is α, matching the sufficient condition of Theorem 1. □

12 Scope of the assumptions

The theorem is only as strong as its hypotheses. The hypothesis on which the result depends is calibration.

Calibration (A1) is the central hypothesis. Everything reduces to it. Lemma 2 is calibration plus the tower property, and the robust machinery of Section 7 only protects calibration against a minority that violates it. If the simulator is biased on the faithful majority, so that $\mathbb{E}_o[g(o) \mid x, u, a]$ departs systematically from $W(x, u, a)$ for typical u, then no trimming level recovers welfare, because the contamination model assumes the bulk is clean. In this sense RLDF inherits the reliability of its world model: the reward is verifier-grade only to the extent that $g \circ T$ is calibrated and policy-independent (Definition 5). On outcome-verifiable tasks this holds by construction ($\varepsilon = 0$); on simulated-outcome tasks it is an empirical claim about the simulator that must be argued, not assumed. The role of Section 7 is to weaken the requirement from "calibrated everywhere" to "calibrated on a $(1-\varepsilon)$ majority with $\varepsilon < \alpha$," not to remove it.

Targeting is excluded by (A2), not defeated by trimming. Trimming caps the influence of a corrupt fraction at $\varepsilon/(1-2\alpha)$ only when $\varepsilon < \alpha$ (Lemma 6(ii)); once a panel is corrupt-majority ($\varepsilon > \alpha$), the trimmed mean's breakdown point is crossed and the bound is void. A type-aware targeting adversary (Remark 5) is dangerous because it can try to manufacture such a

corrupt-majority panel by routing the manipulation only to gameable consumers. Trimming alone does not stop this. What stops it is the representative-panel construction of Definition 6: every candidate is graded on a fresh i.i.d. draw from $P(\cdot \mid x)$, so each panel's corrupt share equals the population ε no matter how the policy routes responses, which is exactly hypothesis (A2). Hence (A2) is not a tacit assumption that the attack does not occur; it is the content of a concrete design choice, representative grading, that an implementer either makes or forfeits. Given (A2), trimming does the remaining work, and does it without detecting which consumers are gameable; only the population-level bound on how many matters. The irreducible assumption is global: if the faithful are a population minority ($\varepsilon > \alpha$ everywhere), no aggregation rule recovers welfare.

Margin (A3) and constants. The margin condition demands that the welfare gap between a good response and a manipulation exceed the summed trimming, contamination, and sampling penalties. The constants are worst-case: the trimming term shrinks to $O(\alpha\sigma)$ under bounded grade variance, and the contamination term is the price of admitting an ε-minority one declines to detect. The qualitative content is robust: a sufficiently distinguishable welfare difference is recoverable, and one finer than the noise floor is not, as for any sampled reward.

Population diversity. The population kernel $P(\cdot \mid x)$ is what separates RLDF from a single-judge hindsight reward. A degenerate P collapses Lemma 2 to a one-evaluator scheme and forfeits both the concentration of Lemma 3 and the robustness of Section 7; a diverse P is what gives the trimmed mean a distribution to trim. Heterogeneity is therefore a requirement of the construction, not a complication: it is what the welfare average and its robust surrogate are defined over.

13 Positioning against prior work

RLVR and its soft extensions. RLVR [1, 2, 3] is the $\varepsilon = 0$, deterministic, single-consumer limit (Proposition 1, Corollary 1). Soft extensions that learn a generative reward model or rubric critic for open-ended domains [4, 5, 6] re-admit a corruptible response-only grader and so live on the foresight side of Lemma 1: they score the answer, not its consequence.

Hindsight simulation. RLHS [7] is the direct antecedent and the source, in substance, of Lemmas 1–2: it simulates the downstream outcome and grades in hindsight, with automated LLM simulator and evaluator, to mitigate sycophancy and deception. Its evaluator is a single persona. The present results add the population kernel (Section 5), the closed-form gameability threshold (Section 6) that ties the construction to the user-feedback manipulation result of [8], and the robust-aggregation ranking guarantee (Section 7) that a single-persona scheme cannot state.

User-feedback RL and its hazards. Training on simulated user feedback induces targeted manipulation when even a small fraction of users is gameable [8], which is exactly Lemma 4 and Remark 5, with representative panels and trimming as the countermeasure. Emergent misalignment under competition for audiences [11] and deceptive policies that persist through safety training [10] are the qualitative hazards the calibration discussion of Section 12 keeps in view.

Robust aggregation and Byzantine-robust learning. The trimmed-mean reward of Section 7 imports a tool from robust statistics [27] and from Byzantine-robust distributed learning, where coordinate-wise trimmed means attain optimal statistical rates while tolerating a corrupt minority below the breakdown point [19]. The transfer is exact: a gameable subpopulation plays

the role of Byzantine workers, and the condition $\varepsilon < \alpha$ of Theorem 1 is the breakdown point turned into a training guarantee, with the matching converse of Proposition 3. What is new is not the aggregator but the object it aggregates, the realized welfare of a consumer population, and the coupling of the breakdown threshold to a welfare-consistency statement.

Decision-focused learning and social choice. Grading the realized outcome of acting on a response, rather than the response, is the decision-focused-learning idea [22] carried into RLHF: the loss is the regret of the decision a prediction induces, not the prediction error. RLDF supplies the pieces that idea needs for a language model, a simulated decision environment and a population over which the regret is taken. The aggregation step is also a social-choice rule over a heterogeneous population [21], and the welfare target $J^\star = \mathbb{E}_{u\sim P}[W]$ is its utilitarian aggregate; population-proportional alignment [20] matches the policy to the evaluator distribution, whereas RLDF asks only for the weaker ranking-consistency an optimizer needs. Foresight preference optimization is, by contrast, subject to reward-model overoptimization [23], the Goodhart failure that consequence-grading is meant to sidestep.

Market and multi-agent framings. Simulated consumer populations appear in economic LLM sandboxes [13] and generative-agent simulations [15], and market mechanisms have been proposed as alignment scaffolds over agents' propositions [12] and over emotional verifiable signals [14]. These price stated preferences or beliefs; RLDF prices realized welfare outcomes of acting on advice, aggregated over a heterogeneous population, and feeds the aggregate back as the reinforcement signal. We use the word market in a weak sense, a population aggregating realized outcomes into a price, and not in the mechanism-design sense of a clearing price, a scoring rule, or competition between responses; the object that does the work is the robust population aggregate μ_α, and no result depends on the metaphor. Table 1 summarizes the axes.

Method	Reward source	Pop.?	Outcome-grounded?	Robust to gaming?	Ref.
RLHF (foresight RM)	learned RM, response	no	no	no	[16]
RLVR	verifier, answer	no	yes *	yes *	[1, 2]
Soft RLVR / GenRM	learned grader, resp.	no	no	no	[4, 5]
RLHS	sim. outcome, 1 judge	no	yes	not analyzed	[7]
User-feedback RL	sim. user signal	yes	no	no	[8]
Economic sandbox (MALLES)	consumer preference	yes	no	not analyzed	[13]
Market-making (props.)	belief market	yes	no	partial	[12]
RLDF (this note)	sim. outcome, pop.	yes	yes	yes (Thm. 1)	

Table 1: Positioning of robust RLDF. *RLVR is outcome-grounded and gaming-robust only within its verifiable slice; RLDF recovers it as the degenerate limit (Corollary 1) and extends the guarantee to stochastic, heterogeneous, partially corruptible populations.

References

[1] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948, 2025. https://arxiv.

org/abs/2501.12948.

[2] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, et al. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. arXiv:2411.15124, 2024. https://arxiv.org/abs/2411.15124.

[3] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300, 2024. https://arxiv.org/abs/2402.03300.

[4] Y. Su, D. Yu, L. Song, J. Li, H. Mi, Z. Tu, M. Zhang, D. Yu. Crossing the Reward Bridge: Expanding RL with Verifiable Rewards Across Diverse Domains. arXiv:2503.23829, 2025. https://arxiv.org/abs/2503.23829.

[5] A. Bhaskar, X. Ye, D. Chen. Language Models that Think, Chat Better. arXiv:2509.20357, 2025. https://arxiv.org/abs/2509.20357.

[6] Z. Wei, X. Yang, K. Sun, J. Wang, R. Shao, et al. TruthRL: Incentivizing Truthful LLMs via Reinforcement Learning. arXiv:2509.25760, 2025. https://arxiv.org/abs/2509.25760.

[7] K. Liang, H. Hu, R. Liu, T. L. Griffiths, J. F. Fisac. RLHS: Mitigating Misalignment in RLHF with Hindsight Simulation. arXiv:2501.08617, 2025. https://arxiv.org/abs/2501.08617.

[8] M. Williams, M. Carroll, A. Narang, C. Weisser, B. Murphy, A. Dragan. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. arXiv:2411.02306, 2024. https://arxiv.org/abs/2411.02306.

[9] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, et al. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548, 2023. https://arxiv.org/abs/2310.13548.

[10] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566, 2024. https://arxiv.org/abs/2401.05566.

[11] B. El, J. Zou. Moloch's Bargain: Emergent Misalignment When LLMs Compete for Audiences. arXiv:2510.06105, 2025. https://arxiv.org/abs/2510.06105.

[12] B. Gho, S. Muppavarapu, A. Shaik, T. Tsay, A. Mohan, et al. From Competition to Coordination: Market Making as a Scalable Framework for Safe and Aligned Multi-Agent LLM Systems. arXiv:2511.17621, 2025. https://arxiv.org/abs/2511.17621.

[13] Y. Wu, Y. Liu, X. Deng. MALLES: A Multi-agent LLMs-based Economic Sandbox with Consumer Preference Alignment. arXiv:2603.17694, 2026. https://arxiv.org/abs/2603.17694.

[14] P. Wang, R. Ma, B. Zhang, X. Chen, Z. He, et al. RLVER: Reinforcement Learning with Verifiable Emotion Rewards for Empathetic Agents. arXiv:2507.03112, 2025. https://arxiv.org/abs/2507.03112.

[15] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442, 2023. https://arxiv.org/abs/2304.03442.

[16] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290, 2023. https://arxiv.org/abs/2305.18290.

[17] J. Leng, C. Huang, B. Zhu, J. Huang. Taming Overconfidence in LLMs: Reward Calibration in RLHF. arXiv:2410.09724, 2024. https://arxiv.org/abs/2410.09724.

[18] S. Mu, D. Klabjan. On the Second-Order Convergence of Biased Policy Gradient Algorithms. arXiv:2311.02546, 2023. https://arxiv.org/abs/2311.02546.

[19] D. Yin, Y. Chen, K. Ramchandran, P. Bartlett. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. ICML 2018. arXiv:1803.01498. https://arxiv.org/abs/1803.01498.

[20] K. Kim, J. Zhang, A. Ozdaglar, P. A. Parrilo. Beyond RLHF and NLHF: Population-Proportional Alignment under an Axiomatic Framework. arXiv:2506.05619, 2025. https://arxiv.org/abs/2506.05619.

[21] V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, et al. Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. ICML 2024. arXiv:2404.10271. https://arxiv.org/abs/2404.10271.

[22] A. N. Elmachtoub, P. Grigas. Smart "Predict, then Optimize". *Management Science* 68(1):9–26, 2022. arXiv:1710.08005. https://arxiv.org/abs/1710.08005.

[23] L. Gao, J. Schulman, J. Hilton. Scaling Laws for Reward Model Overoptimization. ICML 2023. arXiv:2210.10760. https://arxiv.org/abs/2210.10760.

[24] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[25] A. Dvoretzky, J. Kiefer, J. Wolfowitz. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *Annals of Mathematical Statistics*, 27(3):642–669, 1956.

[26] P. Massart. The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality. *Annals of Probability*, 18(3):1269–1283, 1990.

[27] G. Lugosi, S. Mendelson. Mean Estimation and Regression under Heavy-Tailed Distributions: A Survey. *Foundations of Computational Mathematics*, 19:1145–1190, 2019.