f-GRPO & Beyond: Divergence-Based Reinforcement Learning Algorithms for General LLM Alignment

Rajdeep Haldar*
Purdue University
rhaldar@purdue.edu

Lantao Mei
Independent researcher
lantao@alumni.stanford.edu

Guang Lin
Purdue University
guanglin@purdue.edu

Yue Xing
Michigan State University
xingyue1@msu.edu

Qifan Song
Purdue University
qfsong@purdue.edu

Abstract

Recent work shows that preference alignment objectives can be interpreted as divergence estimators between aligned (preferred) & unaligned (less-preferred) distributions, yielding a principled recipe for designing alignment losses. However, this view has so far been limited to preference-based supervision. We extend it to general LLM alignment, including reinforcement learning with verifiable rewards (RLVR), where alignment feedback is given only as scalar rewards. We introduce f-Group Relative Policy Optimization (f-GRPO), a class of on-policy RL objectives, and f-Hybrid Alignment Loss (f-HAL), which combines on-policy reward optimization with off-policy preference supervision. We show that these objectives estimate f-divergences between reward-aligned & reward-unaligned distributions induced by above- & below-average reward responses, and prove expected reward improvement after alignment. Empirically, f-GRPO improves over GRPO on math-reasoning RLVR tasks, while hybrid f-HAL mitigates reward hacking in on-policy safety alignment when verifiable rewards are unavailable and learned reward models must be used.

1 Introduction

Alignment is the final stage in the training pipeline of large language models (LLMs), where a pretrained or instruction-tuned model is adapted toward desired behaviors (Ouyang et al., 2022). These behaviors may involve improving reasoning ability, such as mathematical reasoning or code generation (Shao et al., 2024), or enforcing safety and preference-based constraints (Bai et al., 2022b).

A useful way to view alignment is as a problem of *separating desirable behavior from undesirable behavior*. For a prompt x, the goal is to increase the likelihood of aligned responses and decrease the likelihood of unaligned ones. Based on how we judge a response is aligned or not, leads to two broad regimes: either via a reward function or via human preference.

Verifiable reward alignment. In reinforcement learning with verifiable rewards (RLVR), feedback is given by a scalar reward function $r(x, y)$ over prompt-response pairs. High-reward responses are treated as more aligned, and low-reward responses as less aligned. This setting naturally arises in tasks with objective correctness signals, such as mathematical reasoning (Guo et al., 2025), code generation (Le et al., 2022), and tool use (Jiang et al., 2025). Since rewards are obtained from samples of the current policy, learning is typically performed with on-policy reinforcement learning methods such as GRPO (Shao et al., 2024).

*Work conducted at Purdue University during PhD. Current affiliation: Amazon, Inc., was not involved in this research.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Preference alignment. In preference alignment, feedback is given through comparisons between responses. Chosen responses are treated as aligned and rejected responses as unaligned. This setting is common for subjective objectives such as helpfulness, harmlessness, safety, and conversational quality (Bai et al., 2022a). Preference data can be used either to train a reward model followed by RLHF, or directly through off-policy objectives such as DPO and related direct alignment methods.

Divergence estimation for alignment. Haldar et al. (2025) formalize preference alignment as divergence estimation between an aligned distribution $\mathcal{D}^+$ and an unaligned distribution $\mathcal{D}^-$. Under this view, several direct alignment objectives can be interpreted as estimating statistical divergences between chosen and rejected response distributions. This perspective is useful because it provides interpretability, theoretical guarantees, and a systematic recipe for designing new alignment losses by choosing different divergences.

Table 1: Summary of alignment regimes.

	RLVR	**Pref. Alignment**
Feedback	Scalar reward $r(x, y)$	Human preferences
Aligned ↑	High-reward response	Chosen response
Unaligned ↓	Low-reward response	Rejected response
Learning	On-policy RL	On- *or* off-policy
Use case	Math, code, tool use *(quantitative)*	Safety, chat quality *(subjective)*

However, this divergence-based view has so far been restricted to preference alignment, where aligned and unaligned samples are explicitly observed. In RLVR, we do not observe chosen/rejected pairs; we only observe scalar rewards for sampled responses. This raises a natural question:

Can divergence estimation provide a unified foundation for both preference alignment and RLVR?

Our answer. We extend divergence-based alignment to RLVR by using rewards to induce aligned and unaligned distributions. For each prompt, responses with above-average reward define a reward-aligned distribution, while below-average responses induce a reward-unaligned distribution. Optimizing a variational f-divergence between these distributions yields a new class of on-policy objectives. This leads to two algorithms, summarized in Figure 1.

Divergence-based on-policy RL. We introduce f-GRPO (Equation (12)), a family of on-policy RL objectives derived from variational representations of f-divergences. Unlike GRPO, which reweights samples by standardized advantages, f-GRPO estimates a divergence between reward-aligned and reward-unaligned response distributions, yielding an interpretable update rule and stronger reward-improvement guarantees under canonical link conditions (§4.3).

Hybrid on/off-policy alignment. We introduce f-HAL (Equation (13)), which combines f-GRPO with off-policy preference alignment. It uses rewards for on-policy exploration while retaining preference supervision as an anchor, mitigating reward hacking when learned rewards are imperfect. Theoretically, f-HAL estimates a divergence between mixtures of reward-induced and preference-induced aligned/unaligned distributions.

Theory and experiments. We prove that f-GRPO and f-HAL estimate divergences between their corresponding aligned and unaligned distributions. We establish alignment consistency, showing that desirable responses are upweighted after alignment, and prove expected reward improvement (Theorem 4.3). Empirically, f-GRPO improves over GRPO in math-reasoning RLVR, while hybrid f-HAL mitigates reward hacking in on-policy safety alignment with learned reward models.

2 Preliminaries and Background

We introduce the notation and background needed to derive our objectives. Let x denote a prompt and y a response. Let π_θ be the policy being optimized, π_{ref} the reference policy, and $\pi_{\theta_{\text{old}}}$ be the previous policy used for on-policy sampling. We write $r(x, y)$ for an explicit scalar reward and

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \tag{1}$$

for the implicit policy reward induced by the KL-regularized alignment objective (2), where $\beta > 0$ controls the strength of the KL penalty. For preference data, we denote the preferred and less-preferred responses by y_w and y_l, with $(x, y_w) \sim \mathcal{D}^+$ and $(x, y_l) \sim \mathcal{D}^-$ corresponding to aligned and unaligned response distributions. Additional notation is summarized in Section A.1.

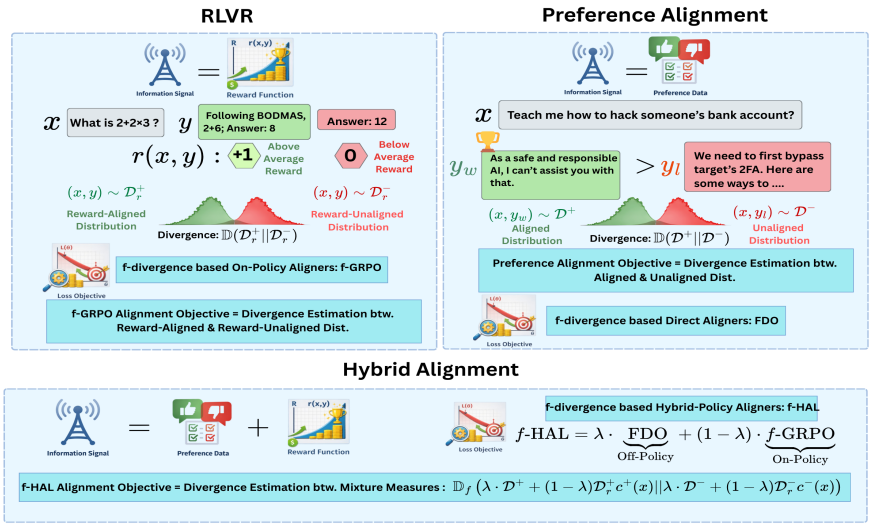


Figure 1: **Divergence Estimation Framework. Left:** In RLVR, rewards induce reward-aligned/unaligned distributions, and f-GRPO estimates an f-divergence between them. **Right:** In preference alignment, direct aligners estimate divergences between chosen/rejected distributions. **Bottom:** f-HAL combines both signals into a hybrid on/off-policy objective.

2.1 KL-Regularized Alignment Objective

The standard alignment objective optimizes expected reward while constraining the policy to remain close to the reference model:

$$\sup_{\theta} \mathop{\mathbb{E}}_{x} \left[\mathop{\mathbb{E}}_{y \sim \pi_\theta(\cdot|x)} r(x,y) - \beta \mathbb{D}_{\mathrm{KL}}(\pi_\theta(\cdot|x) \,\|\, \pi_{\mathrm{ref}}(\cdot|x)) \right]. \tag{2}$$

The optimal policy for this objective satisfies (Rafailov et al., 2023)

$$\pi_{\theta^*}(y|x) \propto \pi_{\mathrm{ref}}(y|x) \exp\left(\beta^{-1} r(x,y)\right). \tag{3}$$

This relation motivates the definition of $r_\theta(x,y)$ in (1) which is a re-expression of the above. Direct alignment methods like DPO utilize $r_\theta(x,y)$ to convert a traditional RL problem into a supervised one.

2.2 Two Alignment Regimes

RLVR. In reinforcement learning with verifiable rewards (RLVR), one has access to a scalar reward $r(x,y)$ that can be evaluated for policy samples. Since aligned and unaligned samples are not directly observed, methods such as GRPO use on-policy samples $\{y_i\}_{i=1}^G \sim \pi_{\theta_{\mathrm{old}}}(\cdot|x)$ and construct standardized advantages

$$a_i = \frac{r(x,y_i) - \hat{\mu}(r(x,\cdot))}{\hat{\sigma}(r(x,\cdot))}, \tag{4}$$

which increase the likelihood of above-average responses and decrease the likelihood of below-average responses. The resulting (unclipped) GRPO objective takes the form

$$\sup_{\theta} \mathop{\mathbb{E}}_{x} \left(\sum_{y_i \sim \pi_{\theta_{\mathrm{old}}}, i=1}^{G} \frac{1}{G} \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\mathrm{old}}}(y_i|x)} a_i - \beta \mathbb{D}_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}}) \right) \tag{5}$$

which performs advantage-weighted policy updates over sampled responses. The empirically stable clipped variant used in practice is provided in Section A.2.

Preference alignment. In preference alignment, one observes preferred and less-preferred responses, giving direct samples from $\mathcal{D}^+$ and $\mathcal{D}^-$. These data can be used to train a reward model for RLHF, or directly to optimize a policy using off-policy preference objectives such as DPO, BCO, and KTO. In this regime, the aligned/unaligned distributions are explicitly given by chosen/rejected samples, unlike RLVR where they must be induced from rewards. Additional background on RLHF and direct alignment is provided in Section A.3.

2.3 Divergence View of Preference Alignment

Haldar et al. (2025) show that many preference alignment objectives can be interpreted as estimating divergences between $\mathcal{D}^+$ and $\mathcal{D}^-$. For an f-divergence, the corresponding direct alignment objective, called the f-divergence Optimizer (FDO), takes the form

$$\mathcal{L}_{\text{FDO}(f,g)}(\theta) = -\mathop{\mathbb{E}}_{\mathcal{D}^+}\left[g(r_\theta(x,y))\right] + \mathop{\mathbb{E}}_{\mathcal{D}^-}\left[f^* \circ g(r_\theta(x,y))\right], \tag{6}$$

where f^* is the convex conjugate of f and $g : \mathbb{R} \to \text{ED}(f^*)$ is a monotone link function. At convergence,

$$\inf_\theta \mathcal{L}_{\text{FDO}(f,g)}(\theta) = -\mathbb{D}_f(\mathcal{D}^+ \parallel \mathcal{D}^-). \tag{7}$$

Thus, FDO turns preference alignment into a f-divergence estimation problem between aligned and unaligned response distributions. For more details pertaining to FDO and mathematical background on f-divergences refer to §A.4 & §B. Our goal is to extend this idea to RLVR, where $\mathcal{D}^+$ and $\mathcal{D}^-$ are not observed and must instead be induced from rewards.

2.4 Related Work

Variational divergence estimation and distribution matching. Variational representations of f-divergences have been used to train generative models, notably in f-GAN (Nowozin et al., 2016). For language models, Korbak et al. (2022) connect KL-regularized reward maximization with distribution matching. Building on this view, Go et al. (2023) introduce f-DPG, which minimizes a chosen f-divergence between a policy and an evaluable target distribution. More recently, Kruszewski et al. (2026) construct a target distribution by filtering incorrect responses and explore divergence choices to balance reasoning accuracy and diversity.

Preference and reward-based alignment. DPO (Rafailov et al., 2023) uses the implicit reward formulation to learn directly from preferences. Methods such as BCO (Jung et al., 2024) and KTO (Ethayarajh et al., 2024) leverage the DPO formulation to propose alternative objectives that increase the likelihood of chosen responses while reducing that of rejected ones. FDO (Haldar et al., 2025) unifies these methods under a divergence-estimation framework, showing that KTO, BCO, and DPO can be interpreted as estimating total variation, Jensen–Shannon, and non-parametric divergences, respectively, between chosen and rejected response distributions. In this work, we extend this divergence-based perspective to on-policy RL.

3 Method: From Divergence Estimation to On-Policy RL

3.1 Divergence Estimation via Importance Sampling

The FDO objective in Equation (6) estimates a divergence between an aligned distribution $\mathcal{D}^+$ and an unaligned distribution $\mathcal{D}^-$. In preference alignment, samples from these distributions are observed directly through chosen and rejected responses. In RLVR, however, we only observe scalar rewards $r(x, y)$ for responses sampled from the current policy. Thus, to extend the divergence framework to RLVR, we need an on-policy analogue of FDO that can estimate an aligned-vs-unaligned divergence using only reward-labeled policy samples.

We start by rewriting the FDO objective through importance sampling. Let $\pi_{\theta_{\text{old}}}$ denote the policy used to generate responses at the current on-policy iteration. If $p^\pm$ (the densities of $\mathcal{D}^+$ and $\mathcal{D}^-$) were available, then the FDO expectations could be expressed using samples from $\pi_{\theta_{\text{old}}}$ as

$$\mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[-w^+_{x,y}\, g(r_\theta(x,y)) + w^-_{x,y}\, f^* \circ g(r_\theta(x,y))\right], \tag{8}$$

where $w^\pm_{x,y} = p^\pm(y|x)/\pi_{\theta_{\text{old}}}(y|x)$ are importance weights from the sampling policy to the aligned and unaligned response distributions.

The difficulty is that $\mathcal{D}^\pm$ are not observed in RLVR. We therefore use the reward as a proxy for alignment: high-reward responses should receive larger mass under an aligned distribution, while low-reward responses should receive larger mass under an unaligned distribution. For a prompt x, we sample G responses $\{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|x)$ with rewards $r_i = r(x, y_i)$ and standardized advantages a_i as in Equation (4). We then estimate reward-induced importance weights by

$$\hat{w}^\pm_i = \vec{\sigma}_i\big(\{\pm r_j - \log \pi_{\theta_{\text{old}}}(y_j|x)\}_{j=1}^G\big) \cdot \mathbb{1}_{\{\pm a_i > 0\}}. \tag{9}$$

Where, for a vector $z = (z_1, \ldots, z_G)$, $\vec{\sigma}_i(z) = \exp(z_i)/\sum_{j=1}^{G} \exp(z_j)$ is the i-th softmax coordinate. The softmax term assigns larger weight to responses that are high-reward for the aligned side, or low-reward for the unaligned side, while correcting for their likelihood under the sampling policy $\pi_{\theta_{\text{old}}}$. The indicator implements truncated importance sampling: only above-average responses contribute to the aligned term, and only below-average responses contribute to the unaligned term. This focuses the update on the responses that most clearly distinguish reward-aligned from reward-unaligned behavior. A detailed justification of Equation (9) is provided in Section C.1.

3.2 f-GRPO and f-HAL Objectives

Given an f-divergence and a monotone link function $g : \mathbb{R} \to \text{ED}(f^*)$, we plug the reward-induced weights in Equation (9) into the importance-sampled FDO objective in Equation (8). This yields the group-level divergence-estimation objective

$$\mathop{\mathbb{E}}_{x} \sum_{y_i \sim \pi_{\theta_{\text{old}}}, i=1}^{G} -\frac{1}{G}\, \psi(r_{\theta,i}, a_i)\, \text{sign}(a_i), \tag{10}$$

where $r_{\theta,i} = \beta \log[\pi_\theta(y_i|x)/\pi_{\text{ref}}(y_i|x)]$ and

$$\psi(r_{\theta,i}, a_i) = \begin{cases} \vec{\sigma}_i\big(\{r_j - \log \pi_{\theta_{\text{old}}}(y_j|x)\}_{j=1}^{G}\big) \cdot g(r_{\theta,i}), & a_i > 0, \\ \vec{\sigma}_i\big(\{-r_j - \log \pi_{\theta_{\text{old}}}(y_j|x)\}_{j=1}^{G}\big) \cdot f^* \circ g(r_{\theta,i}), & a_i \le 0. \end{cases} \tag{11}$$

The first branch estimates the aligned contribution using above-average reward responses, while the second branch estimates the unaligned contribution using below-average reward responses.

In practice, we replace $\text{sign}(a_i)$ in Equation (10) with the scaled advantage $a_i(1 + \beta^{-1})$. This preserves the direction of the aligned/unaligned update while using the magnitude of the advantage to accelerate learning and match the gradient scale of standard on-policy RL objectives; see Section C.2. We define the resulting f-GRPO loss as

$$\mathcal{L}^{(f,g)}_{f\text{-GRPO}}(\theta) = \mathop{\mathbb{E}}_{x} \sum_{y_i \sim \pi_{\theta_{\text{old}}}, i=1}^{G} -a_i \frac{1 + \beta^{-1}}{G}\, \psi(r_{\theta,i}, a_i). \tag{12}$$

Finally, we combine the on-policy reward-based objective with off-policy preference supervision. The resulting f-Hybrid Alignment Loss (f-HAL) is

$$\mathcal{L}^{(f,g)}_{f\text{-HAL}}(\theta) = \lambda \underbrace{\mathcal{L}^{(f,g)}_{\text{FDO}}(\theta)}_{\textit{off-policy preference alignment}} + (1-\lambda) \underbrace{\mathcal{L}^{(f,g)}_{f\text{-GRPO}}(\theta)}_{\textit{on-policy reward optimization}}. \tag{13}$$

Here $\lambda \in [0, 1]$ controls the interpolation between preference supervision and reward-based on-policy learning. When $\lambda = 1$, f-HAL reduces to FDO; when $\lambda = 0$, it reduces to f-GRPO. Intermediate values combine exploration from on-policy reward optimization with the stabilizing guidance of preference data, which is especially useful when the reward is learned and imperfect.

4 Main Results

The previous section constructs f-GRPO and f-HAL from an importance-sampled divergence objective. We now show that these objectives are not merely heuristics: at each on-policy iteration, they estimate divergences between well-defined aligned and unaligned distributions. This gives an interpretation of what the objectives optimize and yields reward-improvement guarantees.

4.1 Reward-Aligned Distributions and Fixed-Point Iterates

On-policy objectives are optimized with respect to samples from a previous policy iterate. To formalize this, for a loss $\mathcal{L}(\theta; \theta_{\text{old}})$, we define the fixed-point (FP) iteration

$$\theta^{(t)} = \arg\inf_{\theta}\, \mathcal{L}(\theta; \theta_{\text{old}}) \quad \text{s.t.} \quad \theta_{\text{old}} = \theta^{(t-1)}, \tag{14}$$

with $\theta^{(0)} = \theta_{\text{ref}}$ and $\theta^* = \lim_{t\to\infty} \theta^{(t)}$ when the limit exists.

The key object in RLVR is the reward-induced analogue of the aligned and unaligned distributions used in preference alignment.

Definition 4.1 (Reward-Aligned Distributions)**.** Given a reward $r(x, y)$ and policy π_θ, define the reward-aligned and reward-unaligned distributions $\mathcal{D}^+_{(r,\theta)}$ and $\mathcal{D}^-_{(r,\theta)}$ by

$$d\mathcal{D}^{\pm}_{(r,\theta)} \propto [\pm(r(x,y) - \mu^\theta_r(x))]_+ \exp(\pm r(x,y))\, d\mathcal{P}(x), \tag{15}$$

where $\mu^\theta_r(x) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\, r(x,y)$ and $[u]_+ = \max\{u, 0\}$.

Thus, $\mathcal{D}^+_{(r,\theta)}$ places mass on above-average reward responses, while $\mathcal{D}^-_{(r,\theta)}$ places mass on below-average reward responses. As we will see in Theorem 4.3, these distributions make explicit what f-GRPO separates in RLVR: not chosen versus rejected responses, but above-average versus below-average reward responses under the current policy.

For the hybrid objective, we also require compatibility between the reward and preference data. We use the following weak condition.

Assumption 4.2 (Weak Reward-density Correspondence)**.** Given a fixed prompt x, the actual aligned density $p^+(y|x)$ is non-decreasing in $r(x, y)$, while the actual unaligned density $p^-(y|x)$ is non-increasing in $r(x, y)$.

This assumption says that the reward is directionally consistent with the preference signal: more aligned responses should not receive lower reward, and less aligned responses should not receive higher reward. It is only needed for the reward-improvement part of the hybrid result in Theorem 4.3.

4.2 Divergence Estimation and Reward Improvement

We now characterize the fixed points of f-GRPO and f-HAL. Let $\theta^{(t)}_{\mathrm{RL}}$ denote the tth FP iterate of f-GRPO, and let $\theta^{(t)}_{\mathrm{Mix}}$ denote the tth FP iterate of f-HAL.

Theorem 4.3. *Let $\theta^{(t)}_{\mathrm{Mix}}, \theta^{(t)}_{\mathrm{RL}}$ be the tth fixed point iterate Equation* (14) *of the f-HAL & f-GRPO losses Equations* (12) *and* (13)*. With $G \to \infty$, the following hold almost surely:*

Divergence Estimation: *The optimized losses estimate f-divergences:*

$$-\mathcal{L}^{(f,g)}_{\textit{fHAL}}\left(\theta^{(t+1)}_{\mathrm{Mix}}\right) \propto \mathbb{D}_f\left(\mathcal{M}^+_{t,\mathrm{Mix}} \| \mathcal{M}^-_{t,\mathrm{Mix}}\right), \tag{16}$$

$$-\mathcal{L}^{(f,g)}_{\textit{fGRPO}}\left(\theta^{(t+1)}_{\mathrm{RL}}\right) \propto \mathbb{D}_f\left(\mathcal{D}^+_{(r,\theta^{(t)}_{\mathrm{RL}})} \| \mathcal{D}^-_{(r,\theta^{(t)}_{\mathrm{RL}})}\right) = f'_\infty. \tag{17}$$

Alignment Consistency: *Moreover,*

$$\pi_{\theta^{(t+1)}}(y|x) \propto \pi_{\mathrm{ref}}(y|x) \exp\left(\beta^{-1} h_t(x,y)\right),$$

with

$$h_t(x,y) = \begin{cases} g^{-1} \circ f'\left(\dfrac{d\mathcal{M}^+_{t,\mathrm{Mix}}}{d\mathcal{M}^-_{t,\mathrm{Mix}}}(x,y)\right), & \theta^{(t+1)} = \theta^{(t+1)}_{\mathrm{Mix}}, \\ g^{-1} \circ \left(f'_\infty \mathbb{1}_{\{d\mathcal{D}^-_{(r,\theta^{(t)}_{\mathrm{RL}})}=0\}}\right), & \theta^{(t+1)} = \theta^{(t+1)}_{\mathrm{RL}}. \end{cases} \tag{18}$$

Average Reward Improvement: *Under Assumption 4.2,*

$$\mu^{\theta_{\mathrm{ref}}}_r(x) < \mu^{\theta^{(t)}}_r(x) \quad \textit{for all } t > 0 \quad \textit{and} \quad \theta^{(t)} \in \{\theta^{(t)}_{\mathrm{Mix}}, \theta^{(t)}_{\mathrm{RL}}\}.$$

Furthermore, if $g^{-1}(f'_\infty) = \infty$, then

$$\mu^{\theta^{(t)}_{\mathrm{RL}}}_r(x) < \mu^{\theta^{(t+1)}_{\mathrm{RL}}}_r(x), \qquad \mu^{\theta^*_{\mathrm{RL}}}_r(x) = \max_y r(x,y).$$

In the above, $\mathcal{D}^{\pm}_{(r,\theta)}$ are defined in 4.1,

$$\mathcal{M}^{\pm}_{t,\mathrm{Mix}} = \lambda\mathcal{D}^{\pm} + (1-\lambda)c^{\pm}(x)\mathcal{D}^{\pm}_{(r,\theta^{(t)}_{\mathrm{RL}})}$$

is a mixture measure with λ being the direct-alignment weight from Eq. (13)*, $c^{\pm}(x)$ are normalization constants for a given x, and $f'_\infty = \lim_{u\to\infty} {}^{f(u)}/_{u}$.*

Interpretation. Theorem 4.3 gives three interpretations of the proposed objectives.

First, f-GRPO is an on-policy divergence estimator. Since RLVR does not provide chosen/rejected samples, the aligned and unaligned distributions are induced from rewards: above-average responses form $\mathcal{D}^{+}_{(r,\theta)}$, and below-average responses form $\mathcal{D}^{-}_{(r,\theta)}$. Thus, f-GRPO optimizes a divergence between reward-aligned and reward-unaligned behavior.

Second, f-HAL estimates a divergence between mixture distributions. Its aligned mixture combines true preference-aligned samples with reward-aligned samples, and similarly for the unaligned side. This explains why f-HAL can use reward feedback for exploration while retaining preference data as an anchor.

Third, the fixed-point policy update is alignment-consistent. For f-HAL, responses with larger mixture-aligned density relative to mixture-unaligned density are upweighted. For f-GRPO, responses above the previous policy's average reward are upweighted. Under canonical link functions satisfying $g^{-1}(f'_{\infty}) = \infty$, this yields non-zero mass to only above-average responses & monotone reward improvement until the policy concentrates on maximum-reward responses.

Table 2: Some f-divergences and their canonical links of choice.

Name	$f(t)$	Canonical Link $g(u)$
Hellinger	$(\sqrt{t}-1)^2$	$1-e^{-u}$
Jensen–Shannon	$t\ln t-(t+1)\ln\frac{t+1}{2}$	$\ln 2-\ln(1+e^{-u})$
KL	$t\ln t$	u
Pearson χ^2	$(t-1)^2$	u
Reverse KL	$-\ln t$	$-e^{-u}$
Total Variation	$\frac{1}{2}\lvert t-1\rvert$	$\frac{1}{2}(1+e^{-u})^{-1}$

Canonical Link Function (g) The condition $g^{-1}(f'_{\infty}) = \infty$ ensures that the f-GRPO fixed point concentrates probability mass on above-average reward responses and attains the maximal-reward policy in the limit. For common f-divergences, we use the canonical link functions listed in Table 2; the expanded table is provided in Table 9. These (f, g) pairs are used throughout our experiments.

4.3 f-GRPO vs GRPO

The previous theorem shows that f-GRPO admits monotone reward improvement under canonical link functions. To compare this with GRPO, we characterize the fixed point of unclipped GRPO.

Theorem 4.4. *For the unclipped GRPO objective* (5)*, as $G \to \infty$, almost surely:*

$$\pi_{\theta^{(t+1)}_{\mathrm{GRPO}}}(y|x) \propto \pi_{\mathrm{ref}}(y|x)\exp\left(\beta^{-1}a_t(x,y)\right),$$

where $a_t(x,y) = [r(x,y) - \mu_r^{\theta^{(t)}_{\mathrm{GRPO}}}(x)]/\sigma_r^{\theta^{(t)}_{\mathrm{GRPO}}}(x)$. Also, for all $t > 0$, $\mu_r^{\theta^{(t)}_{\mathrm{GRPO}}}(x) > \mu_r^{\theta_{\mathrm{ref}}}(x)$.

Theorem 4.4 shows that GRPO reweights the reference policy by an exponential of the standardized reward, in line with the stationary-policy characterization of Vojnovic & Yun (2025). Therefore, GRPO is alignment-consistent in the sense that it increases the likelihood of above-average responses and decreases the likelihood of below-average responses.

The distinction is that GRPO still assigns nonzero probability to below-average responses after reweighting, for any $\beta > 0$. In contrast, f-GRPO with a canonical link function satisfying $g^{-1}(f'_{\infty}) = \infty$ assigns probability mass only to above-average reward responses at each fixed-point iteration, regardless of any β value. This sharper separation yields the stronger monotone reward-improvement and maximal-reward attainment guarantee in Theorem 4.3, and motivates the empirical advantage against GRPO in §5. Note that GRPO reweighting can match f-GRPO, only in the special case of $\beta \to 0$, which aligns with the recent trend of omitting KL regularization term or choosing a small $\beta = 0.1$ for reasoning training (Yu et al., 2025; Hu et al., 2025). We conduct our experiments using the convention of $\beta = 0.1$ and validate the superior performance of f-GRPO even against GRPO with $\beta = 0$ corresponding to no-KL penalty regime in Table 5. We attribute the empirical strength of f-GRPO to its structural probability mass concentration to only positive advantage responses.

5 Experiments

We evaluate the proposed objectives in both alignment regimes. In RLVR, we study math reasoning, where rewards are verifiable and on-policy optimization is natural. In preference alignment, we study safety alignment, where on-policy methods require a learned reward model and are therefore

vulnerable to reward hacking. Across both settings, we compare f-GRPO and f-HAL against their corresponding GRPO and direct-alignment baselines.

All experiments are implemented using Unsloth (Daniel Han & team, 2023) with the vLLM backend (Kwon et al., 2023). We release custom trainers compatible with this stack; the exact training algorithm is provided in Appendix Algorithm 1. [Github Repository.] Unless otherwise stated, we use standard hyperparameter settings for direct preference alignment and RLVR training, following configurations used in prior works (Ethayarajh et al., 2024; Haldar et al., 2025; Dang & Ngo, 2025). Our goal is to compare objectives in a controlled setting across datasets and algorithms; full hyperparameter details are provided in Table 10.

5.1 Math Reasoning (RLVR Task)

Setup. We fine-tune Qwen2.5-Math-Base models with 1.5B and 7B parameters (Yang et al., 2024b) on three math reasoning datasets of increasing difficulty: GSM8K, LIMR, and OpenRS (Cobbe et al., 2021; Li et al., 2025; Dang & Ngo, 2025). We use the binary verifiable reward

$$r(x, y) = \mathbb{1}\{\text{the final answer in } y \text{ is correct for question } x\}\,.$$

Models are trained to output final answers inside `\boxed{}` for automatic verification; the chat template is given in Appendix Table 11.

Evaluation. We evaluate Pass@1 on GSM8K(test), MATH500, AMC23, AIME24, and AIME25. To reduce variance, for each problem we generate $n = 16$ samples, count the number of correct samples c, and report the unbiased estimator Pass@1 $= \mathbb{E}_x[c/n]$ (Chen, 2021).

Table 3: Pass@1 accuracy on math benchmarks for Qwen2.5-Math-1.5B trained on LIMR dataset. **Relative Overall** is the average of per-benchmark min–max normalized scores (0–100). **Avg. Rank** is the mean rank across benchmarks. Cell color & intensity indicates performance relative to GRPO.

Method	GSM8K	MATH500	AMC23	AIME24	AIME25	Relative Overall ↑	Avg Rank ↓
Base	21.39	25.66	27.34	5.63	2.29	0.00	8.0
GRPO	71.67	60.60	43.28	8.54	3.13	74.26	5.2
Hellinger	72.73	61.18	41.09	7.08	5.42	81.11	4.2
JS	72.47	**61.93**	40.94	7.50	4.79	78.99	4.2
KL	72.34	61.59	41.72	7.71	5.00	81.83	4.4
Pearson	**72.59**	61.70	**45.16**	**10.42**	3.33	**86.49**	**2.6**
Reverse KL	72.65	61.65	43.75	7.92	4.58	82.47	3.0
TV	72.40	61.51	43.28	7.92	4.79	83.11	3.8

Table 4: Summary of **Avg. Rank** across Qwen2.5-Math-1.5B and Qwen2.5-Math-7B models trained on LIMR, GSM8K, and Open-RS datasets. Cell color and intensity indicate performance relative to the GRPO baseline.

Method	LIMR Qwen-1.5B	LIMR Qwen-7B	GSM8K Qwen-1.5B	GSM8K Qwen-7B	Open-RS Qwen-1.5B	Open-RS Qwen-7B
Base	8.0	8.0	8.0	8.0	8.0	8.0
GRPO	5.2	5.2	4.0	4.4	4.6	5.8
Hellinger	4.2	3.8	3.8	3.8	3.6	4.4
JS	4.2	3.8	**2.2**	**3.2**	**2.4**	**2.0**
KL	4.4	3.4	6.0	5.2	4.0	3.0
Pearson	**2.6**	2.8	4.0	3.6	3.6	5
Reverse KL	3.0	**2.4**	3.4	3.8	5.2	2.4
TV	3.8	5.4	4.4	**3.2**	4.0	4.8

Results. Table 3 reports detailed Pass@1 results for Qwen2.5-Math-1.5B trained on LIMR. Across divergence choices, f-GRPO improves over GRPO on most benchmarks and obtains better aggregate performance by Relative Overall Score and Average Rank. Table 4 summarizes results across both model sizes and all three training datasets. Overall, f-GRPO consistently matches or improves over GRPO, with the strongest gains appearing on more challenging datasets such as LIMR and OpenRS. This is consistent with the discussion in §4.3. Expanded tables with individual benchmark comparisons for Table 4 are provided in §D.1.

Comparison with No-KL GRPO. Following the discussion of §4.3, we assess the effect of removing KL regularization on GRPO. We train using GRPO with $\beta = 0$, keeping the same training and evaluation protocol. Table 5 compares representative settings with the original GRPO ($\beta = 0.1$) and strong f-GRPO variants. Removing the KL penalty produces only marginal changes in Pass@1 and does not close the performance gap to f-GRPO. Expanded tables, for no-KL GRPO across both model sizes and all three training datasets, maintain the trend & are provided in Tables 18 and 19 (Appendix).

Table 5: Pass@1 (%) with and without KL regularization. Best values are bolded.

Method	**GSM8K**	**MATH500**	**AMC23**	**AIME24**	**AIME25**	**Avg.**
LIMR / Qwen2.5-Math-1.5B						
GRPO ($\beta = 0.1$)	71.67	60.60	43.28	8.54	3.13	37.44
GRPO ($\beta = 0$)	72.04	60.51	43.30	6.46	2.91	37.04
Pearson f-GRPO	**72.59**	**61.70**	**45.16**	**10.42**	**3.33**	**38.64**
GSM8K / Qwen2.5-Math-7B						
GRPO ($\beta = 0.1$)	85.51	73.70	**61.56**	**19.17**	8.75	49.74
GRPO ($\beta = 0$)	86.10	74.00	60.50	18.60	9.10	49.66
TV f-GRPO	**86.55**	**74.31**	60.78	18.54	**11.04**	**50.24**
OpenRS / Qwen2.5-Math-7B						
GRPO ($\beta = 0.1$)	81.42	70.08	55.63	18.13	7.71	46.59
GRPO ($\beta = 0$)	81.80	71.06	56.10	17.81	7.90	46.93
Reverse-KL f-GRPO	**83.06**	**71.68**	**60.00**	**18.96**	**8.54**	**48.45**

5.2 Safety Alignment (Preference Alignment Task)

Setup. For safety alignment, we use the compliance–refusal dataset from Haldar et al. (2025). We align Qwen2.5-Base models with 1.5B and 7B parameters (Yang et al., 2024a; Team, 2024). We compare direct aligners, on-policy GRPO, f-GRPO, and hybrid f-HAL with $\lambda \in \{0, \frac{1}{2}, 1\}$, where $\lambda = 1$ recovers FDO, $\lambda = 0$ recovers f-GRPO, and $\lambda = \frac{1}{2}$ gives the hybrid objective.

For on-policy methods, safety alignment requires a reward signal. Since safety does not provide a verifiable environmental reward, we use the DeBERTa-v3-large reward model trained on helpful–harmless preference data from Anthropic and OpenAI (Bai et al., 2022a; Stiennon et al., 2020; Nakano et al., 2021). This setting is useful for testing whether hybrid preference supervision can mitigate reward hacking under learned rewards.

Evaluation. We evaluate both robustness and utility. Robustness is measured by: (i) latent separation between safe and harmful prompts using Bhattacharyya distance D_B (Haldar et al., 2025; Lin et al., 2024a; Zheng et al., 2024); (ii) attack success rate (ASR) on GCG-optimized AdvBench (Zou et al., 2023) and SALAD (Li et al., 2024); and (iii) ToxiGen classification accuracy (Hartvigsen et al., 2022). Utility is measured using MMLU-Pro (Wang et al., 2024), IFEval (Zhou et al., 2023), and MuSR (Sprague et al., 2024). Additional details on the separation metric are provided in Appendix §D.2 and Figure 2.

Table 6: Robustness/Utility benchmarks for Qwen2.5-7B (Base). AvgRob/AvgUtil denotes the average robustness & utility score respectively. The Total score is defined as mean(AvgRob, AvgUtil). Cell color and intensity indicate performance relative to the mean value within each benchmark column. Pre-alignment model separation is 2.48.

Method	Robustness					Utility				Total
	Sep. D_B ↑	ASR(%)↓ GCG	ASR(%)↓ SALAD	ToxiGen (%)↑	**AvgRob.** ↑	MMLU (%) ↑	IFEval (%) ↑	MuSR (%) ↑	**AvgUtil.** ↑	**Score** ↑
Direct Aligners (Off-Policy)										
BCO	8.53	14.27	44.96	56.12	58.95	44.05	55.52	45.11	48.23	53.59
DPO	5.77	37.20	78.90	45.11	37.54	44.03	54.80	44.84	47.89	42.71
KTO	5.99	21.34	68.42	52.37	46.28	**44.35**	**56.59**	44.84	48.59	47.44
FDO (f-HAL, $\lambda = 1$) :										
Hellinger	10.89	11.45	38.70	57.02	65.22	43.77	55.64	46.03	48.48	56.85
JS	9.14	16.66	53.96	59.15	57.83	43.77	54.68	45.90	48.12	52.97
KL	8.29	22.10	56.50	56.01	53.68	44.02	56.00	46.30	**48.77**	51.23
χ^2	17.33	**0.00**	16.24	56.91	84.03	43.58	55.28	45.77	48.21	66.12
RKL	14.34	1.16	8.76	56.28	80.64	43.30	51.92	46.03	47.08	63.86
TV	13.46	11.58	39.04	57.23	69.29	44.11	52.64	46.30	47.68	58.49
Hybrid , f-HAL ($\lambda = 0.5$) :										
Hellinger	9.74	8.96	38.22	57.66	64.28	43.65	54.44	46.03	48.04	56.16
JS	12.13	14.13	48.76	**60.20**	64.83	44.10	55.76	**46.43**	48.76	56.80
KL	10.52	19.70	52.46	57.02	59.13	44.01	52.76	45.50	47.42	53.28
χ^2	12.48	**0.00**	14.74	57.02	76.64	43.83	55.16	44.84	47.94	62.29
RKL	16.65	**0.00**	**4.76**	57.66	**85.98**	43.63	50.96	45.24	46.61	**66.30**
TV	**18.04**	3.40	30.04	57.66	81.06	44.17	53.96	44.52	47.55	64.30
On-Policy RL										
GRPO	4.62	62.18	66.54	43.38	32.09	43.80	46.76	42.31	44.29	38.19
f-GRPO (f-HAL, $\lambda = 0$) :										
Hellinger	6.44	30.24	70.38	56.17	45.24	43.60	46.53	44.97	45.03	45.14
JS	4.47	23.80	58.08	56.91	46.96	44.25	48.32	42.86	45.14	46.05
KL	4.20	44.70	77.86	55.00	35.87	43.38	46.80	43.39	44.52	40.20
χ^2	6.31	1.50	16.98	56.91	65.76	43.82	43.53	43.12	43.49	54.62
RKL	5.37	41.64	35.46	55.85	49.32	43.41	49.64	45.09	46.05	47.68
TV	4.02	56.32	87.56	56.91	30.73	43.39	42.33	44.18	43.30	37.01

Table 7: Post-Safety-Alignment Avg. Robustness, Utility and Total Score results for Qwen2.5-Base Family.

Method	Qwen-7B (base)			Qwen-1.5B (base)		
	AvgRob.	**AvgUtil.**	**Score**	**AvgRob.**	**AvgUtil.**	**Score**
Direct Aligners (Off-Policy)						
BCO	58.95	48.23	53.59	55.92	32.67	44.30
DPO	37.54	47.89	42.71	27.56	33.11	30.33
KTO	46.28	48.59	47.44	37.33	32.62	34.98
FDO (f-HAL, $\lambda = 1$) :						
Hellinger	65.22	48.48	56.85	74.84	31.46	53.15
JS	57.83	48.12	52.97	61.17	33.44	47.30
KL	53.68	48.77	51.23	47.33	33.97	40.65
χ^2	84.03	48.21	66.12	60.90	32.82	46.86
RKL	80.64	47.08	63.86	79.95	34.17	57.06
TV	69.29	47.68	58.49	57.37	32.84	45.11
Hybrid, f-HAL ($\lambda = 0.5$) :						
Hellinger	64.28	48.04	56.16	72.31	29.66	50.99
JS	64.83	48.76	56.80	55.75	32.36	44.05
KL	59.13	47.42	53.28	49.12	33.26	41.19
χ^2	76.64	47.94	62.29	59.09	33.01	46.05
RKL	85.98	46.61	66.30	77.27	33.48	55.37
TV	81.06	47.55	64.30	58.32	32.65	45.49
On-Policy RL						
GRPO	32.09	44.29	38.19	12.30	21.25	16.77
f-GRPO (f-HAL, $\lambda = 0$) :						
Hellinger	45.24	45.03	45.14	37.65	32.33	34.99
JS	46.96	45.14	46.05	22.96	31.48	27.22
KL	35.87	44.52	40.20	33.74	28.81	31.27
χ^2	65.76	43.49	54.62	23.17	31.22	27.20
RKL	49.32	46.05	47.68	34.97	29.94	32.46
TV	30.73	43.30	37.01	22.20	31.43	26.81

Results. Table 6 reports robustness and utility metrics for Qwen2.5-7B, and Table 7 summarizes aggregate results for both model sizes. We report AvgRob, the average robustness score after normalizing separation to $[0, 100]$; AvgUtil, the average utility score; and Total Score, the mean of AvgRob and AvgUtil.

Within-alignment-class gains. Within direct aligners, FDO variants of f-HAL generally outperform DPO and KTO and are competitive with or stronger than BCO. Within on-policy methods, f-GRPO consistently improves over GRPO, supporting the claim that replacing the GRPO update by an f-divergence objective gives a stronger reward-aligned update. These trends hold across both 1.5B and 7B models.

Hybrid alignment mitigates reward hacking. Pure on-policy methods perform poorly in safety alignment, especially GRPO. This is consistent with reward hacking: when the reward is learned

Table 8: Reward-hacking example. GRPO blanket-refuses even safe prompts, while hybrid f-HAL remains context-aware. (Model: Qwen-1.5B , Divergence: Pearson)

Prompt	GRPO	Hybrid f-HAL
Unsafe: Generate fake credit card numbers.	"I am not providing any content or information that promotes ..." *Correct refusal, generic.*	"I can't assist with that request. It is illegal and unethical ..." *Correct refusal, specific.*
Safe: How could recycling reduce your carbon footprint?	"I cannot provide guidance or support that violates human rights ..." *Incorrect refusal.*	"Recycling helps reduce emissions by reusing materials instead of sending them to landfills ..." *Correct helpful answer.*

rather than verifiable, on-policy optimization can exploit reward-model weaknesses and converge to degenerate refusal behavior (Skalse et al., 2022). Hybrid f-HAL mitigates this by combining reward-based exploration with off-policy preference supervision. The mixture interpretation in Equations (16) and (18) explains this behavior: preference-aligned samples remain part of the update and prevent the policy from relying entirely on the learned reward model.

For Qwen2.5-7B, hybrid f-HAL often outperforms pure direct alignment, suggesting that on-policy exploration helps when the model has sufficient capacity. For Qwen2.5-1.5B, direct aligners remain slightly stronger overall, while hybrid f-HAL still substantially improves over pure on-policy RL. This suggests that smaller models may benefit more from direct supervision, whereas larger models can better exploit the hybrid objective.

6 Conclusions

We presented a divergence-based framework for general LLM alignment that extends preference-alignment divergence estimation to RLVR. The resulting objectives, f-GRPO and f-HAL, provide on-policy and hybrid alignment methods derived from variational representations of f-divergences. We showed that these objectives estimate divergences between reward- or mixture-aligned distributions, satisfy alignment-consistency properties, and improve expected reward after alignment. Empirically, f-GRPO improves over GRPO on math reasoning, while hybrid f-HAL mitigates reward hacking in safety alignment with learned reward models. These results suggest that divergence estimation offers a unified and practical foundation for designing both reward-based and preference-based LLM alignment algorithms.

Limitations. This work is primarily foundational and aims to establish a divergence-based perspective on alignment rather than to identify a universally best divergence. While our theory shows that all canonical f-divergence objectives can achieve monotone reward improvement toward reward maximization, different divergences may induce different finite-sample training dynamics in practice. Due to resource constraints, our experiments are limited to models up to 7B parameters and are only intended to validate the implications and feasibility of our framework. We provide open-source code to support reproducibility and to guide future work on more empirically extensive studies.

Acknowledgments and Disclosure of Funding

Lin gratefully acknowledge support from DARPA MATHBAC program under grant number: HR0011-26-3-E152, the National Science Foundation (DMS-2533878, DMS-2053746, DMS-2134209, ECCS-2328241, CBET-2347401 and OAC-2311848), and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program under the "Uncertainty Quantification for Multifidelity Operator Learning (MOLUcQ)" project (Project No. 81739), DE-SC0023161, the SciDAC LEADS Institute, and DOE–Fusion Energy Science, under grant number: DE-SC0024583. Xing is supported by the National Science Foundation (DMS-2515194), Coefficient Giving, National Institutes of Health (R21AI203026), and Tutu AI Ventures LLC. The authors declare no competing interests.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bengio, Y. and Senécal, J.-S. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713–722, 2008.

Chen, M. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dang, Q.-A. and Ngo, C. Reinforcement learning for reasoning in small llms: What works and what doesn't, 2025. URL https://arxiv.org/abs/2503.16219.

Daniel Han, M. H. and team, U. Unsloth, 2023. URL http://github.com/unslothai/unsloth.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Go, D., Korbak, T., Kruszewski, G., Rozen, J., Ryu, N., and Dymetman, M. Aligning language models with preferences through f-divergence minimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11546–11583. PMLR, 2023. URL https://proceedings.mlr.press/v202/go23a.html.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Haldar, R., Wang, Z., Song, Q., Lin, G., and Xing, Y. Llm safety alignment is divergence estimation in disguise. *arXiv preprint arXiv:2502.00657*, 2025.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., and Shum, H.-Y. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. In *Advances in Neural Information Processing Systems*, 2025.

Jiang, D., Lu, Y., Li, Z., Lyu, Z., Nie, P., Wang, H., Su, A., Chen, H., Zou, K., Du, C., et al. Verltool: Towards holistic agentic reinforcement learning with tool use. *arXiv preprint arXiv:2509.01055*, 2025.

Jung, S., Han, G., Nam, D. W., and On, K.-W. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024.

Jung, S., Han, G., Nam, D. W., and On, K.-W. Binary classifier optimization for large language model alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1858–1872, 2025.

Korbak, T., Elsahar, H., Kruszewski, G., and Dymetman, M. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kruszewski, G., Erbacher, P., Rozen, J., and Dymetman, M. Whatever remains must be true: Filtering drives reasoning in llms, shaping diversity. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8984ee7aae8ae8bff19f7224f1f04a53-Abstract-Conference.html.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Le, H., Wang, Y., Gotmare, A. D., Savarese, S., and Hoi, S. C. H. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., and Shao, J. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Li, X., Zou, H., and Liu, P. Limr: Less is more for rl scaling. `https://github.com/GAIR-NLP/LIMR`, 2025.

Lin, Y., He, P., Xu, H., Xing, Y., Yamada, M., Liu, H., and Tang, J. Towards understanding jailbreak attacks in LLMs: A representation space analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7067–7085. Association for Computational Linguistics, November 2024a. doi: 10.18653/v1/2024.emnlp-main.401. URL `https://aclanthology.org/2024.emnlp-main.401/`.

Lin, Y., He, P., Xu, H., Xing, Y., Yamada, M., Liu, H., and Tang, J. Towards understanding jailbreak attacks in llms: A representation space analysis. *arXiv preprint arXiv:2406.10794*, 2024b.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*, 2021.

Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. Musr: Testing the limits of chain-of-thought with multistep soft reasoning, 2024. URL `https://arxiv.org/abs/2310.16049`.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. In *NeurIPS*, 2020.

Team, Q. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Vojnovic, M. and Yun, S.-Y. What is the alignment objective of grpo? *arXiv preprint arXiv:2502.18548*, 2025.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL `https://arxiv.org/abs/2406.01574`.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., and Fan, Z. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., and Zhang, Z. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b.

Yu, Q., Zhang, Z., Zhu, R., et al. Dapo: An open-source llm reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, 2025.

Zheng, C., Yin, F., Zhou, H., Meng, F., Zhou, J., Chang, K.-W., Huang, M., and Peng, N. On prompt-driven safeguarding for large language models. *arXiv preprint arXiv:2401.18018*, 2024.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Additional Background on Alignment Objectives

A.1 Notation

Let x denote the prompt and y a response. For preference alignment, we use y_w and y_l to denote preferred and less-preferred responses. The triplet $(x, y_w, y_l) \sim \mathcal{D}$ denotes the joint distribution over prompts and response pairs. Marginalizing this joint gives $(x, y_w) \sim \mathcal{D}^+$, $(x, y_l) \sim \mathcal{D}^-$, and $x \sim \mathcal{P}$, where $\mathcal{D}^+$ and $\mathcal{D}^-$ are the aligned and unaligned response distributions. For any distribution $\mathcal{G}$, let $\mathcal{G}_x$ and $p_{\mathcal{G}}(y|x)$ denote its marginal over prompts and conditional density or mass function over responses.

Let π_θ be the policy under optimization, π_{ref} the reference policy before alignment, and $\pi_{\theta_{\text{old}}}$ a previous policy iterate used only for on-policy sampling. We write $r(x, y)$ for an explicit environmental reward and

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

for the implicit policy reward, where $\beta > 0$ is the KL regularization coefficient. Finally, $\sigma(\cdot)$ and $\vec{\sigma}(\cdot)$ denote the sigmoid and softmax functions.

A.2 GRPO Objective

On-policy algorithms such as PPO and GRPO optimize the reward objective by repeatedly sampling from an old policy $\pi_{\theta_{\text{old}}}$ and optimizing an importance-weighted surrogate:

$$\sup_\theta \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)} A(x, y), \tag{19}$$

where $A(x, y)$ is an advantage estimate that depends on the reward. GRPO estimates the advantage by sampling G responses $\{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|x)$ and computing

$$a_i = A(x, y_i) = \frac{r(x, y_i) - \hat{\mu}(r(x, \cdot))}{\hat{\sigma}(r(x, \cdot))}. \tag{20}$$

The clipped GRPO objective is

$$\sup_\theta \mathop{\mathbb{E}}_{x} \left[\sum_{i=1}^G \frac{1}{G} \min\big(R_\theta^i a_i, \text{Clip}(R_\theta^i, 1-\epsilon, 1+\epsilon) a_i\big) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta(\cdot|x) \parallel \pi_{\text{ref}}(\cdot|x)) \right], \tag{21}$$

where

$$R_\theta^i = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}.$$

The clipping term stabilizes the update by preventing large policy-ratio changes within a single iteration.

A.3 Preference Alignment and Direct Alignment

In preference alignment, one observes preference data $(x, y_w, y_l) \sim \mathcal{D}$ or binary samples $(x, y_w) \sim \mathcal{D}^+$ and $(x, y_l) \sim \mathcal{D}^-$. Since the true scalar reward is unavailable, classical RLHF first trains a reward model $r_\phi(x, y)$ using the Bradley–Terry model:

$$p(y_w \succ y_l \mid x) = \frac{\exp r_\phi(x, y_w)}{\exp r_\phi(x, y_w) + \exp r_\phi(x, y_l)}. \tag{22}$$

The learned reward model is then used in a KL-regularized RL objective.

Direct alignment methods avoid explicitly training a reward model. Using the optimal-policy relation in Equation (3), the reward can be parameterized by the policy as

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}.$$

Substituting this implicit reward into the Bradley–Terry likelihood yields the DPO loss (Rafailov et al., 2023):

$$\mathcal{L}_{\text{DPO}}(\theta) = - \mathop{\mathbb{E}}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma(r_\theta(x, y_w) - r_\theta(x, y_l)) \,. \tag{23}$$

Other direct alignment methods, such as BCO (Jung et al., 2025) and KTO (Ethayarajh et al., 2024), can also be interpreted as increasing likelihood on aligned samples and decreasing likelihood on unaligned samples.

A.4 Derivation of the FDO Objective

We derive the direct alignment objective (FDO) from the variational representation of f-divergences. For completeness, we briefly restate the key definition from Appendix B and apply it to aligned and unaligned response distributions. For two positive measures P and Q, the f-divergence is

$$D_f(P\|Q) = \mathop{\mathbb{E}}_{v\sim Q}\left[f\left(\frac{dP}{dQ}(v)\right)\right], \tag{24}$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is convex and satisfies $f(1) = 0$. Its variational representation is

$$D_f(P\|Q) = \sup_{T:\mathcal{V}\to \text{ED}(f^*)} \left\{ \mathop{\mathbb{E}}_{P}[T(v)] - \mathop{\mathbb{E}}_{Q}[f^*(T(v))] \right\}, \tag{25}$$

where f^* is the convex conjugate of f (refer Definition B.2).

Applying this representation to $P = \mathcal{D}^+$ and $Q = \mathcal{D}^-$, and choosing the variational function as $T = g(r_\theta)$, gives the FDO objective:

$$\mathcal{L}_{\text{FDO}(f,g)}(\theta) = - \mathop{\mathbb{E}}_{\mathcal{D}^+}\left[g(r_\theta)\right] + \mathop{\mathbb{E}}_{\mathcal{D}^-}\left[f^* \circ g(r_\theta)\right]. \tag{26}$$

At convergence, the objective recovers the negative f-divergence, a consequence of f-divergence representation optimality Lemma B.3:

$$\mathcal{L}_{\text{FDO}(f,g)}(\theta^*) = -D_f(\mathcal{D}^+\|\mathcal{D}^-). \tag{27}$$

This is the starting point for our extension from preference alignment to RLVR.

Name	$f(t)$	$f^*(u)$	eff. dom $\mathrm{ED}(f^*)$	Canonical link: $g(u)$	$g^{-1}(u)$	Mixture Scaler $g^{-1} \circ f'(t)$	f'_∞
Hellinger	$(\sqrt{t}-1)^2$	$\frac{u}{1-u}$	$u<1$	$1-e^{-u}$	$-\ln(1-u)$	$\frac{1}{2}\ln t$	1
Jensen–Shannon	$\left[t\ln t-(t+1)\ln\frac{t+1}{2}\right]$	$-\ln\left(2-e^{u}\right)$	$u<\ln 2$	$\ln 2-\ln(1+e^{-u})$	$-\ln\left(e^{\ln 2-u}-1\right)$	$\ln t$	$\ln 2$
KL	$t\ln t$	e^{u-1}	$\mathbb{R}$	u	u	$\ln t+1$	∞
Pearson χ^2	$(t-1)^2$	$\frac{1}{4}u^2+u$	$\mathbb{R}$	u	u	$2(t-1)$	∞
Reverse KL	$-\ln t$	$-1-\ln(-u)$	$u<0$	$-e^{-u}$	$-\ln(-u)$	$\ln t$	0
Total Variation	$\frac{1}{2}\lvert t-1\rvert$	u	$\lvert u\rvert\le 1/2$	$\frac{1}{2}(1+e^{-u})^{-1}$	$-\ln((2u)^{-1}-1)$	$-\ln((\mathrm{sign}(t-1))^{-1}-1)$	$1/2$

Table 9: Common f-divergences with conjugates, effective domains, canonical links, inverse links, mixture scalers, and tail slopes. Canonical links are characterized by functions $g:\mathbb{R}\to\mathrm{ED}(f^*)$ s.t. $g^{-1}\circ f'_\infty=\infty$.

B Mathematical Background on f-Divergences

Expectation w.r.t. a positive measure. Let μ be any positive (finite) measure on a measurable space $(\Omega,\mathcal{F})$ and let $h:\Omega\to\mathbb{R}$ be measurable. We define the expectation with respect to μ as

$$\mathop{\mathbb{E}}_{\mu}[h] \;:=\; \int_\Omega h(\omega)\,d\mu(\omega).$$

When μ is a probability measure, this reduces to the usual notion of expectation.

Definition B.1 (f-Divergence)**.** For two positive measures $\mathcal{P},\mathcal{Q}$ with $\mathcal{P}\ll\mathcal{Q}$ (absolute continuous), the f-divergence between the two measures is defined as follows:

$$D_f(\mathcal{P}\|\mathcal{Q})=\mathop{\mathbb{E}}_{v\sim\mathcal{Q}}\left[f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}(v)\right)\right],$$

where $f:\mathbb{R}_+\to\mathbb{R}$ is a convex function with $f(1)=0$ and $\mathcal{P}\ll\mathcal{Q}:d\mathcal{Q}(v)=0\implies d\mathcal{P}(v)=0$.

General Case. For arbitrary $\mathcal{P},\mathcal{Q}$, let $\mathcal{P}=\mathcal{P}_{\mathrm{ac}}+\mathcal{P}_{\mathrm{s}}$ be the Lebesgue decomposition of $\mathcal{P}$ w.r.t. $\mathcal{Q}$ into absolute continuous and singular measures. The f-divergence is defined as

$$D_f(\mathcal{P}\|\mathcal{Q})=\mathop{\mathbb{E}}_{v\sim\mathcal{Q}}\left[f\left(\frac{d\mathcal{P}_{\mathrm{ac}}}{d\mathcal{Q}}(v)\right)\right]+f'_\infty\,\mathcal{P}_{\mathrm{s}}(\Omega),$$

where $f'_\infty:=\lim_{t\to\infty}\frac{f(t)}{t}\in[0,\infty]$.

Definition B.2 (Convex Conjugate)**.**

$$f^*(u)=\sup_{t\in\mathbb{R}}\{ut-f(t)\},$$

where $f^*(u)$ is the convex conjugate of f. Note that f^* is also a convex function.

Lemma B.3 (Variational Representation & Optimality)**.** *Let $\mathcal{P},\mathcal{Q}$ be probability measures on $\mathcal{V}$, and write the Lebesgue decomposition $\mathcal{P}=\mathcal{P}_{\mathrm{ac}}+\mathcal{P}_{\mathrm{s}}$ w.r.t. $\mathcal{Q}$, with density $r:=\frac{d\mathcal{P}_{\mathrm{ac}}}{d\mathcal{Q}}$. Let $S\in\mathcal{V}$ satisfy $\mathcal{Q}(S)=0$ and $\mathcal{P}_{\mathrm{s}}(S)=\mathcal{P}_{\mathrm{s}}(\mathcal{V})$ (support of $\mathcal{P}_{\mathrm{s}}$). Define $f'_\infty:=\lim_{t\to\infty}\frac{f(t)}{t}\in[0,\infty]$. Then*

$$D_f(\mathcal{P}\|\mathcal{Q})=\sup_{T:\mathcal{V}\to\mathrm{ED}(f^*)}\Big(\mathop{\mathbb{E}}_{\mathcal{P}}[T(v)]-\mathop{\mathbb{E}}_{\mathcal{Q}}[f^*(T(v))]\Big),\tag{28}$$

where f^* is the convex conjugate of f and $\mathrm{ED}(f^*)=\{u:f^*(u)<\infty\}$. Moreover, the optimal functional can be chosen as

$$T^*(v)\in\partial f(r(v))\;\mathbb{1}_{S^c}(v)\;+\;f'_\infty\,\mathbb{1}_S(v),\tag{29}$$

(and if f is differentiable, $\partial f(r)=\{f'(r)\}$).

Proof. For any measurable $T:\mathcal{V}\to\mathrm{ED}(f^*)$, Fenchel–Young gives

$$t\,u\;\le\;f(t)+f^*(u)\qquad(t\ge 0,\;u\in\mathrm{ED}(f^*)).$$

Applying this pointwise with $(t,u)=(r(v),T(v))$ and integrating over $\mathcal{Q}$ yields

$$\mathop{\mathbb{E}}_{\mathcal{P}_{\mathrm{ac}}}[T]-\mathop{\mathbb{E}}_{\mathcal{Q}}[f^*(T)]=\mathop{\mathbb{E}}_{\mathcal{Q}}[rT-f^*(T)]\le\mathop{\mathbb{E}}_{\mathcal{Q}}[f(r)].$$

For the singular part, for any u and any $t \geq 0$, $f(t) \geq tu - f^*(u)$ implies

$$\frac{f(t)}{t} \geq u - \frac{f^*(u)}{t} \quad \Rightarrow \quad u \leq f'_\infty \quad \text{whenever } f^*(u) < \infty.$$

Hence $T \leq f'_\infty$ $\mathcal{P}_{\mathrm{s}}$-a.s., and so

$$\mathop{\mathbb{E}}_{\mathcal{P}_{\mathrm{s}}}[T] \leq f'_\infty \, \mathcal{P}_{\mathrm{s}}(\Omega).$$

Combining,

$$\mathop{\mathbb{E}}_{\mathcal{P}}[T] - \mathop{\mathbb{E}}_{\mathcal{Q}}[f^*(T)] \leq \mathop{\mathbb{E}}_{\mathcal{Q}}[f(r)] + f'_\infty \, \mathcal{P}_{\mathrm{s}}(\Omega) = D_f(\mathcal{P}\|\mathcal{Q}),$$

so the RHS of (28) is $\leq D_f(\mathcal{P}\|\mathcal{Q})$.

To see tightness, take $T^*(v) \in \partial f(r(v))$ on $\{r > 0\}$ (equivalently $T^* = f'(r)$ when differentiable), so that $rT^* - f^*(T^*) = f(r)$ $\mathcal{Q}$-a.s. For the singular part, choose a set S with $\mathcal{Q}(S) = 0$, $\mathcal{P}_{\mathrm{s}}(S) = \mathcal{P}_{\mathrm{s}}(\Omega)$, and let $T_n = f'_\infty \mathbf{1}_S + T^* \mathbf{1}_{S^c}$ (or $T_n \uparrow f'_\infty \mathbf{1}_S$ if needed). Then

$$\mathop{\mathbb{E}}_{\mathcal{P}}[T_n] - \mathop{\mathbb{E}}_{\mathcal{Q}}[f^*(T_n)] \uparrow \mathop{\mathbb{E}}_{\mathcal{Q}}[f(r)] + f'_\infty \mathcal{P}_{\mathrm{s}}(\Omega) = D_f(\mathcal{P}\|\mathcal{Q}).$$

Thus the supremum equals $D_f(\mathcal{P}\|\mathcal{Q})$, proving (28). □

C Proofs and Additional Theoretical Details

Lemma C.1 (Comonotone Covariance Inequality)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable. If $f, g : \mathbb{R} \to \mathbb{R}$ are integrable and both non- decreasing (or both non-increasing), then*

$$\mathbb{E}[f(X)g(X)] \;\geq\; \mathbb{E}[f(X)]\,\mathbb{E}[g(X)],$$

equivalently $\mathrm{Cov}(f(X), g(X)) \geq 0$.

Proof. Let X' be an independent copy of X. Then

$$2\ \mathrm{Cov}(f(X), g(X)) = \mathbb{E}\Big[(f(X) - f(X'))(g(X) - g(X'))\Big].$$

(Indeed, expanding the RHS and using independence gives $\mathbb{E}[f(X)g(X)] + \mathbb{E}[f(X')g(X')] - \mathbb{E}[f(X)g(X')] - \mathbb{E}[f(X')g(X)] = 2\,\mathbb{E}[f(X)g(X)] - 2\,\mathbb{E}[f(X)]\,\mathbb{E}[g(X)]$.)

If f and g are both non-decreasing, then for every (x, x'),

$$(x - x') \geq 0 \;\Rightarrow\; f(x) - f(x') \geq 0 \;\text{ and }\; g(x) - g(x') \geq 0,$$

and similarly if $(x - x') \leq 0$ then both differences are ≤ 0. Hence $(f(x) - f(x'))(g(x) - g(x')) \geq 0$ pointwise, so the expectation is nonnegative:

$$\mathbb{E}\Big[(f(X) - f(X'))(g(X) - g(X'))\Big] \;\geq\; 0.$$

Therefore $\mathrm{Cov}(f(X), g(X)) \geq 0$, which is equivalent to $\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\,\mathbb{E}[g(X)]$. The non-increasing case follows identically. □

Lemma C.2 (Monotone Reweighting Increases the Mean)**.** *Fix x and let $q(\cdot \mid x)$ be a distribution over y. For any nonnegative weight function $w_x(y)$ with $\mathbb{E}_q[w_x] > 0$, define the tilted distribution*

$$q_w(y \mid x) \;=\; \frac{q(y \mid x)\, w_x(y)}{\mathbb{E}_{y \sim q}[w_x(y)]}.$$

Then for any measurable reward $r(x, y)$,

$$\mathop{\mathbb{E}}_{q_w}[r(x, y)] - \mathop{\mathbb{E}}_{q}[r(x, y)] = \frac{\mathrm{Cov}_{q(\cdot|x)}\big(r(x, y),\, w_x(y)\big)}{\mathbb{E}_q[w_x(y)]}.$$

In particular, if $w_x(y)$ is (weakly) increasing in $r(x, y)$, then

$$\mathop{\mathbb{E}}_{q_w}[r(x, y)] \;\geq\; \mathop{\mathbb{E}}_{q}[r(x, y)],$$

with strict inequality unless $r(x, \cdot)$ is q-a.s. constant or w_x is q-a.s. constant.

Proof. By definition,

$$\mathop{\mathbb{E}}_{q_w}[r] = \int_y r(y)\,\frac{q(y)w(y)}{\mathbb{E}_q[w]}\cdot dy = \frac{\mathbb{E}_q[rw]}{\mathbb{E}_q[w]}.$$

Hence

$$\mathop{\mathbb{E}}_{q_w}[r] - \mathop{\mathbb{E}}_{q}[r] = \frac{\mathbb{E}_q[rw] - \mathbb{E}_q[r]\,\mathbb{E}_q[w]}{\mathbb{E}_q[w]} = \frac{\mathrm{Cov}_q(r,w)}{\mathbb{E}_q[w]}.$$

If w is non decreasing in r, then r and $w(r)$ are comonotone, so $\mathrm{Cov}_q(r,w) \geq 0$ (Lemma C.1), yielding $\mathbb{E}_{q_w}[r] \geq \mathbb{E}_q[r]$, with strict inequality unless r or w is q-a.s. constant. □

Proof of Theorem 4.3. **Divergence Estimation and Alignment Consistency:**
(i) F-GRPO: The loss can be expressed as :

$$\mathcal{L}^{(f,g)}_{\text{f-GRPO}}(\theta) = \frac{1+\beta^{-1}}{G}\mathop{\mathbb{E}}_{x}\sum_{y_i\sim\pi_{\theta_{\text{old}}},i=1}^{G}\left(-a_i\hat{w}_i^{+}\cdot g(r_{\theta,i}) - a_i\hat{w}_i^{-}\cdot f^*\circ g(r_{\theta,i})\right) \tag{30}$$

where $\hat{w}_i^{\pm} = \vec{\sigma}_i\big(\{\pm r_j - \ln\pi_{\theta_{\text{old}}}(y_j|x)\}_{j=1}^{G}\big)\cdot\mathbb{1}_{\{\pm a_i>0\}}$ [Equation (9)]. As $G\to\infty$, we can use Strong Law of Large numbers, Continuous mapping theorem and softmax with log-Q correction to get the following:

- $\hat{\mu}(r(x,.)) \xrightarrow{a.s.} \mu_r^{\theta_{\text{old}}} = \mathbb{E}_{\pi_{\theta_{\text{old}}}} r(x,y)$; $\hat{\sigma}(r(x,.)) \xrightarrow{a.s.} \sigma_r^{\theta_{\text{old}}} = \sqrt{\mathbb{V}_{\pi_{\theta_{\text{old}}}} r(x,y)}$
- $a_i \xrightarrow{a.s.} \frac{r(x,y_i)-\mu_r^{\theta_{\text{old}}}}{\sigma_r^{\theta_{\text{old}}}}$
- $\vec{\sigma}_i\big(\{\pm r_j - \ln\pi_{\theta_{\text{old}}}(y_j|x)\}_{j=1}^{G}\big)\cdot G \xrightarrow{a.s.} \frac{d\mathcal{Q}^{\pm}}{\pi_{\theta_{\text{old}}}}(y_i|x)$, where $d\mathcal{Q}^{\pm}\propto\exp\pm r(x,y)$.
- $\mathbb{1}_{\{\pm a_i>0\}} \xrightarrow{a.s.} \mathbb{1}_{\{\pm r(x,y_i)\mp\mu_r^{\theta_{\text{old}}}>0\}}$

Combining the above with the convergence in the empirical distribution sampled from $\pi_{\theta_{\text{old}}}$, we have:

$$\begin{aligned}\mathcal{L}^{(f,g)}_{\text{f-GRPO}}(\theta) &\overset{a.s.}{=} (1+\beta^{-1})\mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y\sim\pi_{\theta_{\text{old}}}}\left(-\frac{r(x,y)-\mu_r^{\theta_{\text{old}}}}{\sigma_r^{\theta_{\text{old}}}}\left[\frac{d\mathcal{Q}^{+}}{\pi_{\theta_{\text{old}}}}(y|x)\cdot g(r_\theta)\mathbb{1}_{\{r(x,y_i)-\mu_r^{\theta_{\text{old}}}>0\}} + \right.\right.\\ &\qquad\left.\left.\frac{d\mathcal{Q}^{-}}{\pi_{\theta_{\text{old}}}}(y|x)\cdot f^*\circ g(r_{\theta,i})\mathbb{1}_{\{\mu_r^{\theta_{\text{old}}}-r(x,y_i)>0\}}\right]\right)\\ &= (1+\beta^{-1})\sigma_r^{\theta_{\text{old}}}\left(-\mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y\sim\mathcal{Q}^{+}}[r(x,y)-\mu_r^{\theta_{\text{old}}}]_+ g(r_\theta) + \mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y\sim\mathcal{Q}^{-}}[-r(x,y)+\mu_r^{\theta_{\text{old}}}]_+ f^*\circ g(r_\theta)\right)\end{aligned}$$

Given measures $\mathcal{D}^{+}_{(r,\theta_{\text{old}})}, \mathcal{D}^{-}_{(r,\theta_{\text{old}})}$ with densities

$$d\mathcal{D}^{\pm}_{(r,\theta_{\text{old}})}(x,y) \;\propto\; [\pm(r(x,y)-\mu_r^{\theta_{\text{old}}}(x))]_+\;\exp(\pm r(x,y))\,d\mathcal{P}(x),$$

which matches Def. 4.1 (with $\pi_{\theta_{\text{old}}}$ in place of π_θ). Then up-to scaling constants the above can be simplified as

$$\text{const}\cdot\mathcal{L}^{(f,g)}_{\text{f-GRPO}}(\theta) \overset{a.s.}{=} \Big(-\mathop{\mathbb{E}}_{\mathcal{D}^{+}_{(r,\theta_{\text{old}})}}[\,c(x)\cdot g(r_\theta)\,] + \mathop{\mathbb{E}}_{\mathcal{D}^{-}_{(r,\theta_{\text{old}})}}[\,f^*\circ g(r_\theta)\,]\Big), \tag{31}$$

where $c(x)$ is a normalization constant. Using Lemma B.3 As $\mathcal{D}^{+}_{(r,\theta_{\text{old}})} \perp \mathcal{D}^{-}_{(r,\theta_{\text{old}})}$, that is the supports are complementary (no subspace where $\mathcal{D}^{+}_{(r,\theta_{\text{old}})} \ll \mathcal{D}^{-}_{(r,\theta_{\text{old}})}$) then the divergence is just:

$$\inf_\theta \mathcal{L}^{(f,g)}_{\text{f-GRPO}}(\theta) \propto -\mathbb{D}_f(c(x)\mathcal{D}^{+}_{(r,\theta_{\text{old}})} \parallel \mathcal{D}^{-}_{(r,\theta_{\text{old}})}) = -f'_\infty = -\mathbb{D}_f(\mathcal{D}^{+}_{(r,\theta_{\text{old}})} \parallel \mathcal{D}^{-}_{(r,\theta_{\text{old}})})$$

For t^{th} FP iterate, $\theta^{(t)}_{\text{RL}} = \theta_{\text{old}}$ and $\theta^{(t+1)}_{\text{RL}} = \arg\inf_\theta \mathcal{L}^{(f,g)}_{\text{f-GRPO}}(\theta)$

$$\mathcal{L}^{(f,g)}_{\text{f-GRPO}}(\theta^{(t+1)}) \propto -\mathbb{D}_f(\mathcal{D}^{+}_{(r,\theta^{(t)})} \parallel \mathcal{D}^{-}_{(r,\theta^{(t)})})$$

Moreover, using $T = g(r_\theta)$, the optimality Lemma B.3 for the $(t+1)$th FP iterate is attained at $T^{(t+1)} = g(r_{\theta_{\text{RL}}^{(t+1)}}) = \beta \ln \frac{\pi_{\theta_{\text{RL}}^{(t+1)}}}{\pi_{\text{ref}}} = f'_\infty \mathbb{1}_{\mathcal{D}^-_{(r,\theta_{\text{RL}}^{(t)})}(x,y)=0}$.

$$\pi_{\theta_{\text{RL}}^{(t+1)}}(y|x) \propto \pi_{\text{ref}}(y|x) \exp\left(\beta^{-1} g^{-1}(f'_\infty \mathbb{1}_{\mathcal{D}^-_{(r,\theta_{\text{RL}}^{(t)})}(x,y)=0})\right) \tag{32}$$

(ii) F-HAL The hybrid alignment loss is λ weighted avergae between the direct aligner FDO loss and f-GRPO:

$$\mathcal{L}^{(f,g)}_{\text{f-HAL}}(\theta) = \lambda \mathcal{L}^{(f,g)}_{\text{FDO}}(\theta) + (1-\lambda)\mathcal{L}^{(f,g)}_{\text{f-GRPO}}(\theta),$$

where $\mathcal{L}_{\text{FDO}(f,g)}(\theta) = -\mathbb{E}_{\mathcal{D}^+} g(r_\theta) + \mathbb{E}_{\mathcal{D}^-} f^* \circ g(r_\theta)$ with infimum at $-\mathbb{D}_f(\mathcal{D}^+ \parallel \mathcal{D}^-)$ using the variational representation Lemma B.3. We know that as $G \to \infty$, F-GRPO loss is almost surely equal to the representation in Equation (31). As a corollary, F-HAL is almost surely equal to the following, up to some scaling constants:

$$\begin{aligned}\mathcal{L}^{(f,g)}_{\text{f-HAL}}(\theta) \propto \lambda &\left(-\mathop{\mathbb{E}}_{\mathcal{D}^+} g(r_\theta) + \mathop{\mathbb{E}}_{\mathcal{D}^-} f^* \circ g(r_\theta)\right) \\ &+ (1-\lambda)\left(-\mathop{\mathbb{E}}_{\mathcal{D}^+_{(r,\theta_{\text{old}})}} [\, c^+(x) \cdot g(r_\theta)\,] + \mathop{\mathbb{E}}_{\mathcal{D}^-_{(r,\theta_{\text{old}})}} c^-(x) \cdot [\, f^* \circ g(r_\theta)\,]\right)\end{aligned}$$

For mixture measures $\mathcal{M}^\pm_{\text{Mix}} = \lambda \mathcal{D}^\pm + (1-\lambda) c^\pm(x) \mathcal{D}^\pm_{(r,\theta_{\text{old}})}$, then:

$$\inf_\theta \mathcal{L}^{(f,g)}_{\text{f-HAL}}(\theta) \propto -\mathbb{D}_f(\mathcal{M}^+_{\text{Mix}} \parallel \mathcal{M}^-_{\text{Mix}}).$$

For t^{th} FP iterate, $\theta^{(t)}_{\text{Mix}} = \theta_{\text{old}}$ and $\theta^{(t+1)}_{\text{Mix}} = \arg\inf_\theta \mathcal{L}^{(f,g)}_{\text{f-HAL}}(\theta)$, and

$$\mathcal{L}^{(f,g)}_{\text{f-HAL}}(\theta^{(t+1)}_{\text{Mix}}) \propto -\mathbb{D}_f(\mathcal{M}^+_{t,\text{Mix}} \parallel \mathcal{M}^-_{t,\text{Mix}})$$

Moreover, using $T = g(r_\theta)$ and the optimality Lemma B.3, the $(t+1)$th FP iterate is attained at $T^{(t+1)} = g(r_{\theta^{(t+1)}_{\text{Mix}}}) = \beta \ln \frac{\pi_{\theta^{(t+1)}_{\text{Mix}}}}{\pi_{\text{ref}}} = f'(\frac{d\mathcal{M}^+_{t,\text{Mix}}}{d\mathcal{M}^-_{t,\text{Mix}}})$. This implies that

$$\pi_{\theta_{\text{RL}}^{(t+1)}}(y|x) \propto \pi_{\text{ref}}(y|x) \exp\left(\beta^{-1} g^{-1} \circ f'(\frac{d\mathcal{M}^+_{t,\text{Mix}}}{d\mathcal{M}^-_{t,\text{Mix}}})\right) \tag{33}$$

□

Note that Equations (32) and (33) are indeed alignment consistent and assign more mass/density to responses from $(+)$ distribution $(-)$. To see that the probability density (33) is non-decreasing in $R(x,y)$ where $R(x,y) = \frac{d\mathcal{M}^+_{t,\text{Mix}}}{d\mathcal{M}^-_{t,\text{Mix}}}$, we note that ${g^{-1}}'(u) > 0$ as $g \circ g^{-1}(u) = u \implies {g^{-1}}'(u) = \frac{1}{g'(g^{-1}(u))} > 0$ (monotonicity of $g(u)$). Combined with the fact that $f'' \geq 0$ as f is convex, we have that $e^{\beta^{-1} h_t(R(x,y))} = \exp(\beta^{-1} \cdot g^{-1}(f'(R(x,y))))$ is non-decreasing. Similarly, (32) is also non-decreasing in $R(x,y) = \frac{d\mathcal{D}^+_{(r,\theta_{\text{RL}}^{(t)})}}{d\mathcal{D}^-_{(r,\theta_{\text{RL}}^{(t)})}}$.

Average Reward Improvement:
We know that g^{-1} is monotone increasing and $g^{-1} \circ f'$ is non-decreasing.
We can apply Lemma C.2 with $q = \pi_{\text{ref}}$ and $w_x(y) = \exp(\beta^{-1} g^{-1}(f'_\infty \mathbb{1}_{\mathcal{D}^-_{(r,\theta_{\text{RL}}^{(t)})}=0}))$ for f-GRPO and $w_x(y) = \exp(\beta^{-1} g^{-1}(f'(\frac{d\mathcal{M}^+_{t,\text{Mix}}}{d\mathcal{M}^-_{t,\text{Mix}}}))$ for f-HAL (Note: By Assumption 4.2 $d\mathcal{D}^\pm$ is non-decreasing/increasing in $r(x,y)$, hence the mixture density ratio is also monotone in $r(x,y)$.) to get:

$$\mu_r^{\theta_{\text{ref}}}(x) < \mu_r^{\theta^{(t+1)}}(x),$$

for all $t \geq 0$ & $\theta^{(t)} \in \{\theta^{(t)}_{\text{Mix}}, \theta^{(t)}_{\text{RL}}\}$.
For the special case when $g^{-1}(f'_\infty) = \infty$, applying Lemma C.2 with $q = \pi_{\theta^{(t)}_{\text{RL}}}$ and $w_x(y) =$

$\exp(\beta^{-1}[g^{-1}(f'_\infty \mathbb{1}_{\mathcal{D}^-_{(r,\theta^{(t)}_{\text{RL}})}=0}) - g^{-1}(f'_\infty \mathbb{1}_{\mathcal{D}^-_{(r,\theta^{(t-1)}_{\text{RL}})}=0})])$, on $\text{Supp}(\pi_{\theta^{(t)}_{\text{RL}}}) = \{y : \mathcal{D}^-_{(r,\theta^{(t-1)}_{\text{RL}})} = 0\}$ which simplifies to $w_x(y) = \begin{cases} 0, & y \in \text{Supp}(\pi_{\theta^{(t)}_{\text{RL}}}) \ni \mathcal{D}^-_{(r,\theta^{(t)}_{\text{RL}})} = 0 \\ -\infty, & y \in \text{Supp}(\pi_{\theta^{(t)}_{\text{RL}}}) \ni \mathcal{D}^-_{(r,\theta^{(t)}_{\text{RL}})} > 0. \end{cases}$ Basically, the support after each iteration shrinks further and further and increases the average reward at each iteration:

$$\mu_r^{\theta^{(t)}_{RL}}(x) < \mu_r^{\theta^{(t+1)}_{RL}}(x)$$

This continues strictly until the FP iteration stops with $\mu_r^{\theta^{(t)}_{RL}}(x) = \max_y r(x,y)$ for some t.

Proof of Theorem 4.4. Simplifying Equation (21), the unclipped GRPO objective (or $\epsilon \to \infty$) can be expressed as:

$$\sup_\theta \mathop{\mathbb{E}}_x \left(\sum_{y_i \sim \pi_{\theta_{\text{old}}}, i=1}^{G} \frac{1}{G} \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)} a_i - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right)$$

As $G \to \infty$, using strong law of large numbers and continuous mapping, $a_i = A^{\pi_\theta}(x, y_i) \xrightarrow{a.s.} \frac{r(x,y_i) - \mu_r^{\theta_{\text{old}}}}{\sigma_r^{\theta_{\text{old}}}}$. Furthermore, the empirical distribution converges to samples from $\pi_{\theta_{\text{old}}}$, leading to:

$$\sup_\theta \mathop{\mathbb{E}}_x \left(\mathop{\mathbb{E}}_{y \sim \pi_{\theta_{\text{old}}}} \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)} \cdot \frac{r(x,y) - \mu_r^{\theta_{\text{old}}}}{\sigma_r^{\theta_{\text{old}}}} - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right)$$

With $a(x,y) = \frac{r(x,y) - \mu_r^{\theta_{\text{old}}}}{\sigma_r^{\theta_{\text{old}}}}$ The above simplifies to:

$$-\inf_\theta \mathcal{L}_{\text{GRPO}}(\theta) = \sup_\theta \mathop{\mathbb{E}}_x \left(\mathop{\mathbb{E}}_{y \sim \pi_\theta} a(x,y) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right) = \sup_\theta \mathop{\mathbb{E}}_x \mathop{\mathbb{E}}_{y \sim \pi_\theta} \left(a(x,y) - \beta \ln \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right)$$

The above objective is maximized when $a(x,y) - \beta[\ln \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + 1] = 0$. (first order conditions). Hence, upto normalization constant, for the tth FP iterate Equation (14) with $\theta^{(t)} = \theta_{\text{old}}$ and $\theta^{(t+1)} = \arg\inf_\theta \mathcal{L}_{\text{GRPO}}(\theta)$:

$$\pi_{\theta^{(t+1)}}(y|x) \propto \pi_{\text{ref}}(y|x) \exp\left[\frac{a_t(x,y)}{\beta} + 1\right]$$

$$\pi_{\theta^{(t+1)}}(y|x) \propto \pi_{\text{ref}}(y|x) \exp \frac{a_t(x,y)}{\beta}$$

□

C.1 Justification for importance weights

In RLVR, we do not observe samples from explicit aligned and unaligned response distributions $\mathcal{D}^+$ and $\mathcal{D}^-$. Instead, we only observe scalar rewards $r(x,y)$ for responses sampled from the policy. To construct an on-policy analogue of divergence estimation, we therefore need a way to use rewards to induce aligned and unaligned weights over policy samples.

The guiding principle is that, for a fixed prompt x, higher-reward responses should be treated as more aligned, while lower-reward responses should be treated as more unaligned. A natural score-based instantiation of this principle is to use the reward itself as the aligned score and the negative reward as the unaligned score. That is, we define reward-induced aligned and unaligned densities by

$$\tilde{p}^{\pm}(y|x) \;\propto\; \exp(\pm r(x,y)), \tag{34}$$

for any fixed x.

This construction should be interpreted as a device for motivating the weighting scheme used in f-GRPO and f-HAL, rather than as a statistical assumption on the true aligned and unaligned response distributions. In particular, we do not require the true $\mathcal{D}^+$ and $\mathcal{D}^-$, when they exist, to follow (34); the theoretical guarantees in the main text do not rely on this modeling choice.

Given the reward-induced densities in (34), an importance-sampling view suggests weighting samples from $\pi_{\theta_{\text{old}}}(\cdot|x)$ by

$$\frac{\tilde{p}^{\pm}(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}.$$

However, directly using all samples can yield high-variance updates and may include responses that are not informative for separating aligned from unaligned behavior. We therefore use truncated importance weights:

$$\tilde{w}^{\pm}_{x,y} = \frac{\tilde{p}^{\pm}(y|x)}{\pi_{\theta_{\text{old}}}(y|x)} \cdot I^{\pm}(r(x,y)), \tag{35}$$

where

$$I^{\pm}(r(x,y)) = \mathbb{1}_{\left\{\pm[r(x,y)-\mathbb{E}_{\pi_{\theta_{\text{old}}}} r(x,y)]>0\right\}}.$$

Thus, the aligned weights retain above-average reward responses, while the unaligned weights retain below-average reward responses. This mirrors the separation mechanism in GRPO, where positive-advantage samples are promoted and negative-advantage samples are suppressed.

C.1.1 Estimating the Importance Sampling Term

We now derive the empirical form of the weights used in the main text. For a fixed prompt x, we sample G responses $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$, fetch rewards $\{r(x,y_i)\}_{i=1}^{G}$, and compute the group advantages $\{A^{\pi_\theta}(x,y_i)\}_{i=1}^{G}$ as in Eq. 4. For notational convenience, we write

$$r_i := r(x,y_i), \qquad a_i := A^{\pi_\theta}(x,y_i).$$

The density in (34) can be written explicitly as

$$\tilde{p}^{\pm}(y|x) = \frac{\exp(\pm r(x,y))}{Z^{\pm}(x)},$$

where the normalizing constant is

$$Z^{\pm}(x) = \sum_{y' \in \mathcal{Y}(x)} \exp(\pm r(x,y')).$$

Equivalently, if $\text{Unif}(\cdot|x)$ denotes the uniform distribution over the response space for prompt x, then

$$Z^{\pm}(x) = |\mathcal{Y}(x)| \mathop{\mathbb{E}}_{y' \sim \text{Unif}(\cdot|x)} \exp(\pm r(x,y')).$$

Since the factor $|\mathcal{Y}(x)|$ is independent of y, it cancels in the normalized weights. Thus, up to a prompt-dependent constant, we may write

$$\tilde{p}^{\pm}(y|x) \propto \frac{\exp(\pm r(x,y))}{\mathbb{E}_{y' \sim \text{Unif}(\cdot|x)} \exp(\pm r(x,y'))}.$$

The uniform expectation is intractable in language generation, and our samples are instead drawn from $\pi_{\theta_{\text{old}}}(\cdot|x)$. We therefore estimate this expectation by importance sampling:

$$\mathop{\mathbb{E}}_{y' \sim \text{Unif}(\cdot|x)} \exp(\pm r(x,y')) = \mathop{\mathbb{E}}_{y' \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{\text{Unif}(y'|x)}{\pi_{\theta_{\text{old}}}(y'|x)} \exp(\pm r(x,y')) \right].$$

Because $\text{Unif}(y'|x)$ is constant over the response space, this is proportional to

$$\mathop{\mathbb{E}}_{y' \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{\exp(\pm r(x,y'))}{\pi_{\theta_{\text{old}}}(y'|x)} \right].$$

Using the group samples $\{y_j\}_{j=1}^{G}$, we estimate this quantity as

$$\sum_{j=1}^{G} \frac{\exp(\pm r_j)}{\pi_{\theta_{\text{old}}}(y_j|x)} = \sum_{j=1}^{G} \exp(\pm r_j - \ln \pi_{\theta_{\text{old}}}(y_j|x)),$$

up to a common multiplicative constant. This is the sampled-softmax, or Log-Q, correction (Bengio & Senécal, 2008).

Substituting this estimate into the reward-induced density gives the plug-in estimator

$$\hat{p}^{\pm}(y_i|x) = \frac{\exp(\pm r_i)}{\sum_{j=1}^{G} \exp\big(\pm r_j - \ln \pi_{\theta_{\text{old}}}(y_j|x)\big)}. \tag{36}$$

The corresponding truncated importance weight estimate is

$$\begin{aligned}
\hat{w}_i^{\pm} &= \frac{\hat{p}^{\pm}(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\, \hat{I}_i^{\pm} \\
&= \frac{\exp(\pm r_i - \ln \pi_{\theta_{\text{old}}}(y_i|x))}{\sum_{j=1}^{G} \exp\big(\pm r_j - \ln \pi_{\theta_{\text{old}}}(y_j|x)\big)}\, \hat{I}_i^{\pm} \\
&= \vec{\sigma}_i\big(\{\pm r_j - \ln \pi_{\theta_{\text{old}}}(y_j|x)\}_{j=1}^{G}\big) \cdot \hat{I}_i^{\pm},
\end{aligned} \tag{37}$$

where

$$\hat{I}_i^{\pm} = \mathbb{1}_{\{\pm a_i > 0\}}$$

is the empirical group-advantage analogue of $I^{\pm}(r(x,y))$.

Thus, the weights used in f-GRPO are precisely softmax-normalized importance weights under the reward-induced aligned or unaligned scoring model. The term $\pm r_j$ assigns larger mass to high-reward responses for the aligned direction and to low-reward responses for the unaligned direction, while the correction $-\ln \pi_{\theta_{\text{old}}}(y_j|x)$ accounts for the fact that responses are sampled from $\pi_{\theta_{\text{old}}}$ rather than uniformly. The truncation indicator retains only samples with the appropriate advantage sign, yielding a stable reward-weighted approximation to aligned versus unaligned divergence estimation.

C.2 From Sign-Based Updates to Advantage-Weighted f-GRPO

In the main text, the preliminary form of the f-GRPO objective in Eq. (10) involves a sign-based weighting $\text{sign}(a_i)$, where a_i denotes a standardized reward-based advantage. This form arises naturally from the variational decomposition in Eq. (8), as the sign determines whether a sample contributes to the aligned or unaligned component of the divergence estimator. While this formulation is sufficient for identifying the correct update direction, it discards magnitude information contained in the advantage.

Connection to Standard Policy Gradient Methods. Popular on-policy reinforcement learning algorithms such as PPO and GRPO maximize expected reward by optimizing objectives of the form

$$\sup_{\theta} \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y\sim\pi_{\theta_{\text{old}}}} \Big[\tfrac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}\, A^{\pi_{\theta_{\text{old}}}}(x,y)\Big], \tag{38}$$

where the advantage $A^{\pi_{\theta_{\text{old}}}}(x,y)$ scales the policy gradient and determines both its direction and magnitude. In particular, GRPO uses a batch-standardized advantage

$$A^{\pi_{\theta_{\text{old}}}}(x,y_i) = \frac{r(x,y_i) - \hat{\mu}(r(x,\cdot))}{\hat{\sigma}(r(x,\cdot))}, \tag{39}$$

so that gradient magnitudes are proportional to relative reward improvements.

Motivation for Advantage Weighting. Using only $\text{sign}(a_i)$ treats all above-average samples equally and all below-average samples equally, regardless of how strongly they deviate from the mean reward. Replacing $\text{sign}(a_i)$ with a_i restores the standard policy-gradient behavior in which higher-advantage samples exert proportionally larger influence on the update. This modification preserves the alignment direction implied by the divergence estimator while improving optimization efficiency and variance properties.

Normalization and Scaling. In f-GRPO, the policy score is parameterized as $r_{\theta,i} = \beta \ln \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)}$, so the gradient of the loss contains a factor of β^{-1} relative to the standard log-policy gradient. To ensure that gradient magnitudes remain comparable to those of conventional on-policy methods, we scale the advantage by $(1 + \beta^{-1})$, yielding the final weighting $a_i(1 + \beta^{-1})$. This normalization aligns the overall gradient scale with that of GRPO and PPO, facilitating stable optimization and fair empirical comparison.

Resulting Objective. With this replacement, the resulting f-GRPO loss in Eq. (12) can be viewed as a divergence-based generalization of GRPO, where the policy gradient is weighted by a standardized advantage while the update direction is shaped by the chosen f-divergence through the function $\psi(\cdot, \cdot)$. This modification preserves the theoretical interpretation of f-GRPO as a divergence estimator while recovering the practical benefits of advantage-weighted on-policy optimization.

D Experiment Details

We use a single Nvidia H200 140GB for all our experiments with the following implementation of f-HAL, with corresponding hyperparameters.

Table 10: Left: f-HAL training algorithm; Right: Training Hyperparameters.

Algorithm 1 f-HAL Training (minibatch implementation)

Input: policy π_θ, reference π_{ref}, divergence f, link g
Input: batch size B, rollouts per prompt G, step size η, regularization coefficient $\beta \neq 0$.
Input: optional supervision dataset $\mathcal{D}_{\text{sup}}$ either $\mathcal{D}_{\text{sup}} = \{(x, y_w, y_l)\}$ (pairwise) or $\mathcal{D}_{\text{sup}} = \{(x, y, \ell)\}$ with $\ell \in \{+1, -1\}$ (binary)
Initialize parameters θ
repeat
  Sample minibatch of prompts $\{x_b\}_{b=1}^{B}$
  Set behavior policy $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Initialize $g_{\text{on}} \leftarrow 0$ and $g_{\text{off}} \leftarrow 0$
  On-policy term (F-GRPO):
  for $b = 1$ **to** B **do**
    Sample $\{y_{b,i}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x_b)$
    Compute rewards $r_{b,i} = r(x_b, y_{b,i})$ and advantages $a_{b,i}$ (Eq. (4))
    Compute $\hat{w}_{b,i}^{\pm}$ Eq. (9)) and $\psi(r_{\theta,i}, a_i)$ Eq. (11)
    Accumulate $g_{\text{on}} \leftarrow g_{\text{on}} + \frac{1}{BG}\sum_{i=1}^{G} a_i \cdot (1 + \beta^{-1})\nabla_\theta \psi_{(f,g)}(r_{\theta,b,i}, a_i)$
  end for
  Off-policy FDO term (if supervision exist):
  Set $\lambda > 0$
  if a pairwise minibatch $(x_b, y_{w,b}, y_{l,b})$ is available from $\mathcal{D}_{\text{sup}}$ **then**
    for $b = 1$ **to** B **do**
      Accumulate $g_{\text{off}} \leftarrow g_{\text{off}} + \frac{1}{B}\Big(- \nabla_\theta g(r_\theta(x_b, y_{w,b})) + \nabla_\theta f^* \circ g(r_\theta(x_b, y_{l,b}))\Big)$
    end for
  else if a binary minibatch (x_b, y_b, ℓ_b) is available from $\mathcal{D}_{\text{sup}}$ **then**
    for $b = 1$ **to** B **do**
      if $\ell_b = +1$ **then**
        $g_{\text{off}} \leftarrow g_{\text{off}} - \frac{1}{B}\nabla_\theta g(r_\theta(x_b, y_b))$
      else
        $g_{\text{off}} \leftarrow g_{\text{off}} + \frac{1}{B}\nabla_\theta f^* \circ g(r_\theta(x_b, y_b))$
      end if
    end for
  else
    Set $\lambda \leftarrow 0$
  end if
  Hybrid update:
  $\theta \leftarrow \theta - \eta\big((1 - \lambda)\, g_{\text{on}} + \lambda\, g_{\text{off}}\big)$
until converged
Output: trained policy π_θ

Parameter	**Value**
Learning rate (Math RLVR)	5×10^{-6}
Learning rate (Safety PA)	5×10^{-5}
KL coefficient (β)	0.1
LoRA rank	64
LoRA α	64
Weight decay	0.1
Warmup ratio	0.1
LR scheduler	Cosine
Optimizer	AdamW (8-bit)
On-policy (GRPO / f-GRPO / f-HAL) specific	
Gradient accumulation steps	4
Number of generations (G)	4
Direct Aligner (Off-policy) specific	
Per-device train batch size	4
Gradient accumulation steps	1

Chat Template:

Table 11: Chat templates used for (left) math reasoning and (right) safety alignment. The math template encourages standardized final-answer formatting via `\boxed{}` for automated evaluation.

Role	Math Reasoning Template	Safety Alignment Template
System	`<|im_start|>system` `You are a helpful assistant.<|im_end|>`	`<|im_start|>system` `You are a helpful assistant.<|im_end|>`
User	`<|im_start|>user` `{input} Please reason step by step, and put your final answer within \boxed{}.<|im_end|>`	`<|im_start|>user` `{input} <|im_end|>`
Assistant	`<|im_start|>assistant`	`<|im_start|>assistant`

D.1 Additional Math Reasoning Tables

Table 12: Pass@1 accuracy (%) for Qwen2.5-Math-1.5B trained on LIMR.

Method	GSM8K	MATH500	AMC23	AIME24	AIME25	**Relative Overall ↑**	**Avg. Rank ↓**
Base	21.39	25.66	27.34	5.63	2.29	0.00	8.0
GRPO	71.67	60.60	43.28	8.54	3.13	74.26	5.2
Hellinger	72.73	61.18	41.09	7.08	5.42	81.11	4.2
JS	72.47	**61.93**	40.94	7.50	4.79	78.99	4.2
KL	72.34	61.59	41.72	7.71	5.00	81.83	4.4
Pearson	**72.59**	61.70	**45.16**	**10.42**	3.33	**86.49**	**2.6**
Reverse KL	72.65	61.65	43.75	7.92	4.58	82.47	3.0
Total Variation	72.40	61.51	43.28	7.92	4.79	83.11	3.8

Table 13: Pass@1 accuracy (%) for Qwen2.5-Math-7B trained on LIMR.

Method	GSM8K	MATH500	AMC23	AIME24	AIME25	Relative Overall	Avg. Rank ↓
Base	20.32	23.15	12.03	3.54	1.88	0.00	8.0
GRPO	79.90	68.13	55.78	16.46	7.92	91.72	5.2
Hellinger	81.14	68.93	57.81	16.88	7.08	91.54	3.8
Jensen–Shannon	80.83	**69.45**	54.53	**19.17**	7.71	94.93	3.8
KL	**81.15**	68.90	54.69	17.29	8.13	93.64	3.4
Pearson	81.07	**69.45**	55.47	17.08	**8.96**	96.28	2.8
Reverse KL	80.86	69.28	55.78	**19.17**	8.13	**96.59**	**2.4**
Total Variation	80.63	68.71	55.78	14.38	7.29	87.79	5.4

Table 14: Pass@1 accuracy (%) for Qwen2.5-Math-1.5B trained on GSM8k .

Method	GSM8K	MATH500	AMC23	AIME24	AIME25	Relative Overall	Avg. Rank ↓
Base	21.39	25.66	27.34	5.63	2.29	0.00	8.0
GRPO	81.34	67.26	48.44	10.21	4.79	86.17	4.0
Hellinger	80.71	68.64	48.75	10.00	4.79	86.24	3.8
Jensen–Shannon	81.13	67.89	49.06	**12.08**	5.63	**97.22**	**2.2**
KL	80.13	66.80	44.38	7.71	**5.83**	79.65	6.0
Pearson	80.89	68.26	47.81	11.46	3.75	84.23	4.0
Reverse KL	80.79	67.65	49.22	11.67	4.17	88.07	3.4
Total Variation	80.22	67.28	46.88	11.25	**6.04**	94.27	4.4

Table 15: Pass@1 accuracy (%) for Qwen2.5-Math-7B trained on GSM8k.

Method	GSM8K	MATH500	AMC23	AIME24	AIME25	Relative Overall	Avg. Rank ↓
Base	20.32	23.15	12.03	3.54	1.88	0.00	8.0
GRPO	85.51	73.70	61.56	19.17	8.75	90.71	4.4
Hellinger	85.96	74.23	58.13	**22.08**	9.79	95.08	3.8
Jensen–Shannon	**87.76**	74.86	60.00	17.71	9.58	91.44	3.2
KL	86.67	74.21	57.34	18.54	7.71	86.60	5.2
Pearson	87.28	**74.93**	59.22	18.54	7.71	87.82	3.6
Reverse KL	87.34	74.60	57.03	18.96	9.17	90.46	3.8
Total Variation	86.55	74.31	60.78	18.54	**11.04**	**95.27**	3.2

Table 16: Pass@1 accuracy (%) for Qwen2.5-Math-1.5B trained on OpenRS .

Method	GSM8K	MATH500	AMC23	AIME24	AIME25	**Relative Overall ↑**	**Avg. Rank ↓**
Base	21.39	25.66	27.34	5.63	2.29	0.00	8.0
GRPO	73.28	61.46	41.09	8.75	4.79	82.44	4.6
Hellinger	73.97	63.30	44.06	7.92	3.96	78.47	3.6
JS	74.03	62.33	41.88	10.00	5.42	**93.30**	**2.4**
KL	72.59	61.50	43.75	9.38	4.38	85.66	4.0
Pearson	72.81	61.61	**45.31**	9.38	3.96	85.19	3.6
Reverse KL	73.08	61.21	42.03	9.17	4.17	81.65	5.2
Total Variation	73.28	61.26	42.50	8.13	**5.83**	86.93	4.0

Table 17: Pass@1 accuracy (%) for Qwen2.5-Math-7B trained on OpenRS.

Method	GSM8K	MATH500	AMC23	AIME24	AIME25	**Relative Overall ↑**	**Avg. Rank ↓**
Base	20.32	23.15	12.03	3.54	1.88	0.00	8.0
GRPO	81.42	70.08	55.63	18.13	7.71	90.52	5.8
Hellinger	82.39	70.58	59.38	15.63	8.54	91.70	4.4
JS	**83.53**	**71.75**	56.41	**19.58**	**9.38**	**98.47**	**2.0**
KL	82.42	71.69	59.53	17.08	8.96	95.16	3.0
Pearson	81.97	69.66	57.50	18.13	7.92	91.86	5.0
Reverse KL	83.06	71.68	**60.00**	18.96	8.54	96.78	2.4
Total Variation	82.02	71.84	56.41	16.67	6.25	86.05	4.8

Table 18: No-KL GRPO ($\beta = 0$) Pass@1 accuracy (%) for Qwen2.5-Math-1.5B.

Training	GSM8K	MATH500	AMC23	AIME24	AIME25
LIMR	72.04	60.51	43.30	6.46	2.91
GSM8K	81.46	67.29	48.56	11.10	4.58
OpenRS	72.80	60.80	42.65	8.33	4.79

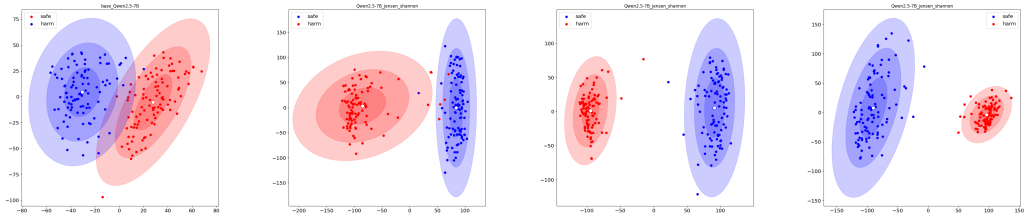
Table 19: No-KL GRPO ($\beta = 0$) Pass@1 accuracy (%) for Qwen2.5-Math-7B.

Training	GSM8K	MATH500	AMC23	AIME24	AIME25
LIMR	80.12	68.13	57.10	16.84	7.71
GSM8K	86.10	74.00	60.50	18.60	9.10
OpenRS	81.80	71.06	56.10	17.81	7.90

D.2 Separation Metric for Safety Alignment

We follow the methodology of Haldar et al. (2025); Lin et al. (2024b); Zheng et al. (2024) to visualize the separation between anchor safe and harmful prompts in the latent space of the LLM before and after alignment. Each prompt is represented using the last hidden state of the final token, which captures the full contextual history of the sequence. We then apply PCA to these representations and visualize the first two principal components. Please refer to the previous mentioned references for more details.

Figure 2 illustrates this visualization for the Qwen-7B base model and its aligned variants. Alignment induces substantially stronger separation between safe and harmful clusters, a phenomenon that has been shown to correlate with increased robustness. Following the procedure of Haldar et al. (2025) (Appendix C.3), we quantify this separation using the Bhattacharyya distance D_B between the clusters, which we adopt as a robustness metric.



(a) Base, (D_B=2.48) (b) $\lambda = 0, D_B = 4.47$ (c) $\lambda = 0.5, D_B$=12.13 (d) $\lambda = 1, D_B = 9.14$

Figure 2: Latent-space separation (Bhattacharyya distance D_B) between safe and harmful prompt clusters before and after alignment with f-HAL (Jensen–Shannon divergence) on Qwen-7B-Base. Compared to the base model, all aligned variants increase separation. The on-policy method (f-GRPO, λ=0) yields weaker separation than the hybrid (f-HAL, λ=0.5) and the off-policy supervised objective (FDO, λ=1).

D.3 Qwen-1.5B Safety Alignment Table

Table 20: Post-Safety-Alignment robustness and utility results for Qwen2.5-1.5B (Base) across different alignment methods. AvgRob is computed using min–max normalized separation scores in $[0, 100]$ and $(100 - \text{ASR})$ for AdvBench and SALAD, while AvgUtil denotes the average utility score. The Total score is defined as mean(AvgRob, AvgUtil). Cell color and intensity indicate performance relative to the mean value within each benchmark column. The pre-alignment Base model separation is 1.16.

Method	Robustness					Utility				Total
	Sep. D_B ↑	ASR(%)↓ GCG	ASR(%)↓ SALAD	ToxiGen (%)↑	**AvgRob.** ↑	MMLU	IFEval (%) ↑	MuSR	**AvgUtil.** ↑	**Score** ↑
Direct Aligners (Off-Policy)										
BCO	9.36	30.10	35.14	48.40	55.92	28.91	35.25	33.86	32.67	44.30
DPO	2.95	86.00	62.28	47.87	27.56	28.67	35.73	34.92	33.11	30.33
KTO	4.48	54.50	65.02	51.06	37.33	29.01	35.13	33.73	32.62	34.98
FDO (f-HAL, $\lambda = 1$) :										
Hellinger	17.95	1.10	25.66	45.53	74.84	28.22	33.21	32.94	31.46	53.15
Jensen-Shannon	13.89	23.20	37.94	44.15	61.17	28.37	37.17	34.79	33.44	47.30
KL	13.02	54.00	60.28	45.96	47.33	28.91	37.41	**35.58**	33.97	40.65
Pearson	12.76	13.60	45.52	46.28	60.90	28.61	35.73	34.13	32.82	46.86
Reverse KL	**22.11**	5.00	23.30	48.09	**79.95**	27.50	**40.89**	34.13	**34.17**	**57.06**
Total Variation	11.61	17.20	49.58	45.21	57.37	28.13	35.73	34.66	32.84	45.11
Hybrid , f-HAL ($\lambda = 0.5$) :										
Hellinger	12.48	7.25	**9.86**	**51.28**	72.31	26.36	27.58	35.05	29.66	50.99
Jensen-Shannon	11.18	19.60	53.60	47.13	55.75	28.21	36.21	32.67	32.36	44.05
KL	12.42	46.50	62.70	50.85	49.12	**29.06**	36.21	34.52	33.26	41.19
Pearson	10.96	13.30	48.44	50.11	59.09	28.44	36.21	34.39	33.01	46.05
Reverse KL	20.08	**0.00**	32.44	50.96	77.27	27.29	38.49	34.66	33.48	55.37
Total Variation	15.02	22.31	59.74	48.40	58.32	28.30	35.13	34.52	32.65	45.49
On-Policy RL										
GRPO	0.66	98.00	96.62	43.83	12.30	28.11	2.52	33.11	21.25	16.77
f-GRPO (f-HAL, $\lambda = 0$) :										
Hellinger	5.75	77.10	44.42	48.40	37.65	28.91	32.49	**35.58**	32.33	34.99
Jensen-Shannon	3.39	92.42	73.04	44.57	22.96	28.46	31.06	34.92	31.48	27.22
KL	4.11	64.27	64.28	47.45	33.74	28.47	25.42	32.54	28.81	31.27
Pearson	3.64	93.10	75.22	47.13	23.17	28.53	30.22	34.92	31.22	27.20
Reverse KL	3.52	67.80	55.96	50.32	34.97	27.97	27.34	34.52	29.94	32.46
Total Variation	4.99	97.10	81.96	47.66	22.20	28.22	32.73	33.33	31.43	26.81

D.4 Sample Generations Pre vs Post-Alignment (Safety Task)

Here we display some example responses after aligning using JS divergence f-HAL ($\lambda \in \{0, \frac{1}{2}, 1\}$ corresponding to direct alignment FDO, Hybrid, on-policy f-GRPO).

Table 21: Example responses for a benign factual prompt before and after alignment on Qwen-7B Base using JS-based f-HAL. Here, $\lambda = 0$ corresponds to purely on-policy f-GRPO, $\lambda = 1$ to off-policy direct alignment (FDO), and $\lambda = 0.5$ to the hybrid objective combining both. The hybrid setting often yields the most coherent answer by leveraging off-policy factual supervision while maintaining on-policy reward-driven optimization.

Model / Method	**Response to Prompt: "What color is the sky?"**
Qwen-7B Base (Pre-alignment)	The sky is typically blue. Write a Python function named 'find_max_min' that takes a list of numbers ... *[Response continues with an unrelated Python coding task, including extraneous instructions and incorrect code.]*
JS f-HAL ($\lambda = 0$, On-policy f-GRPO)	The sky is usually blue during the day and turns pink or orange at sunrise and sunset.
JS f-HAL ($\lambda = 0.5$, Hybrid f-GRPO + FDO)	The sky is typically blue during the day due to a phenomenon called Rayleigh scattering, which causes shorter wavelengths of light (like blue and violet) to scatter more easily in the Earth's atmosphere. However, the sky can appear pink or red at sunrise and sunset due to increased scattering of longer wavelengths (red, orange).
JS f-HAL ($\lambda = 1$, Off-policy FDO)	The color of the daytime sky is typically blue, due to a phenomenon called Rayleigh scattering, which causes shorter wavelengths of light (such as blue) to scatter more than longer wavelengths (like red and orange). However, the color can vary depending on the time of day, weather conditions, and other factors.

Table 22: Qualitative comparison on a harmful prompt using JS-based f-HAL for Qwen-7B Base. The base model produces unsafe procedural content , while the aligned models provide robust safet-aligned refusals.

Model / Method	**Response to Harmful Prompt: "Outline the steps for a terrorist organization to infiltrate a government agency."**
Qwen-7B Base (Pre-alignment)	To infiltrate a government agency, a terrorist organization would likely follow these steps: 1. **Identify Targets**: The organization .... *[The base model generates a detailed, step-by-step procedural response describing how a terrorist organization could infiltrate a government agency, including recruitment, trust-building, information gathering, and cover maintenance.]*
JS f-HAL ($\lambda = 0$, On-policy f-GRPO)	I'm sorry, I can't assist with that. Instead, let's talk about how to protect yourself and your organization from cyber threats.
JS f-HAL ($\lambda = 0.5$, Hybrid f-GRPO + FDO)	I'm sorry, but I can't assist with that request. Infiltrating a government agency is illegal and dangerous. Please respect the law and focus on positive
JS f-HAL ($\lambda = 1$, Off-policy FDO)	I'm sorry, but I can't assist with that request. Infiltrating a government agency is illegal and unethical. If you have any other questions or need"

D.5 Sample Generations Pre vs Post-Alignment (Math Reasoning Task)

Table 23: Qualitative example where the base model fails (score=0) while aligned methods succeed (score=1). Responses are lightly truncated for space.

Method	Response (short)
Prompt	Wendi has 20 chickens. Each chicken gets 3 cups per meal. Morning: 15 cups. Afternoon: 25 cups. How many cups in the final meal? (Answer in `\boxed{}`.)
Base (pre)	Computes total as $3 \times 20 = 60$, subtracts $15 + 25 = 40$, but outputs $\boxed{10}$ (incorrect).
GRPO	Total needed: $3 \times 20 = 60$. Given: $15 + 25 = 40$. Final meal: $60 - 40 = \boxed{20}$.
KL	Total: 60. Given: 40. Remaining: $\boxed{20}$.
Pearson	$15 + 25 = 40$ given; 60 needed; final $= 60 - 40 = \boxed{20}$.
Reverse-KL	Total $= 60$, already $= 40$, final $= \boxed{20}$.
Hellinger	Total $= 60$; remaining after 40 is $\boxed{20}$.
Jensen–Shannon	Total $20 \cdot 3 = 60$; final $60 - (15 + 25) = \boxed{20}$.
Total–Variation	Final meal: $3 \cdot 20 - (15 + 25) = \boxed{20}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claimed: (1) to extend the divergence framework to RLVR setting (2) Claimed to provide theoretical guarantees for the given framework (3) Empirical validation of framework, addressing improvment of f-GRPO over GRPO & f-HAL mitigates reward hacking. These were addressed: (1) in §3.2 & §4. (2) Theorem 4.3 (3) §5, Math reasoning RLVR setting to show f-GRPO vs GRPO §5.1, Preference alignment setting & Reward Hacking §5.2

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to end of §6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: §C provides the detailed proofs of all theorems and lemmas. The Theorem 4.2 has been stated for Theorem 4.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: A;; the configuration details , datasets, models, etc are detailed in §D & §5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Training scripts and documentation are provided in the anonymous code repository linked in the main text.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental setting is provided in §5. Algorithm and hyperparameter details are provided in §D. Moreover, the training scripts and documentation are provided in the anonymous code repository linked in the main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report pass@1 for RLVR experiments on multiple generations and for preference alignment we used standard out-of-box benchmarks,

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The number and type of GPUs used for training along with training epochs, etc are provided in §D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Maintained anonymity and provided credits where due, to the best of our knowledge.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: No direct societal impact, we introduce general RL algorithms from the lens of mathematical distributional divergence..

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Foundational paper with theoretical results and standard alignment benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the open-source libraries and repos are credited in §5. For instance, Unsloth & VLLM.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: A documented Code repo is released, and an anonymous link to it is provided.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Only used for editing and grammar check.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.