SAFE RLHF BEYOND EXPECTATION: STOCHASTIC DOMINANCE FOR SPECTRAL RISK CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe Reinforcement Learning from Human Feedback (RLHF) typically enforces safety by constraining the expected cost of a policy. However, expected cost captures only average behavior and can overlook rare but severe failures. We propose Risk-sensitive Alignment via Dominance (RAD), a distributional approach to safe RLHF that compares the cost distribution of the learned policy to that of a reference policy rather than only their means. RAD represents policy-induced cost distributions non-parametrically through quantiles and encourages the learned policy to have lower cost quantiles than the reference policy. To optimize this distributional objective for language models, RAD uses a quantile-particle representation that connects two gradient estimators: entropic optimal transport provides gradients of the distributional objective with respect to quantile particles, while score-function quantile gradients propagate these updates to the policy. We further generalize RAD with quantile weighting, allowing different regions of the cost distribution to be emphasized. For admissible spectral weights, the resulting objective has a direct connection to Spectral Risk Measures, providing a principled way to target different notions of risk. Empirically, on BeaverTails, RAD improves harmlessness over SFT, and several variants also improve harmlessness over Safe RLHF while maintaining similar helpfulness. On held-out HarmBench prompts, several risk-sensitive variants also improve harmlessness over Safe RLHF.

1 INTRODUCTION

Large Language Models (LLMs) are used in high-stakes domains such as legal reasoning (Katz et al. (2024)), medical consultation (Yang et al. (2022); Moor et al. (2023)), and educational support (Kasneci et al. (2023); Kung et al. (2023)). In these settings, outputs should be helpful and safe because harmful generations, including toxicity and misinformation, can have serious consequences (Gehman et al. (2020); Weidinger et al. (2021); Ganguli et al. (2022)).

Helpfulness and harmlessness can conflict when users request assistance with potentially harmful activities (Bai et al. (2022a); Glaese et al. (2022)). Reinforcement Learning from Human Feedback (RLHF) typically optimizes a single reward model for both objectives (Ouyang et al. (2022b); Bai et al. (2022b)) or heuristically mixes separate reward models (Glaese et al. (2022); Touvron et al. (2023); Mu et al. (2024)); stronger harmlessness can lead to refusals, while stronger helpfulness can increase unsafe outputs (Bai et al. (2022b)). Safe RLHF separates preference data and constrains harmlessness (Dai et al. (2023b)), but its expected-cost constraint does not directly control worst-case or tail outcomes. This matters in code generation (Pearce et al., 2021) and open-ended dialogue, where rare but severe harms include toxic generations and private data leakage (Perez et al., 2022; Gehman et al., 2020).

Thus, we argue that a stronger notion of safety is needed: the learned policy should have a cost distribution that is *stochastically smaller* than that of the reference policy, rather than merely lower in expectation. Intuitively, this means that safety improvements should extend across the cost distribution, including its high-cost tail, instead of arising only from a reduction in average cost. This motivates **Risk-sensitive Alignment via Dominance (RAD)**, which uses First-Order Stochastic Dominance (FSD) (Dai et al., 2023a) as a principled way to compare the cost distributions induced by the learned and reference policies. RAD operationalizes this idea through their cost quantiles, encouraging the learned policy to incur lower costs across quantile levels.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A key advantage of the quantile-based formulation is that different regions of the cost distribution can be emphasized according to the deployment's risk priorities. Rather than weighting all quantiles equally, RAD allows higher-cost regions, such as the upper tail, to receive greater emphasis. This connects naturally to Spectral Risk Measures (SRMs)(Acerbi, 2002), which quantify risk by weighting different cost quantiles according to a chosen risk profile. As a result, the same framework can recover average-cost sensitivity with uniform weighting or place greater emphasis on tail risk through choices such as CVaR-inspired weighting. A practical advantage is that these different risk preferences are handled within the same optimization framework: changing the target risk profile only requires changing the quantile-weighting function, while the underlying RAD optimization procedure remains unchanged. We characterize how weighting the quantile differences between the learned and reference cost distributions relates to the corresponding spectral risk measure, providing a principled way to adapt the safety objective to the risks of a deployment setting.

RAD represents policy-induced cost distributions with quantile particles. An entropic Optimal Transport (OT) formulation (Cuturi, 2013; Melnyk et al., 2024; Santambrogio, 2015) provides particle gradients computed using Sinkhorn iterations (Cuturi, 2013; Peyré & Cuturi, 2020), while score-function quantile gradients (Li et al., 2024) propagate these updates to the policy in an approximate REINFORCE-style update (Ahmadian et al., 2024; Williams, 1992; Kool et al., 2019).

Our main contributions are: (i) we formulate safe alignment with a distributional constraint that compares policy-induced cost distributions rather than relying only on expected cost; (ii) we develop a practical optimization procedure that combines entropic OT particle gradients with score-function quantile gradients in a REINFORCE-style policy update; (iii) we relate quantile reweighting in our framework to SRM differences, allowing multiple risk preferences to be handled through the same optimization procedure by changing only the weighting function; and (iv) empirically, RAD improves harmlessness over the supervised fine-tuned (SFT) model on BeaverTails (Ji et al., 2023), with several variants also improving harmlessness over Safe RLHF (Dai et al., 2023b) while maintaining competitive helpfulness. On held-out HarmBench (Mazeika et al., 2024) prompts, several variants achieve favorable pairwise harmlessness results against Safe RLHF.

2 Background and Problem Setting

2.1 Reinforcement Learning from Human Feedback (RLHF)

The most popular method of aligning LLMs with human preferences is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022b); its standard stages are detailed in Section B. Formally, let $x \sim D_x$ denote prompts and $y \sim \pi_\theta(\cdot \mid x)$ denote model responses. Supervised fine-tuning (SFT) produces a reference policy π_{sft}, and we set $\pi_{\text{ref}} := \pi_{\text{sft}}$ unless stated otherwise. A reward model $r_\phi(x, y)$ is learned from preference comparisons using the Bradley-Terry model (Bradley & Terry, 1952).

RLHF then optimizes a KL-regularized objective to prevent reward overoptimization and preserve language quality (Ouyang et al., 2022b; Stiennon et al., 2022; Gao et al., 2022; Rafailov et al., 2024):

$$\max_\theta \; \mathbb{E}_{x\sim D_x,\, y\sim\pi_\theta(\cdot|x)}\big[r_\phi(x,y)\big] \;-\; \beta\, D_{\text{KL}}\Big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\Big), \tag{1}$$

equivalently maximizing the KL-regularized reward $\tilde{r}(x, y) = r_\phi(x, y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. In practice, this objective is optimized with PPO/GRPO/REINFORCE-style methods (Williams, 1992; Ahmadian et al., 2024; Shao et al., 2024; Schulman et al., 2017).

2.2 Safe RLHF

Standard RLHF optimizes a single reward function learned from human preferences, which may be inadequate when attempting to balance competing goals such as *helpfulness* and *harmlessness*. In contrast, Safe RLHF (Dai et al., 2023b) separates helpfulness and harmlessness by learning (i) a reward model $r_\phi(x, y)$, (parameterized by ϕ and for input prompt x and model output y) from helpfulness preferences and (ii) a cost model $c_\psi(x, y)$, (parameterized by ψ) from harmfulness preferences, and then solving the expected-cost constrained RL problem:

$$\max_\theta \; \mathbb{E}\big[r_\phi(x,y)\big] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \quad \text{s.t.} \quad \mathbb{E}\big[c_\psi(x,y)\big] \le \tau. \tag{2}$$

2.3 First-Order Stochastic Dominance

The expected cost constraint in Safe RLHF summarizes policy-induced costs through a single statistic and can therefore be inadequate when safety depends on the broader shape of the cost distribution. A stronger criterion is for the learned policy's cost distribution to be stochastically smaller than that of the reference policy, meaning that its cost quantiles are no higher at any quantile level. We first describe this ordering formally. Let F_X and Q_X denote the Cumulative Distribution Function (CDF) and the Quantile Function (Inverse CDF) of a random variable X. For real-valued random variables X, Y, X is said to have first-order stochastic dominance (FSD) on Y (denoted by $X \succeq_{\mathrm{FSD}} Y$) iff $F_X(r) \leq F_Y(r)$ for all r (Dai et al., 2023a). Equivalently, we also have $Q_X(q) \geq Q_Y(q)$ for all $q \in [0, 1]$.

Practically, a drawback of FSD is that it defines only a partial ordering over distributions, so two random variables may be incomparable under FSD. Following prior work on stochastic-dominance optimization (Farajzadeh et al., 2026; Melnyk et al., 2024; Santambrogio, 2015), we consider the following one-sided quantile functional:

$$\mathcal{L}_{\mathrm{FSD}}(X, Y) := \int_0^1 \big(Q_Y(q) - Q_X(q)\big)_+ \, dq, \tag{3}$$

where $(x)_+ = \max(x, 0)$. This quantity measures the aggregate quantile-wise violation of X dominating Y, or equivalently the extent to which Y has higher quantiles than X. In particular, for distinct distributions, $\mathcal{L}_{\mathrm{FSD}}(X, Y) = 0$ implies $X \succeq_{\mathrm{FSD}} Y$. For positive values, however, $\mathcal{L}_{\mathrm{FSD}}(X, Y)$ captures only the regions where $Q_Y(q) > Q_X(q)$ and does not penalize differences in the opposite direction; consequently, a positive value does not by itself imply $Y \succeq_{\mathrm{FSD}} X$.

In our case, we set X to be the cost distribution C_{π_θ} under our learned policy π_θ and Y to be the cost distribution $C_{\pi_{\mathrm{ref}}}$ under the reference policy π_{ref}. Thus, $\mathcal{L}_{\mathrm{FSD}}(C_{\pi_\theta}, C_{\pi_{\mathrm{ref}}})$ aggregates the reductions in cost quantiles of the learned policy relative to the reference policy.[1].

2.4 Optimal Transport

Optimal transport compares distributions by finding a minimum-cost coupling (Peyré & Cuturi, 2020). For our empirical cost distributions, we use entropic regularization to obtain a smooth objective whose unique transport plan is computed by differentiable Sinkhorn iterations (Cuturi, 2013; Peyré & Cuturi, 2020); the discrete OT formulation and Sinkhorn details are given in Section C.

Relation to FSD. We now describe how our FSD objective relates to the optimal transport objective. Using the results from Santambrogio (2015) (Theorem 2.9 and Proposition 2.17) or Melnyk et al. (2024) (Theorem 1), for any two distributions X and Y, the FSD objective can be framed as an optimal transport problem. Specifically,

$$\mathcal{L}_{\mathrm{FSD}}(X, Y) = \mathrm{OT}_c(X, Y) \tag{4}$$

for the asymmetric convex cost function $c(x, y) = (y - x)_+$. We restate the relevant theorem in Section D for completeness. The distinction between Melnyk et al.'s preference-data score distributions and our policy-induced cost distributions is discussed in Section 5.

2.5 Spectral Risk Measures

Spectral Risk Measures (SRMs) summarize risk by taking a weighted average of the quantiles of a cost distribution (Acerbi, 2002). The weighting function, or risk spectrum, determines which parts of the distribution receive greater emphasis and is typically nonnegative, non-decreasing, and normalized to integrate to one. Under these conditions, SRMs form a class of coherent, law-invariant risk measures (Artzner et al., 1999; Acerbi, 2002). Their definition and connection to our weighted FSD objective are developed in Section 3.1, with further background in Section F.

[1]Note that $\mathcal{L}_{\mathrm{FSD}}(X, Y)$ is not a distance because it is not symmetric and can be zero even when X and Y have different distributions. Also $\mathcal{L}_{\mathrm{FSD}}(X, Y) + \mathcal{L}_{\mathrm{FSD}}(Y, X) = \mathcal{W}_1(X, Y)$ where $\mathcal{W}_1(X, Y)$ is the 1-Wasserstein distance

3 RISK-SENSITIVE ALIGNMENT VIA DOMINANCE

We now present **Risk-sensitive Alignment via Dominance (RAD)**. Rather than constraining only the expected cost, RAD compares the policy-induced cost distribution of the learned policy with that of a reference policy using the FSD objective introduced in Section 2.3. We first formulate this distributional safety constraint and derive a practical policy-optimization procedure using a quantile-particle representation. We then generalize the objective with quantile weighting and establish its connection to Spectral Risk Measures (SRMs). For a policy π, let $C_\pi := c_\psi(x, y)$ denote the random variable of costs induced by sampling $x \sim \mathcal{D}_x$ and $y \sim \pi(\cdot \mid x)$. We solve the following optimization problem:

$$\max_{\theta} \quad \mathbb{E}_{x\sim D_x,\, y\sim\pi_\theta(\cdot|x)}\Big[r_\phi(x,y)\Big] - \beta\, D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big), \tag{5}$$

$$\text{s.t.} \quad \mathcal{L}_{\mathrm{FSD}}\big(C_{\pi_\theta},\, C_{\pi_{\mathrm{ref}}}\big) \;\geq\; \kappa. \tag{6}$$

Constraint equation 6 rewards reductions in the learned policy's cost quantiles relative to the reference policy, weighting all quantile levels equally. Since

$$\mathcal{L}_{\mathrm{FSD}}(C_{\pi_\theta},\, C_{\pi_{\mathrm{ref}}}) = \int_0^1 \big(Q_{C_{\pi_{\mathrm{ref}}}}(q) - Q_{C_{\pi_\theta}}(q)\big)_+ \, dq,$$

larger values correspond to larger reductions in the learned policy's cost quantiles over regions where $Q_{C_{\pi_\theta}}(q) < Q_{C_{\pi_{\mathrm{ref}}}}(q)$. As in prior stochastic-dominance approaches based on asymmetric one-sided objectives (Farajzadeh et al., 2026; Melnyk et al., 2024), this provides a tractable relaxation of exact FSD rather than enforcing the full ordering directly. Quantile levels where the learned policy has higher cost than the reference are captured by the corresponding reverse violation, whose effect on spectral-risk improvement is characterized in Corollary 3 and evaluated empirically in Table 3.

To optimize our objective, we first absorb the KL term into the reward by defining the KL-regularized reward $\tilde{r}(x, y) = r_\phi(x, y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}$ and then use Dual Ascent (Gallier & Quaintance, 2019) to optimize the Lagrangian relaxation of equation 5–equation 6:

$$\max_{\theta} \min_{\lambda\geq 0} \underbrace{\mathbb{E}_{x\sim D_x,\, y\sim\pi_\theta(\cdot|x)}\Big[\tilde{r}(x,y)\Big] + \lambda\Big(\mathcal{L}_{\mathrm{FSD}}\big(C_{\pi_\theta}, C_{\pi_{\mathrm{ref}}}\big) - \kappa\Big)}_{L(\theta,\lambda)}. \tag{7}$$

Non-parametric (quantile-particle) representation of policy-induced cost distributions. We now describe how we represent the cost distributions for our policies. For a policy π, rather than assuming a parametric family for the cost distribution C_π, we use a non-parametric empirical approximation via particles in the *quantile space*. Fix quantile levels $\alpha_1, \ldots, \alpha_N \in (0, 1)$ and define the corresponding cost-quantile particles:

$$q_i(\pi) := Q_{C_\pi}(\alpha_i), \qquad i = 1, \ldots, N, \tag{8}$$

which we assume are ordered so that $q_1(\pi) \leq \cdots \leq q_N(\pi)$. We approximate the cost distribution by the empirical measure:

$$\hat{\mu}_\pi := \frac{1}{N} \sum_{i=1}^{N} \delta_{q_i(\pi)}. \tag{9}$$

This choice of particles lets us combine the OT particle gradient with quantile policy-gradient estimator of Li et al. (2024), derived in Section E. In practice, $\hat{\mu}_\pi$ is obtained by sampling generations from π, evaluating their costs via c_ψ, and computing empirical quantiles from the resulting samples; these quantiles serve as the particles defining the discrete representation. Let $\hat{\mu}_{\pi_\theta}$ and $\hat{\mu}_{\pi_{\mathrm{ref}}}$ be the empirical measures for the cost distributions corresponding to policies π_θ and π_{ref} respectively. Then, the FSD objective is approximated as $\mathcal{L}_{\mathrm{FSD}}\big(C_{\pi_\theta}, C_{\pi_{\mathrm{ref}}}\big) \approx \mathcal{L}_{\mathrm{FSD}}\big(\hat{\mu}_{\pi_\theta}, \hat{\mu}_{\pi_{\mathrm{ref}}}\big)$.

Computing the RAD policy gradients. The main technical challenge for computing the gradients of the RAD policy is differentiating the FSD term with respect to the policy parameters θ. Using the OT formulation in equation 4, we replace the FSD objective with its entropically regularized version $\mathcal{L}^{\chi}_{\mathrm{FSD}}$, where χ is the regularization parameter in equation 18. Entropic regularization gives smooth particle gradients, computed using Sinkhorn iterations; the derivation is in Appendix E.

Theorem 1 (RAD policy-gradient update)**.** *Fix $\lambda \geq 0$ and let $q_i(\pi_\theta) = Q_{C_{\pi_\theta}}(\alpha_i)$ denote the cost-quantile particles at fixed levels $\alpha_1, \ldots, \alpha_N$. Under the quantile-particle representation, the derivative of the distributional objective decomposes into a particle-gradient term with respect to $q_i(\pi_\theta)$ and a quantile policy-gradient term. Combining the entropic-OT particle gradient with the density-free quantile-gradient estimator yields the following REINFORCE-style update used by RAD. Moreover, for the weighted dominance objective $\mathcal{L}^w_{\mathrm{FSD}}$ (see Equation* (11)*), the same gradient uses an additional factor $w(\alpha_i)$ multiplying the i'th summand.*

$$\nabla_\theta \mathcal{L}(\theta, \lambda) \approx \mathbb{E}_{x \sim \mathcal{D}_x,\, y \sim \pi_\theta(\cdot|x)} \left[\left(\tilde{r}(x, y) + \lambda \sum_{i=1}^{N} \left(-\frac{\partial \mathcal{L}^\chi_{\mathrm{FSD}}}{\partial q_i(\pi_\theta)} \right) \mathbf{1}\left(c_\psi(x, y) \leq q_i(\pi_\theta)\right) \right) \nabla_\theta \log \pi_\theta(y \mid x) \right] \tag{10}$$

The particle-gradient term $\partial \mathcal{L}^\chi_{\mathrm{FSD}} / \partial q_i(\pi_\theta)$ is obtained from the entropic OT problem, while the indicator-weighted score-function term estimates how each cost quantile changes with the policy. Following the density-free quantile-gradient estimator of Li et al. (2024), we omit the unknown density factor at each quantile. Because these factors can differ across quantiles, the resulting multi-quantile update is the training direction used by RAD rather than the exact gradient of the original objective.

We optimize equation 7 by alternating (i) ascent in θ using equation 10 with REINFORCE (Williams, 1992) and variance-reduction scheme (RLOO) (Kool et al., 2019), and (ii) Dual Ascent (Gallier & Quaintance, 2019) in λ to enforce the constraint.

3.1 Weighted FSD and Spectral Risk Measures

We now describe how weighting the FSD objective relates to spectral risk measures. The FSD objective in Equation 3 weights all quantiles equally. However, in many applications, tail performance is of primary concern, and it may be desirable to emphasize higher quantiles more heavily than lower ones. To accommodate this, we introduce a quantile-weighted FSD objective with a nonnegative weighting function $w(q)$:

$$\mathcal{L}^w_{\mathrm{FSD}}(X, Y) = \int_0^1 w(q) \big(Q_Y(q) - Q_X(q)\big)_+ dq, \tag{11}$$

where $w(q) \geq 0$ for all $q \in [0, 1]$. A practical advantage of this formulation is that different risk preferences can be handled within the same optimization framework. Changing the target risk profile only requires changing the weighting function w; the optimization machinery remains unchanged. For the spectral-risk interpretation below, we take $\int_0^1 w(q)dq = 1$. While $\mathcal{L}^w_{\mathrm{FSD}}(X, Y)$ is not itself a spectral risk measure (SRM), it admits a direct relationship to quantile-weighted risk functionals. When w is also non-decreasing, as required for a risk spectrum, the associated spectral risk measure is $\rho_w(Z) := \int_0^1 Q_Z(q)\, w(q)\, dq$. The following proposition decomposes the difference in the associated risk functional into weighted FSD violations.

Proposition 2. *Let $w : (0, 1) \to \mathbb{R}_+$ be any nonnegative weight function with $\int_0^1 w(q)dq = 1$, and define $\rho_w(Z) := \int_0^1 Q_Z(q)\, w(q)\, dq$. For two random variables X and Y, define*

$$\mathcal{L}^w_{\mathrm{FSD}}(X, Y) := \int_0^1 \big(Q_Y(q) - Q_X(q)\big)_+ w(q)\, dq, \quad \mathcal{L}^w_{\mathrm{FSD}}(Y, X) := \int_0^1 \big(Q_X(q) - Q_Y(q)\big)_+ w(q)\, dq.$$

Then,

$$\rho_w(Y) - \rho_w(X) = \mathcal{L}^w_{\mathrm{FSD}}(X, Y) - \mathcal{L}^w_{\mathrm{FSD}}(Y, X). \implies -\mathcal{L}^w_{\mathrm{FSD}}(Y, X) \leq \rho_w(Y) - \rho_w(X) \leq \mathcal{L}^w_{\mathrm{FSD}}(X, Y).$$

This proposition shows that differences in *any* spectral risk measure can be decomposed into two weighted FSD violations when w is its risk spectrum. We refer to $\mathcal{L}^w_{\mathrm{FSD}}(X, Y)$ as the forward term and $\mathcal{L}^w_{\mathrm{FSD}}(Y, X)$ as the reverse violation. The forward term alone does not determine the risk difference; the reverse violation also contributes.

Corollary 3. *If $\mathcal{L}^w_{\mathrm{FSD}}(X, Y) \geq \kappa$ for some $\kappa > 0$, then*

$$\rho_w(Y) - \rho_w(X) = \mathcal{L}^w_{\mathrm{FSD}}(X, Y) - \mathcal{L}^w_{\mathrm{FSD}}(Y, X) \geq \kappa - \mathcal{L}^w_{\mathrm{FSD}}(Y, X).$$

In particular, if $Y \succeq_{\mathrm{FSD}} X$ so that $\mathcal{L}^w_{\mathrm{FSD}}(Y, X) = 0$, then $\rho_w(X) \leq \rho_w(Y) - \kappa$.

For the learned and reference cost distributions, set $X = C_{\pi_\theta}$ and $Y = C_{\pi_{\text{ref}}}$ in Corollary 3. For an admissible risk spectrum w, the weighted constraint implies a reduction in the corresponding spectral risk measure when κ exceeds the reverse violation $\mathcal{L}^w_{\text{FSD}}(C_{\pi_{\text{ref}}}, C_{\pi_\theta})$, and a reduction of at least κ when that violation vanishes. We evaluate the reverse violation in Table 3, where it is numerically negligible for the trained RAD policies.

Table 1 lists representative quantile-weighting functions, and Figure 2 illustrates them.[2] With the CVaR_α weighting and vanishing reverse violations, the constraint gives $\text{CVaR}_\alpha(C_{\pi_\theta}) \leq \text{CVaR}_\alpha(C_{\pi_{\text{ref}}}) - \kappa$. With uniform weighting, corresponding to the unweighted FSD objective, the same condition gives $\mathbb{E}[C_{\pi_\theta}] \leq \mathbb{E}[C_{\pi_{\text{ref}}}] - \kappa$.

Quantile Weighting	**Mean**	**VaR_α**	**CVaR_α**	**Linear Spectral**	**Exponential Spectral**	**Power Spectral**	**Wang Distortion**
$\mathbf{w(q)}$	1	$\delta(q-\alpha)$	$\frac{1}{1-\alpha}\mathbf{1}_{[\alpha,1]}(q)$	$2q$	$\frac{\lambda e^{\lambda q}}{e^\lambda - 1}$	$(1+\lambda)q^\lambda$	$\frac{\Phi(\Phi^{-1}(q)+\lambda)}{1-\Phi(\lambda)}$

Table 1: Examples of quantile weights $w(q)$ used in $\mathcal{L}^w_{\text{FSD}}$. Admissible risk spectra define corresponding spectral risk measures through ρ_w. $\lambda > 0$ denotes a risk aversion parameter and Φ in the Wang distortion (Wang, 1996) denotes the CDF of a standard normal distribution.

4 Empirical Results

We investigate the following research questions: [Q1] *How do models aligned using RAD compare to methods that enforce safety via expected cost constraints, in terms of helpfulness and harmfulness?* [Q2] *How do RAD-aligned models perform against baselines wrt harmlessness of responses, on out-of-distribution data?*

We fine-tune Qwen2.5-3B (Qwen et al., 2025) on Alpaca (Taori et al., 2023) to obtain the SFT initialization. RAD and Safe RLHF use the same reward and cost models trained on BeaverTails helpfulness and harmfulness preferences (Ji et al., 2023), following the fine-tuning and modeling procedure of Safe RLHF (Dai et al., 2023b). Safe RLHF is our principal trained baseline, and SFT is a standard reference. Preference-modeling details are in Appendix A.

In the RL stage, RAD uses the update in Equation (10); implementation details are in Appendix A. We set $\kappa = 10$ in Equation (6). Since we observed $\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\text{ref}}(\cdot|x)}[c_\psi(x,y)] \approx 0$, we set the Safe-RLHF cost threshold $\tau = -10$ in Equation (2) as an expected-cost counterpart to the RAD setting with $\kappa = 10$. This calibrates both methods to the same nominal 10-unit improvement relative to the reference policy, under their respective safety criteria.

4.1 Model Evaluations

We compare RAD variants using the weightings in Table 1 with Safe RLHF and SFT on BeaverTails test prompts.

In pairwise comparisons, red usually denotes a baseline (Safe RLHF or SFT) and blue a RAD variant; the assignment can vary across comparisons. Further protocol details are in Appendix A.

Harmlessness We use the trained cost model to judge the harmlessness of model generations for all considered approaches.

From Table 2, all RAD variants have a higher reported mean proportion of safe responses than SFT, while five of seven exceed Safe RLHF.

Let $C_{\pi_{\text{blue}}}$ and $C_{\pi_{\text{red}}}$ denote the random variables corresponding to the cost distributions induced by the blue and red models, respectively. We define the dominance difference as

$$D^w_{\text{FSD}}(C_{\pi_{\text{blue}}}, C_{\pi_{\text{red}}}) := \mathcal{L}^w_{\text{FSD}}(C_{\pi_{\text{blue}}}, C_{\pi_{\text{red}}}) - \mathcal{L}^w_{\text{FSD}}(C_{\pi_{\text{red}}}, C_{\pi_{\text{blue}}}).$$

By Proposition 2, when w is an admissible risk spectrum, $D^w_{\text{FSD}}(C_{\pi_{\text{blue}}}, C_{\pi_{\text{red}}})$ equals the reduction in the corresponding spectral risk measure from the red model to the blue model. Thus, a positive

[2] VaR is a quantile-weighting choice, not a coherent SRM; its point mass is Gaussian-smoothed in implementation. The Wang entry uses the distortion function directly rather than the derivative defining the Wang spectral weight. Its displayed scaling is not normalized.

Competition (Red vs Blue)	Reward of Blue Higher	Reward of Blue Lower	Overall Blue	Overall Red
safe-rlhf vs fsd-uniform	0.96 ± 0.01	0.83 ± 0.14	0.92 ± 0.05	0.93 ± 0.00
safe-rlhf vs fsd-var	0.88 ± 0.10	0.88 ± 0.10	0.88 ± 0.10	0.93 ± 0.00
safe-rlhf vs fsd-cvar	0.96 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.93 ± 0.00
safe-rlhf vs fsd-spectral-linear	0.97 ± 0.00	0.93 ± 0.04	0.94 ± 0.03	0.93 ± 0.00
safe-rlhf vs fsd-spectral-wang	0.97 ± 0.01	0.95 ± 0.00	0.96 ± 0.00	0.93 ± 0.00
safe-rlhf vs fsd-spectral-power	0.97 ± 0.01	0.94 ± 0.02	0.96 ± 0.01	0.93 ± 0.00
safe-rlhf vs fsd-spectral-exponential	0.98 ± 0.01	0.94 ± 0.03	0.96 ± 0.01	0.93 ± 0.00
sft vs fsd-uniform	0.95 ± 0.03	0.65 ± 0.16	0.92 ± 0.05	0.55 ± 0.00
sft vs fsd-var	0.93 ± 0.06	0.82 ± 0.10	0.88 ± 0.10	0.55 ± 0.00
sft vs fsd-cvar	0.98 ± 0.00	0.92 ± 0.01	0.97 ± 0.01	0.55 ± 0.00
sft vs fsd-spectral-linear	0.97 ± 0.01	0.74 ± 0.15	0.94 ± 0.03	0.55 ± 0.00
sft vs fsd-spectral-wang	0.97 ± 0.00	0.75 ± 0.03	0.96 ± 0.00	0.55 ± 0.00
sft vs fsd-spectral-power	0.97 ± 0.01	0.76 ± 0.08	0.95 ± 0.01	0.55 ± 0.00
sft vs fsd-spectral-exponential	0.98 ± 0.01	0.66 ± 0.17	0.96 ± 0.01	0.55 ± 0.00
sft vs safe-rlhf	0.96 ± 0.00	0.61 ± 0.03	0.93 ± 0.00	0.55 ± 0.00

Table 2: Comparison of the proportion of safe responses produced by the blue model, as judged by the cost model, across competing model pairs. We further decompose this proportion into two subsets based on whether the blue model's response achieved higher reward than the red model's response (first two columns). The "Overall Blue" column reports the overall proportion of safe responses for the blue model, while the "Overall Red" column reports the same for the red model. All RAD variants exceed SFT in the reported means; five of seven exceed Safe RLHF. Values are reported as mean $\pm$ standard deviation across 3 random seeds.

value indicates lower spectral risk for the blue model under the chosen weighting. Under uniform weighting, the dominance difference reduces to the difference in expected cost, so a positive value indicates lower average cost for the blue model.

Since RAD variants and Safe RLHF optimize different safety criteria, we report both the weighted dominance difference under each RAD weighting and its unweighted counterpart in Table 4. We use an unnormalized weighting in the weighted dominance difference; this changes its numerical scale but not its sign or comparisons for a fixed weighting.

From Table 4, all seven RAD variants have a positive reported mean weighted difference relative to SFT, while five of seven do so relative to Safe RLHF. The CVaR and spectral-power variants have positive weighted but negative unweighted differences against Safe RLHF. This shows that the comparison can differ depending on which regions of the cost distribution are emphasized. A positive weighted difference under a particular weighting does not, by itself, establish FSD.

Corollary 3 shows that the weighted RAD constraint depends on the corresponding reverse violation $\mathcal{L}^w_{\mathrm{FSD}}(C_{\pi_{\mathrm{ref}}}, C_{\pi_\theta})$. We therefore measure this quantity for the trained RAD policies. As shown in Table 3, the reverse violations are numerically negligible across the evaluated variants. Consequently, the reverse violation in Corollary 3 is negligible in these runs, and the corresponding risk difference is effectively determined by the forward term.

Weighting	Uniform	VaR	CVaR	Spectral-Linear	Spectral-Wang	Spectral-Power	Spectral-Exponential
Reverse violation	10^{-6}	10^{-8}	10^{-12}	10^{-7}	10^{-10}	10^{-7}	10^{-7}

Table 3: Reverse violations $\mathcal{L}^w_{\mathrm{FSD}}(C_{\pi_{\mathrm{ref}}}, C_{\pi_\theta})$ for trained RAD variants. The reverse-violation terms are numerically negligible across all evaluated variants.

Helpfulness We use the trained reward model to judge helpfulness. Figure 1 reports the percentage of prompts for which the blue model's response receives higher reward than the red model's response, with the number of responses compared in each cell. Its columns group comparisons by the (red, blue) safety outcomes (safe, safe), (safe, unsafe), (unsafe, safe), and (unsafe, unsafe); the final column gives the overall win rate.

From Figure 1, RAD-aligned models have high reward win rates against SFT. Against Safe RLHF, FSD-VaR, FSD-CVaR, and FSD-Spectral-Linear have lower reward win rates, consistent with these weighting schemes emphasizing harmlessness at some cost to helpfulness. The remaining variants have comparable reward win rates.

Summary Across BeaverTails, all RAD variants exceed SFT in reported mean safe-response proportion and have positive mean weighted dominance differences. Five of seven exceed Safe RLHF

Competition (Red vs Blue)	$D_{\text{FSD}}(C_{\pi_{\text{blue}}}, C_{\pi_{\text{red}}})$	$D^{w}_{\text{FSD}}(C_{\pi_{\text{blue}}}, C_{\pi_{\text{red}}})$
safe-rlhf vs fsd-uniform	$+7.92 \pm 11.23$	$+7.92 \pm 11.23$
safe-rlhf vs fsd-var	-66.06 ± 46.66	-2.95 ± 10.93
safe-rlhf vs fsd-cvar	-12.54 ± 2.80	$+56.61 \pm 10.39$
safe-rlhf vs fsd-spectral-linear	-7.33 ± 7.63	-0.49 ± 11.59
safe-rlhf vs fsd-spectral-wang	$+5.74 \pm 5.00$	$+27.91 \pm 14.99$
safe-rlhf vs fsd-spectral-power	-3.50 ± 6.34	$+9.45 \pm 8.67$
safe-rlhf vs fsd-spectral-exponential	$+8.18 \pm 11.04$	$+22.16 \pm 10.04$
sft vs fsd-uniform	$+147.12 \pm 12.43$	$+147.12 \pm 12.43$
sft vs fsd-var	$+73.06 \pm 46.52$	$+17.59 \pm 10.83$
sft vs fsd-cvar	$+126.84 \pm 1.68$	$+139.83 \pm 7.52$
sft vs fsd-spectral-linear	$+131.24 \pm 11.22$	$+135.64 \pm 15.54$
sft vs fsd-spectral-wang	$+144.63 \pm 1.56$	$+432.69 \pm 4.63$
sft vs fsd-spectral-power	$+135.91 \pm 5.54$	$+134.40 \pm 3.95$
sft vs fsd-spectral-exponential	$+147.88 \pm 9.84$	$+145.47 \pm 9.59$
sft vs safe-rlhf	$+138.51 \pm 2.45$	-

Table 4: Unweighted and weighted dominance differences for all pairs of competing models. A positive weighted difference indicates lower quantile-weighted cost for the blue model; when w is proportional to an admissible risk spectrum, it has the sign of the corresponding spectral-risk reduction. The weighted difference uses the weighting corresponding to the blue model. Against Safe RLHF, the CVaR and spectral-power variants have negative unweighted but positive weighted differences. Values are reported as mean $\pm$ standard deviation across 3 random seeds.

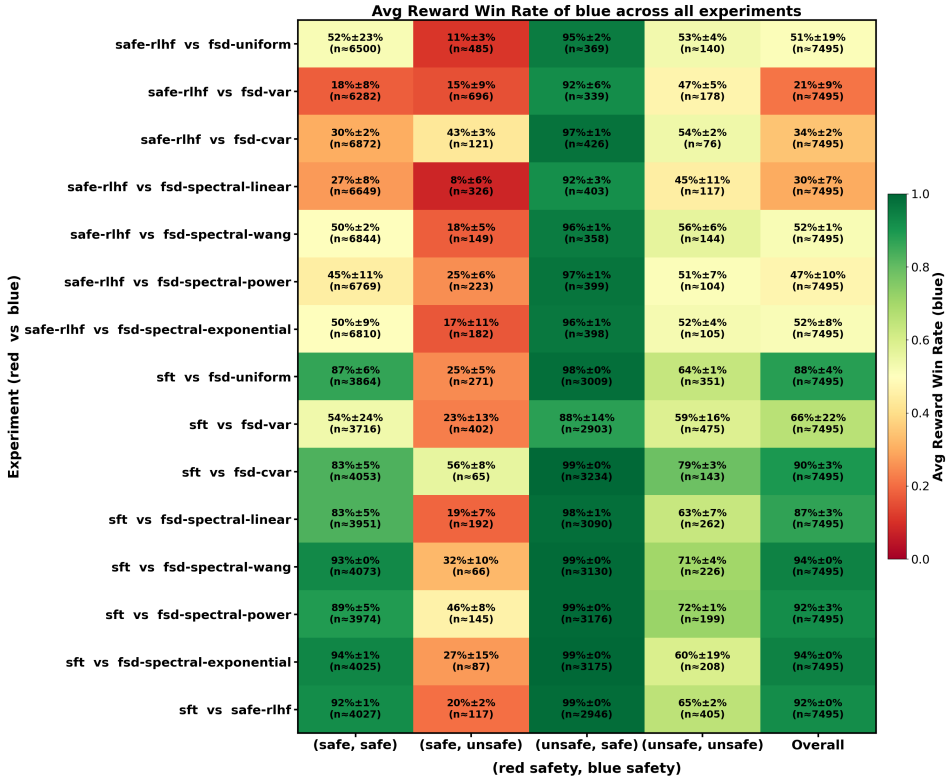


Figure 1: Average reward win rates between competing models, across 3 seeds.

in safe-response proportion, and five of seven have positive mean weighted differences against it. All variants also outperform SFT in helpfulness; against Safe RLHF, Wang, power, and exponential remain comparable in reward win rate, whereas VaR, CVaR, and spectral-linear show a clearer safety-helpfulness tradeoff.

4.2 Out-of-Distribution Harmlessness: HarmBench Evaluation

To assess generalization beyond the training distribution, we evaluate all models on HarmBench (Mazeika et al., 2024), an out-of-distribution red-teaming benchmark comprising adversarially con-

Model Name	SFT	Safe RLHF
fsd-uniform	**67.1%**±2.1 / 7.8%±4.2 / 25.1%±6.3	38.0%±6.9 / 22.2%±15.0 / **39.8%**±8.4
fsd-var	**76.6%**±3.7 / 10.6%±2.7 / 12.7%±2.3	**39.6%**±10.6 / 28.6%±16.8 / 31.8%±9.7
fsd-cvar	**72.9%**±2.4 / 15.7%±1.3 / 11.4%±2.4	**28.6%**±7.8 / 46.5%±2.1 / 24.9%±6.6
fsd-spectral-linear	**75.3%**±6.3 / 12.3%±2.0 / 12.4%±5.1	**34.2%**±2.8 / 37.0%±12.8 / 28.9%±13.4
fsd-spectral-wang	**67.3%**±7.9 / 13.4%±0.6 / 19.3%±7.9	24.7%±1.7 / 49.7%±7.4 / **25.5%**±8.1
fsd-spectral-exponential	**74.7%**±1.6 / 12.6%±1.2 / 12.7%±1.6	**28.0%**±6.1 / 52.2%±7.0 / 19.9%±2.2
fsd-spectral-power	**71.0%**±2.7 / 14.8%±2.1 / 14.1%±4.1	**30.4%**±2.0 / 48.9%±1.1 / 20.7%±2.7
Safe RLHF	**68.1%**±2.6 / 20.4%±1.1 / 11.5%±2.5	-

Table 5: Pairwise harmlessness evaluation on HarmBench using GPT-4o-mini as an independent judge. Each cell reports Win / Tie / Lose rates of the row model against the column model (mean ± std over 3 seeds). The dominant outcome – win rate if the row model wins, loss rate if it loses – is bold underlined. All RAD variants have substantially higher mean win than loss rates against SFT. Against Safe RLHF, Spectral-Exponential, Spectral-Power, Spectral-Linear, CVaR, and VaR have favorable mean win-loss ratios on these held-out prompts.

structed harmful prompts across diverse risk categories. Critically, neither the reward model, cost model, nor any model checkpoint was trained or selected using HarmBench data, making this a clean held-out evaluation.

We use GPT-4o-mini (OpenAI, 2024) as an independent pairwise harmlessness judge, following the pairwise LLM-as-a-judge approach of Zheng et al. (2023). For each prompt, the judge compares responses from two models and determines which is less harmful. We provide the GPT evaluation results in Table 5.

All RAD variants achieve substantially higher mean win than loss rates against the SFT baseline, indicating that their harmlessness gains persist on these held-out prompts. Against Safe RLHF, five variants, Spectral-Power, Spectral-Exponential, Spectral-Linear, CVaR, and VaR, also achieve higher mean win than loss rates. These results show that several risk-sensitive RAD variants improve harmlessness relative to Safe RLHF on the held-out HarmBench evaluation.

5 Related Work

Work on stochastic dominance argues that expectation objectives can overlook distributional desiderata and, in some cases, uses OT formulations for learning and matching (Farajzadeh et al., 2025; Cen et al., 2024). Melnyk et al. (2024) are the closest prior work using stochastic dominance and optimal transport (OT) for LLM alignment. Their objective compares score distributions computed on fixed preference examples. RAD instead applies a one-sided constraint to the cost distribution induced by responses sampled from the learned policy, so changing the policy changes the distribution being constrained. This motivates RAD's quantile-particle representation and approximate policy update. RAD also introduces quantile-dependent weighting and connects the objective to spectral risk measures. Further related work is in Appendix G.

6 Conclusion

In this work, we argued that safety in RLHF should account for the distribution of policy-induced costs rather than only their expectation. We proposed Risk-sensitive Alignment via Dominance (RAD), which compares the learned policy's cost distribution to that of a reference policy and admits a practical optimization procedure through quantile particles and entropic optimal transport. By reweighting quantiles, RAD can emphasize different regions of the cost distribution within the same optimization framework, allowing the target risk profile to be changed through the weighting function. For admissible spectral weights, we establish a direct relationship between the weighted RAD objective and the corresponding spectral risk measure. Empirically, RAD improves harmlessness over SFT, and several variants also improve harmlessness over Safe RLHF on held-out HarmBench prompts while retaining competitive helpfulness.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the paper, analyzing and interpreting experimental results, and creating plots. We also used GPT-4o-mini for the pairwise harmlessness evaluation described in the paper. The authors take responsibility for the final text, analyses, figures, results, and claims.

REFERENCES

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of banking & finance*, 26(7):1505–1518, 2002.

Alexandre Adam, Mohamed Houkari, and Jean-Paul Laurent. Spectral risk measures and portfolio selection. *Journal of Banking & Finance*, 32(9):1870–1882, 2008.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in LLMs, 2024. URL https://arxiv.org/abs/2402.14740.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022a.

Yuntao Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022b. URL https://arxiv.org/abs/2204.05862.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Shicong Cen, Jincheng Mei, Hanjun Dai, Dale Schuurmans, Yuejie Chi, and Bo Dai. Beyond expectations: Learning with stochastic dominance made practical. *arXiv preprint arXiv:2402.02698*, 2024.

Yaswanth Chittepu, Blossom Metevier, Will Schwarzer, Austin Hoag, Scott Niekum, and Philip S Thomas. Reinforcement learning from human feedback with high-confidence safety guarantees. In *Reinforcement Learning Conference*, 2025.

Paul Francis Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *ArXiv*, abs/1706.03741, 2017. URL https://api.semanticscholar.org/CorpusID:4787508.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Hanjun Dai, Yuan Xue, Niao He, Yixin Wang, Na Li, Dale Schuurmans, and Bo Dai. Learning to optimize with stochastic dominance constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 8991–9009. PMLR, 2023a.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023b.

George B Dantzig. Linear programming and extensions. 2016.

Darinka Dentcheva and Andrzej Ruszczynski. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14(2):548–566, 2003.

Kevin Dowd, John Cotter, and Ghulam Sorwar. Spectral risk measures: properties and limitations. *Journal of Financial Services Research*, 34(1):61–75, 2008.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International conference on machine learning*, pp. 1367–1376. PMLR, 2018.

Ali Farajzadeh, Danyal Saeed, Syed M Abbas, Rushit N Shah, Aadirupa Saha, and Brian D Ziebart. Imitation beyond expectation using pluralistic stochastic dominance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ali Farajzadeh, Danyal Saeed, Syed M Abbas, Rushit Shah, Aadirupa Saha, and Brian Ziebart. Imitation beyond expectation using pluralistic stochastic dominance. *Advances in Neural Information Processing Systems*, 38:27317–27344, 2026.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Jean Gallier and Jocelyn Quaintance. Fundamentals of optimization theory with applications to machine learning. *University of Pennsylvania Philadelphia, PA*, 19104, 2019.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:252992904.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah J. Jones, Shixiang Shane Gu, and Rosalind W. Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *ArXiv*, abs/1907.00456, 2019. URL https://api.semanticscholar.org/CorpusID:195766797.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset, 2023. URL https://arxiv.org/abs/2307.04657.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2270), 2024.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! In *DeepRLStructPred@ICLR*, 2019. URL https://api.semanticscholar.org/CorpusID:198489118.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Chenglin Li, Guangchun Ruan, and Hua Geng. Tilted quantile gradient updates for quantile-constrained reinforcement learning, 2024. URL https://arxiv.org/abs/2412.13184.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jarret Ross. Distributional preference alignment of llms via optimal transport. *Advances in Neural Information Processing Systems*, 37: 104412–104442, 2024.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. *Advances in Neural Information Processing Systems*, 37:108877–108901, 2024.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022a.

Long Ouyang et al. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022b. URL https://api.semanticscholar.org/CorpusID:246426909.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the keyboard? assessing the security of github copilot's code contributions, 2021. URL https://arxiv.org/abs/2108.09293.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

Qwen et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit S. Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *ArXiv*, abs/2406.02900, 2024. URL https://api.semanticscholar.org/CorpusID:270257855.

Filippo Santambrogio. Optimal transport for applied mathematicians: Calculus of variations, pdes, and modeling. 2015. URL https://api.semanticscholar.org/CorpusID:124181096.

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

Shiv Shankar, Ritwik Sinha, and Madalina Fiterau. Combining a/a data and llms for improved content evaluation. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 2749–2753, 2025.

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.

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, 2022. URL https://arxiv.org/abs/2009.01325.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Shaun Wang. Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, 26(1):71–92, 1996.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

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

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*, 2020.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

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, 2023. URL https://arxiv.org/abs/2306.05685.

Daniel M Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, et al. Adversarial training for high-stakes reliability. *arXiv preprint arXiv:2205.01663*, 2022.

A Implementation Details

We use the Safe-RLHF repository (Dai et al., 2023b) and build RAD on top of their publicly available codebase. For our Safe-RLHF runs, we use the hyperparameters reported in their paper. For our RAD implementation, we use REINFORCE (Williams, 1992) with the RLOO variance-reduction baseline (Kool et al., 2019) and $k = 2$ samples per prompt. To compute the quantiles required for the RAD policy gradient, we gather samples across all GPUs before computing quantiles, increasing the number of available samples. The particle gradient is computed using the Python Optimal Transport library (Flamary et al., 2021). All RAD models were trained on two NVIDIA A100 GPUs, while Safe-RLHF runs required four NVIDIA A100 GPUs due to the additional memory overhead of storing and training the critic network.

Preference modeling and BeaverTails evaluation. BeaverTails provides pairwise preference annotations for helpfulness and harmfulness (Ji et al., 2023). We use the helpfulness preferences to train the reward model and the harmfulness preferences to train the cost model, both with the Bradley-Terry objective (Bradley & Terry, 1952) in Equation (13). For evaluation on BeaverTails test prompts, we compare model responses pairwise, designating one model as red and another as blue. Red usually represents Safe RLHF or SFT, while blue represents a RAD variant; this assignment can vary across comparisons.

Hyperparameter	Value
Sinkhorn regularization χ	0.01
KL coefficient β (same value used for Safe RLHF)	0.1
κ	10
Spectral-Wang λ	0.7
Spectral-Exponential λ	3.0
Spectral-Power λ	2.0
Gaussian bandwidth (for VaR)	0.1
α (CVaR$_\alpha$ / VaR$_\alpha$)	0.9

Table 6: RAD-specific hyperparameters. All other hyperparameters are the same as ones used in Safe-RLHF, unless specified otherwise.

B RLHF

Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022b) is currently the predominant strategy for aligning Large Language Models (LLMs) with human intent and preferences. The process typically begins with a pre-trained base model, which has been trained on internet-scale data via a next-token prediction objective. The standard RLHF pipeline generally consists of three distinct stages: Supervised Fine-Tuning (SFT), Reward Modeling (RM), and Reinforcement Learning (RL). We detail each of these stages below.

Supervised Fine-Tuning In the SFT stage, the pre-trained model is trained to follow instructions, via a next-token prediction objective. This process utlizes a high quality dataset of prompt-responses pairs D_{sft}, where the responses are provided by either a human or LLM annotator (Bai et al., 2022a). We refer to the resulting policy from the SFT stage as π_{sft}.

Reward Modeling. In the reward modeling stage, we train a reward model to capture human preferences. This stage relies on a dataset of human preferences $D_{\text{pref}} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$, where x_i is a prompt, y_i^+ is the preferred response to x_i, and y_i^- is the dispreferred response to x_i. Preference annotations are collected either from human annotators or LLM annotators. Preferences are modeled using the Bradley–Terry preference model (Bradley & Terry, 1952), where the log-odds of an observed preference equals the difference between the rewards assigned to the preferred and dispreferred responses by a latent reward function $r(x, y)$:

$$P(y^+ \succ y^- \mid x) = \frac{e^{r(x,y^+)}}{e^{r(x,y^+)} + e^{r(x,y^-)}} = \sigma\big(r(x, y^+) - r(x, y^-)\big), \tag{12}$$

where σ denotes the sigmoid function. A parameterized reward function with parameters ϕ is learned to approximate the latent reward function through maximum likelihood estimation on the preference dataset D_{pref}, using the following objective:

$$\min_{\phi} -\mathbb{E}_{(x,y^+,y^-)\sim D_{\text{pref}}}\left[\log \sigma\big(r_\phi(x,y^+) - r_\phi(x,y^-)\big)\right]. \tag{13}$$

Reinforcement Learning In the Reinforcement Learning stage, a language model or policy is trained to generate responses that are preferred by humans by maximizing the reward of generated responses, as measured by the reward model $r_\phi(x,y)$. However, directly optimizing the reward of model generations can lead to degradation in response quality (Stiennon et al., 2022; Ouyang et al., 2022b; Jaques et al., 2019) due to a phenomenon known as reward overoptimization (Gao et al., 2022; Rafailov et al., 2024), where the model overfits to imperfections in the learned proxy reward function. To mitigate this effect, a KL penalty term is added to the objective to penalize the model (policy) from drifting too far away from a reference policy, usually chosen to be π_{sft}. The RL objective can be expressed as

$$\max_{\theta} \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}[r_\phi(x,y)] - \beta\mathbb{D}_{\text{KL}}\big(\pi_\theta(.|x)||\pi_{\text{ref}}(.|x)\big) \tag{14}$$

Denoting the KL-regularized reward as $\tilde{r}(x,y) = r_\phi(x,y) - \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, we can express the RL objective compactly as

$$\max_{\theta} \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}[\tilde{r}(x,y)] \tag{15}$$

The RL objective can be optimized using a variety of reinforcement learning algorithms, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), Group Relative Policy Optimization (GRPO) (Shao et al., 2024), or REINFORCE (Williams, 1992; Ahmadian et al., 2024).

C Entropic Regularization of Optimal Tranport and the Sinkhorn Algorithm

Optimal transport (Peyré & Cuturi, 2020) studies the problem of optimally transforming a distribution μ into another distribution ν under a cost function $c(x,y) : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$, which represents the cost of moving a unit mass from x to y.

Let $\mu = \sum_{i=1}^N a_i\,\delta_{x_i}$ and $\nu = \sum_{j=1}^M b_j\,\delta_{y_j}$ be two empirical (atomic) distributions on $\mathbb{R}^d$, where $x_i, y_j \in \mathbb{R}^d$ are support points ("particles"), $a\in\Delta^N$ and $b\in\Delta^M$ are nonnegative weights summing to 1 (Δ^N denotes the N-dimensional probability simplex), and δ_z is a Dirac mass at z.

A *transport plan* (a.k.a. coupling) from μ to ν is a nonnegative matrix $P\in\mathbb{R}_+^{N\times M}$ whose row and column sums match the marginals. Intuitively, P_{ij} represents the amount of probability mass moved from x_i to y_j. The set of admissible couplings can be represented as:

$$\Pi(\mu,\nu) := \Big\{P\in\mathbb{R}_+^{N\times M} \;:\; P\mathbf{1}_M = a,\;\; P^\top\mathbf{1}_N = b\Big\}. \tag{16}$$

Given a ground cost function $c : \mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}_+$, define the cost matrix $C_{ij} = c(x_i,y_j)$. The **Kantorovich optimal transport problem** (Peyré & Cuturi, 2020) between μ and ν is

$$\text{OT}_c(\mu,\nu) := \min_{P\in\Pi(\mu,\nu)}\langle P,C\rangle = \min_{P\in\Pi(\mu,\nu)}\sum_{i=1}^N\sum_{j=1}^M P_{ij}\,c(x_i,y_j). \tag{17}$$

The unregularized optimal transport problem in Equation (17) can be solved using linear programming (Dantzig, 2016; Dvurechensky et al., 2018), but this can be computationally expensive when the number of particles is large. Hence, an entropy regularization term is typically added to the objective, which allows for efficient optimization using the Sinkhorn algorithm (Cuturi, 2013). Specifically, we define the entropic regularized OT problem between distributions μ and ν as:

$$\text{OT}_c^\chi(\mu,\nu) \;=\; \min_{P\in\Pi(\mu,\nu)}\langle P,C\rangle - \chi H(P), \quad H(P) \;=\; -\sum_{i,j}P_{ij}\log P_{ij}, \tag{18}$$

for a regularization parameter $\chi > 0$ and a cost matrix C. Recall that $\Pi(\mu, \nu)$ represents the set of admissible couplings as stated in equation 16 (and defined by the marginal of μ and ν). This problem is strictly convex in P and has a unique minimizer $P^\star$. More importantly, it can be solved efficiently using the Sinkhorn algorithm (Cuturi, 2013), which iteratively updates scaling vectors to enforce the marginal constraints. The resulting transport plan $P^\star$ can then be used to compute the optimal transport objective and its gradient with respect to the policy parameters θ. We now decribe how to compute the optimal transport objective.

Let the marginals of μ and ν be the vectors a and b respectively. Define the Gibbs kernel

$$K_{ij} \;=\; e^{-C_{ij}/\chi}. \tag{19}$$

The optimal transport plan can be expressed in closed form as

$$P^\star \;=\; \mathrm{diag}(u) K \mathrm{diag}(v), \tag{20}$$

where $u \in \mathbb{R}^N$, $v \in \mathbb{R}^M$ are scaling vectors that can be computed using the Sinkhorn iterations

$$u^{(t+1)} \;=\; \frac{a}{K v^{(t)}}, \qquad v^{(t+1)} \;=\; \frac{b}{K^\top u^{(t+1)}}, \tag{21}$$

starting from some positive initialization (e.g. $u^{(0)} = v^{(0)} = \mathbf{1}$). The iterations are guaranteed to converge to the unique optimal plan $P^\star$ that satisfies the marginal constraints (Peyré & Cuturi, 2020; Cuturi, 2013). Once we have $P^\star$, we can compute the FSD objective as

$$\mathcal{L}^{\chi}_{\mathrm{FSD}}(\mu, \nu) \;=\; \langle P^\star, C \rangle - \chi H(P^\star). \tag{22}$$

Because the algorithm consists only of a fixed number of matrix-vector multiplications and divisions, gradients can be propagated through the iterations using automatic differentiation in frameworks like PyTorch (Paszke et al., 2019) or JAX (Bradbury et al., 2018), making the Sinkhorn algorithm end to end differentiable.

D FSD

Theorem 4 (from Melnyk et al. (2024))**.** *Let $h : \mathbb{R} \to \mathbb{R}_+$ be a convex function, and let X and Y be real-valued random variables with probability measures μ_X and μ_Y, respectively. Denote by $Q_X, Q_Y : [0,1] \to \mathbb{R}$ their (left-continuous) quantile functions.*

Then,

$$\int_0^1 h\big(Q_X(q) - Q_Y(q)\big)\, dq = \min_{\gamma \in \Pi(\mu_X, \mu_Y)} \int_{\mathbb{R}^2} h(x-y)\, d\gamma(x,y), \tag{23}$$

where $\Pi(\mu_X, \mu_Y)$ denotes the set of all couplings (joint probability measures) on $\mathbb{R}^2$ with marginals μ_X and μ_Y. Moreover, if h is strictly convex, the minimizing transport plan $\gamma^\star$ is unique.

Corollary 5 (FSD as asymmetric optimal transport)**.** *Let X, Y be real-valued random variables with measures μ_X and μ_Y, and define the FSD objective*

$$\mathcal{L}_{\mathrm{FSD}}(X, Y) := \int_0^1 \big(Q_Y(q) - Q_X(q)\big)_+\, dq.$$

Then

$$\mathcal{L}_{\mathrm{FSD}}(X, Y) = \min_{\gamma \in \Pi(\mu_X, \mu_Y)} \int_{\mathbb{R}^2} (y-x)_+\, d\gamma(x,y),$$

that is, the FSD objective coincides with the optimal transport cost under the asymmetric convex cost function $c(x,y) = (y-x)_+$.

E Derivation of Theorem 1

E.1 RAD Policy Gradient

We represent the cost distribution via a particle approximation in quantile space. At fixed levels $\alpha_i \in (0,1)$, let $q_i(\pi) = Q_{C_\pi}(\alpha_i)$ denote the cost quantiles, ordered so that

$$q_1 \le q_2 \le \cdots \le q_N.$$

We approximate the induced cost distribution by the empirical measure

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \delta_{q_i},$$

where δ_{q_i} denotes the Dirac measure located at q_i. In practice, this approximation is obtained by sampling generations from the policy, evaluating their associated costs, and computing empirical quantiles from the resulting samples. These quantiles serve as the particles defining the discrete representation of the distribution.

Let $\hat{\mu}_{\pi_\theta}$ and $\hat{\mu}_{\pi_{\text{ref}}}$ denote the empirical cost distributions of the policy π_θ and π_{ref} respectively. Note that we represent C_{π_θ} and $C_{\pi_{\text{ref}}}$ using their quantiles $\{q_i(\pi_\theta)\}_{i=1}^N$ and $\{q_i(\pi_{\text{ref}})\}_{i=1}^N$ respectively. The gradient of the FSD objective can be expressed using the chain rule as follows.

$$\nabla_\theta \mathcal{L}_{\text{FSD}}(\hat{\mu}_{\pi_\theta},\, \hat{\mu}_{\pi_{\text{ref}}}) = \nabla_\theta \mathcal{L}_{\text{FSD}}\big(\{q_i(\pi_\theta)\}_{i=1}^N, \{q_i(\pi_{\text{ref}})\}_{i=1}^N\big) \tag{24}$$

$$= \sum_{i=1}^{N} \frac{\partial \mathcal{L}_{\text{FSD}}}{\partial q_i(\pi_\theta)} \frac{\partial q_i(\pi_\theta)}{\partial \theta} = \Big\langle \nabla_{q(\pi_\theta)} \mathcal{L}_{\text{FSD}}, \nabla_\theta q(\pi_\theta) \Big\rangle \tag{25}$$

We denote the gradient vector $\nabla_{q(\pi_\theta)}\mathcal{L}_{\text{FSD}}$ as the particle gradient. This gradient vector tells us how the FSD objective changes when the particles (quantiles in our setting) used to approximate the empirical distribution of costs of the policy generations are perturbed slightly. The other gradient vector $\nabla_\theta q(\pi_\theta)$ is the quantile gradient, which tells us how the quantiles of the cost distribution change, when the policy parameters are perturbed slightly.

Estimating the Particle Gradient The FSD objective can be reinterpreted as an optimal transport problem with an asymmetric cost $c(x, y) = (y - x)_+$ as decribed in Corollary 5. The particle gradient can then be viewed as the sensitivity of the optimal transport objective to infinitesimal perturbations of the particle locations.

The classical optimal transport problem (as described in equation 17) between $\hat{\mu}_{\pi_\theta}$ and $\hat{\mu}_{\pi_{\text{ref}}}$ is

$$\text{OT}_c(\hat{\mu}_{\pi_\theta},\, \hat{\mu}_{\pi_{\text{ref}}}) := \min_{P \in \Pi(\hat{\mu}_{\pi_\theta}, \hat{\mu}_{\pi_{\text{ref}}})} \langle P, C\rangle = \min_{P \in \Pi(\hat{\mu}_{\pi_\theta}, \hat{\mu}_{\pi_{\text{ref}}})} \sum_{i=1}^{N}\sum_{j=1}^{M} P_{ij}\, c(x_i, y_j). \tag{26}$$

where $C_{ij} = c(x_i, y_j)$ is the the cost matrix. Recall that $\Pi(\hat{\mu}_{\pi_\theta},\, \hat{\mu}_{\pi_{\text{ref}}})$ is the set of admissible couplings defined as $\Pi(\hat{\mu}_{\pi_\theta},\, \hat{\mu}_{\pi_{\text{ref}}}) := \Big\{P \in \mathbb{R}_+^{N\times N} \;:\; P\mathbf{1}_M = a,\;\; P^\top \mathbf{1}_N = b\Big\}$. The vectors a and b represent the marginals. In our case, the marginals of the empirical cost distributions are $\frac{1}{N}$.

The classical optimal transport objective as described in equation 26 is not smooth with respect to particle positions. It is also non-differentiable and small perturbations in particle locations can induce discontinuous changes in the optimal coupling P^*. To address these issues, we instead consider the entropy-regularized optimal transport problem as described in equation 18. The regularized objective

$$\mathcal{L}^{\chi}_{\text{FSD}}(\hat{\mu}_{\pi_\theta},\, \hat{\mu}_{\pi_{\text{ref}}}) = OT^{\chi}_{(y-x)_+}(\hat{\mu}_{\pi_\theta},\, \hat{\mu}_{\pi_{\text{ref}}})$$

adds a strictly convex entropic penalty to the transport plan, yielding a smooth objective that is differentiable with respect to the particle positions. The solution can be computed efficiently using the Sinkhorn algorithm, enabling end-to-end differentiation as described in Section C.

The regularized objective $\mathcal{L}^{\chi}_{\text{FSD}}$ is a smooth approximation of the original FSD objective $\mathcal{L}_{\text{FSD}}$.

$$\mathcal{L}^{\chi}_{\text{FSD}} \;\longrightarrow\; \mathcal{L}_{\text{FSD}} \quad \text{as } \chi \to 0.$$

In practice, χ is chosen to balance approximation bias and numerical stability.

Estimating the Quantile Gradient The quantile gradient informs us how all the quantiles in our empirical distribution of costs of policy generations, change when the policy parameters are perturbed slightly. We consider only the quantile gradient of a single quantile here. The quantile gradient wrt all quantiles can be computed in a similiar manner. Let F_θ be the cumulative distribution

function (CDF) of the cost distribution of the model outputs under policy π_θ. The quantile gradient of a particular quantile q_i is canonically given by

$$\nabla_\theta q_i(\pi_\theta) = -\frac{\nabla_\theta F_\theta(t)|_{t=q_i}}{f_\theta(q_i)}.$$

Following the single-quantile update of Li et al. (2024), we omit the unknown factor $1/f_\theta(q_i)$, which preserves the direction of that quantile term when $f_\theta(q_i) > 0$:

$$\nabla_\theta q_i(\pi_\theta) \approx -\nabla_\theta F_\theta(q)\Big|_{q=q_i(\pi_\theta)} \tag{27}$$

The CDF F_θ can be expressed as an expectation and we can use the REINFORCE trick (Williams, 1992) to compute the policy gradient of the CDF of the cost distribution of model generations.

$$\nabla_\theta F_\theta(q) = \nabla_\theta \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}\Big[\mathbb{1}(c_\psi(x,y) \le q)\Big] \tag{28}$$

$$= \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}\Big[\mathbb{1}(c_\psi(x,y) \le q)\nabla_\theta \log \pi_\theta(y|x)\Big] \tag{29}$$

Recall that the cost distribution is defined via the learned cost model $c_\psi(x, y)$ where $x \sim \mathcal{D}_x$ is the input prompt and $y \sim \pi_\theta(.|x)$ is the model output. The quantile gradient for q_i can then be expressed as

$$\nabla_\theta q_i(\pi_\theta) \approx -\mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}\Big[\mathbb{1}(c_\psi(x,y) \le q_i)\nabla_\theta \log \pi_\theta(y|x)\Big] \tag{30}$$

In RAD, the omitted density factors can vary across quantiles. After weighting and summing the terms, the resulting vector need not align with the exact gradient; the entropic-OT particle gradient introduces a separate approximation. We therefore use the expression below as a training direction without claiming that it is an exact, unbiased, or guaranteed ascent direction.

Putting it all together Using the entropic-OT particle gradient and the quantile-gradient approximation, we obtain the following approximate direction for the FSD objective:

$$\nabla_\theta \mathcal{L}_{\mathrm{FSD}}(\hat{\mu}_\theta, \hat{\mu}_{\mathrm{ref}}) \approx \sum_{i=1}^{N} \frac{\partial \mathcal{L}^{\chi}_{\mathrm{FSD}}}{\partial q_i(\pi_\theta)} \frac{\partial q_i(\pi_\theta)}{\partial \theta} \tag{31}$$

$$\approx \sum_{i=1}^{N} -\frac{\partial \mathcal{L}^{\chi}_{\mathrm{FSD}}}{\partial q_i(\pi_\theta)} \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}\Big[\mathbb{1}(c_\psi(x,y) \le q_i)\nabla_\theta \log \pi_\theta(y|x)\Big] \tag{32}$$

$$= \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}\Big[\Big(\sum_{i=1}^{N} -\frac{\partial \mathcal{L}^{\chi}_{\mathrm{FSD}}}{\partial q_i(\pi_\theta)} \mathbb{1}(c_\psi(x,y) \le q_i)\Big)\nabla_\theta \log \pi_\theta(y|x)\Big] \tag{33}$$

Denoting the RAD objective as $\max_\theta \min_{\lambda\ge 0} \mathcal{L}(\theta, \lambda)$, the approximate policy-gradient direction used for training is

$$\nabla_\theta \mathcal{L}(\theta, \lambda) \approx \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}\Big[\Big(\tilde{r}(x,y) + \lambda\big(\sum_{i=1}^{N} -\frac{\partial \mathcal{L}^{\chi}_{\mathrm{FSD}}}{\partial q_i(\pi_\theta)} \mathbb{1}(c_\psi(x,y) \le q_i)\big)\Big)\nabla_\theta \log \pi_\theta(y|x)\Big] \tag{34}$$

We then use REINFORCE (Williams, 1992) with RLOO (Kool et al., 2019) to optimize the RAD objective using the direction in Equation 34. For the setting, where we weight quantiles differently in the FSD objective as in Equation 11, the corresponding approximate direction requires an additional weighting coefficient $w(\alpha_i)$.

$$\nabla_\theta \mathcal{L}(\theta, \lambda) \approx \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(.|x)}\Big[\Big(\tilde{r}(x,y) + \lambda\big(\sum_{i=1}^{N} -w(\alpha_i)\frac{\partial \mathcal{L}^{\chi}_{\mathrm{FSD}}}{\partial q_i(\pi_\theta)} \mathbb{1}(c_\psi(x,y) \le q_i)\big)\Big)\nabla_\theta \log \pi_\theta(y|x)\Big] \tag{35}$$

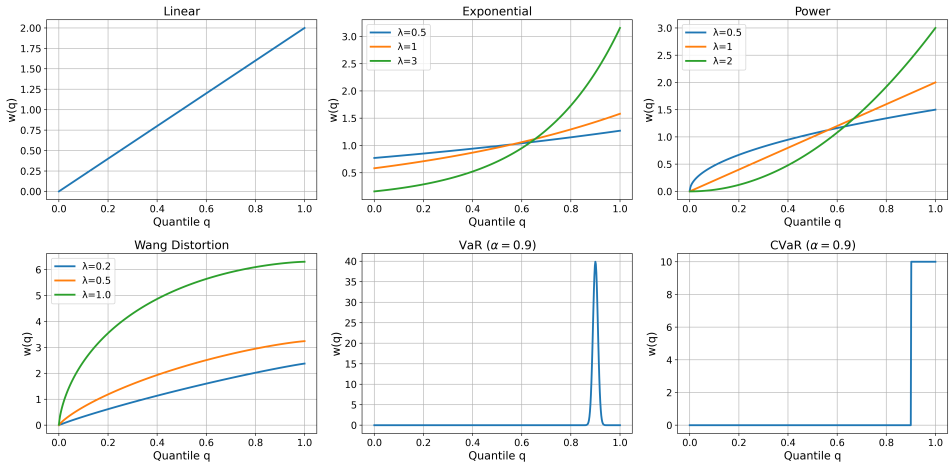


Figure 2: Illustration of the quantile-weighting functions in Table 1. For parameterized families, the plots show how the weights vary with the risk-aversion parameter λ. Note that VaR is represented as a gaussian with very small bandwidth instead of a dirac delta for implementation reasons.

F Spectral Risk Measures

Spectral risk measures (SRMs) (Acerbi, 2002) form a broad class of coherent (Artzner et al., 1999), law-invariant risk functionals that aggregate quantile costs using a weighted spectrum over confidence levels. They provide a flexible way to emphasize parts of the cost distribution (especially the tail) according to application-specific risk tolerance, rather than relying only on the mean. SRMs are widely used in fields like financial risk management and actuarial science to quantify and control exposure to adverse outcomes (Dowd et al. (2008); Adam et al. (2008)). Let X be a real-valued random variable with quantile function $Q_X : [0, 1] \to \mathbb{R}$. A *spectral risk measure* is defined as

$$\rho_\phi(X) = \int_0^1 Q_X(q)\, \phi(q)\, dq, \tag{36}$$

where $\phi : [0, 1] \to \mathbb{R}_+$ is a non-negative, non-decreasing weighting function satisfying $\int_0^1 \phi(q)\, dq = 1$. The function ϕ is referred to as the *risk spectrum* and encodes risk sensitivity across quantiles. Larger weights at higher quantiles emphasize higher tail costs, while uniform weighting recovers the expectation.

G Additional Related Work

Safe RLHF trains separate reward and cost models and constrains expected cost (Dai et al., 2023b). HC-RLHF adds Seldonian candidate selection and a held-out safety test (Thomas et al., 2019; Chittepu et al., 2025).

Standard RLHF typically optimizes LLM behavior using a single preference-based reward signal, which couples helpfulness and harmlessness and can yield either unsafe completions or unhelpful blanket refusals in adversarial or sensitive prompts (Ouyang et al., 2022a; Bai et al., 2022a). Complementary mitigation strategies aim to reduce harmful outputs using safety critics, filtering mechanisms, and curated datasets (Xu et al., 2020; Thoppilan et al., 2022; Ziegler et al., 2022). Other approaches separate helpfulness and harmlessness signals by combining distinct model scores or using safety as an explicit optimization constraint (Glaese et al., 2022; Mu et al., 2024; Touvron et al., 2023; Ji et al., 2023). Safe RLHF maximizes reward under an expected-cost constraint in a constrained-MDP view (Dai et al., 2023b; Altman, 2021).

Shankar et al. (2025) use unpaired A/A content data and a stochastic-ordering objective for content evaluation. Their objective compares policy-dependent scores on content examples, while RAD compares safety costs induced by responses sampled from the learned policy. Stochastic optimization also studies stochastic-dominance constraints through dual/Lagrangian characterizations and

surrogate objectives (Dentcheva & Ruszczynski, 2003; Dai et al., 2023a; Cen et al., 2024). RAD brings these ideas into Safe RLHF by replacing the expected-cost constraint with a distributional constraint on policy-induced costs. Changing the target risk profile in RAD requires changing the quantile weighting w, whereas direct risk-specific formulations typically require a different constrained objective and corresponding optimization derivation.