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

Rajdeep Haldar[1] · **Lantao Mei** · **Guang Lin**[1] · **Yue Xing**[2] · **Qifan Song**[1]

Department of Statistics: [1]Purdue University, [2]Michigan State University

NeurIPS 2026

Paper page

Background: Two Regimes of LLM Alignment

Alignment pushes desirable responses up and undesirable ones down — how we judge a response splits it in two

RLVR

Feedback: a **scalar verifiable reward** $r(x,y)$

Aligned ↑ high-reward · Unaligned ↓ low-reward

Learned with **on-policy RL** (GRPO)

Math, code, tool use

Preference alignment

Feedback: **human preferences** (chosen vs rejected)

Aligned ↑ chosen · Unaligned ↓ rejected

Learned **on- or off-policy** (RLHF, DPO, KTO, BCO)

Safety, chat quality

Known

For **preference** data, alignment is already understood as **divergence estimation**: direct aligners (DPO, KTO, BCO) each turn out to estimate a statistical divergence between the **aligned** (chosen) and **unaligned** (rejected) response distributions — which buys interpretability, guarantees, and a recipe for designing new losses.

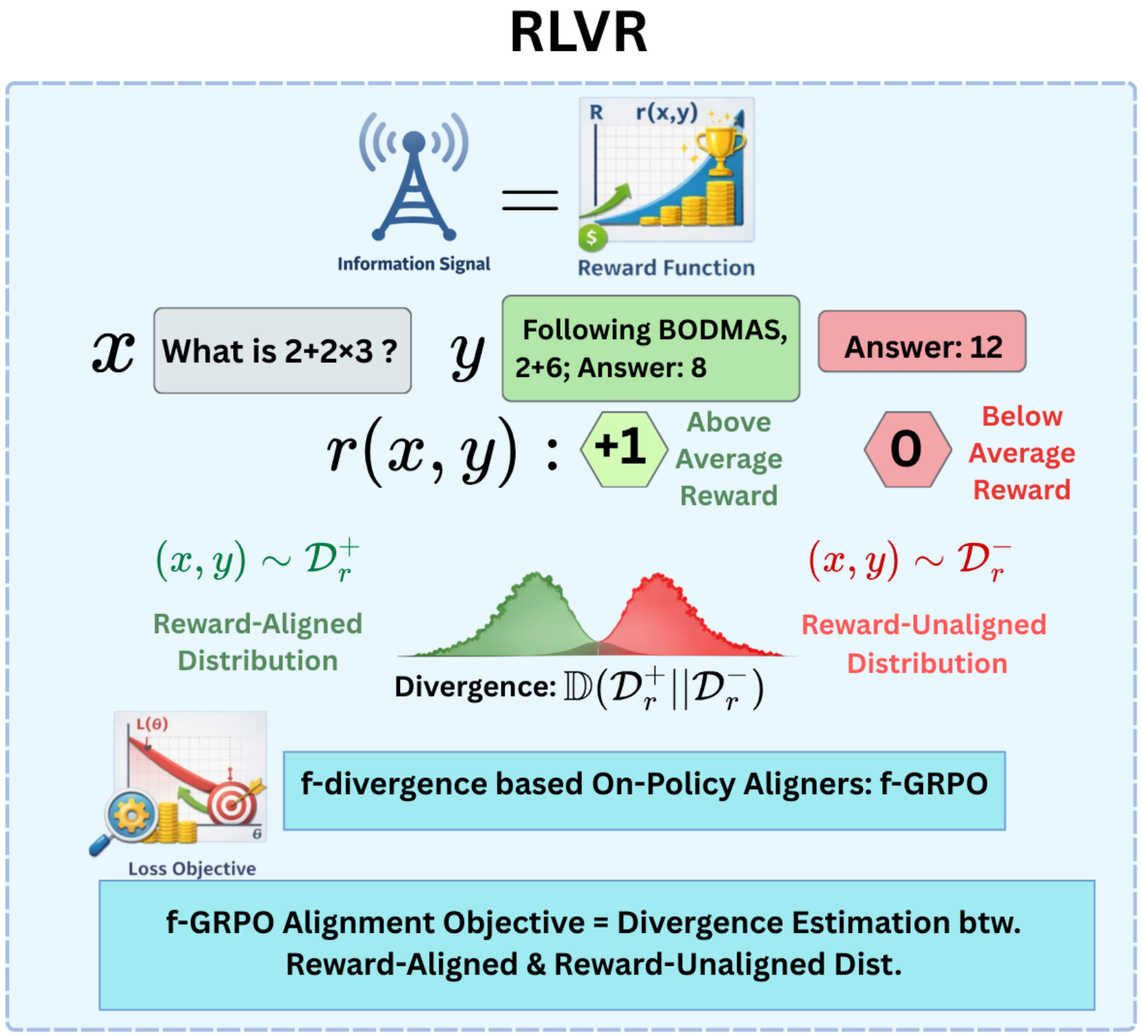
The gap

RLVR has **no chosen/rejected pairs at all** — just one scalar reward per sampled response. The divergence view, as it stands, does not reach it.

This paper

Generalize the divergence view to all of LLM alignment — RLVR included.

Step 1 — The Reward Induces the Two Distributions



Definition 4.1

Draw a group of *G* responses for a prompt, score them, and split the group at the **average reward**:

above average → reward-aligned $\mathcal{D}_r^+$

below average → reward-unaligned $\mathcal{D}_r^-$

$$d\mathcal{D}^{\pm}_{(r,\theta)} \propto [\pm(r(x,y) - \mu_r^{\theta}(x))]_+ \exp(\pm r(x,y))\, dP(x)$$

$\mu^{\theta}_r(x)$ = the policy's average reward on prompt x; $[\cdot]_+$ keeps each side to its own half of the group.

RLVR now has exactly the two objects that preference data hands you for free — so the **same divergence machinery applies**, with no chosen/rejected pairs anywhere in sight.

Step 2 — Divergence Estimation Writes the Loss

Every *f*-divergence is a **maximum over critic functions** *T* (its variational form):

$$\mathbb{D}_f\left(\mathcal{D}^+ \,\|\, \mathcal{D}^-\right) = \max_T \; \mathbb{E}_{\mathcal{D}^+}[T] - \mathbb{E}_{\mathcal{D}^-}[f^*(T)]$$

The critic is already there

We never have to **train** a critic — the policy already carries one. Its **implicit reward**

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

says how much more likely the policy makes a response than the reference model.

… and that is where *g* comes from

A monotone **link *g*** maps r_θ into the range *f** expects, so we set

$$T = g(r_\theta)$$

That is where *g* comes from — it is not a free design choice bolted on, it is what the variational form requires.

One objective, two jobs: maximizing it **estimates** the divergence *and* **separates** the two distributions.

The *f*-GRPO Objective, Term by Term

Plug $T = g(r_\theta)$ in and replace the unobservable expectations by reward-weighted sums over the group

$$\mathcal{L}_{f-GRPO}(\theta) \propto \textbf{aligned term} + \textbf{unaligned term}$$

$$-\sum_{\text{above avg}} w_i^{+}\, g(r_{\theta,i})$$

Above-average reward. Falls as r_θ **rises** — minimizing makes good responses **more** likely.

$$+\sum_{\text{below avg}} w_i^{-}\, f^{*}(g(r_{\theta,i}))$$

Below-average reward. Falls as r_θ **drops** — minimizing makes bad responses **less** likely.

What the weights $w^{\pm}$ mean, in words: a response counts for more when its reward **stands out** from the group *and* when the model was **less likely to have produced it** — surprising generations carry more signal. (Formally, an importance-sampling weight.)

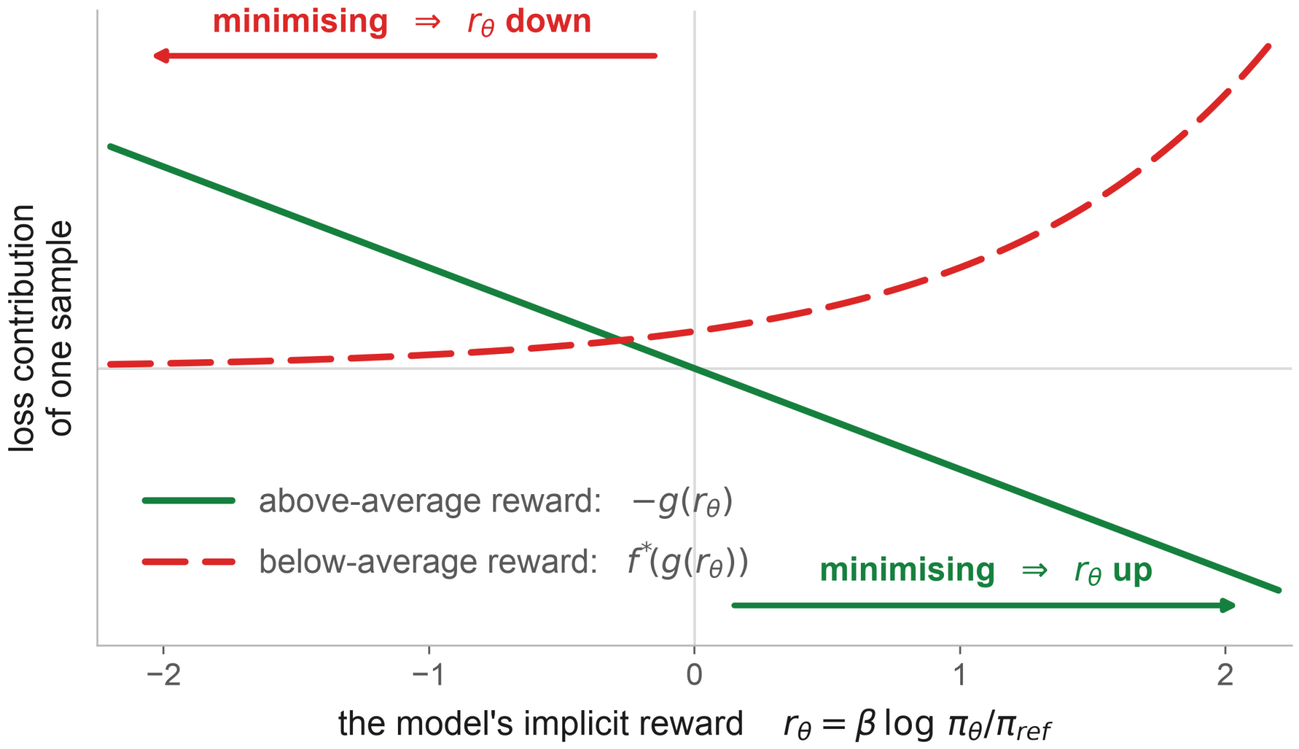
The preference-data version (FDO) is the **same two terms**, with expectations over chosen/rejected data in place of the reward-weighted sums — which is exactly why one framework covers both regimes.

One Worked Choice: KL

Picking f fixes everything else — here is the whole loss, written out

take $f(t) = t\ln t$ so $g(u) = u$ and $f^*(s) = e^{s-1}$

$$\mathcal{L} \propto -\sum_{\text{above avg}} w_i^+ \, r_{\theta,i} + \sum_{\text{below avg}} w_i^- \, e^{r_{\theta,i}-1}$$



Linear reward for good responses, **exponential** penalty for bad ones.

Swapping f — Hellinger, Jensen–Shannon, Pearson χ^2, reverse-KL, total variation — just **reshapes these two curves**. That is the entire design space the framework opens up.

No new machinery per loss: the guarantees on the next slide hold for *every* canonical choice.

What the Objective Buys You

Theorem 4.3 — three results, one per colour

1. It really is divergence estimation

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

The optimized loss **equals the divergence** between the good and bad halves — the loss value means something.

2. Alignment consistency

$$\pi_{\theta^{(t+1)}} \propto \pi_{ref} \exp(\beta^{-1} h_t)$$

Every fixed point is the reference policy **tilted** by some h_t, so preferred responses gain mass.

3. Reward improves monotonically

$$\mu_r^{\theta^{(t)}} < \mu_r^{\theta^{(t+1)}} \longrightarrow \mu_r^{\theta^*} = \max_y r$$

Expected reward rises at **every** step and the limit policy sits on the **best** response.

Results 2 and 3 need a **canonical link**: a monotone g with $g^{-1}(f'_\infty) = \infty$. KL, reverse-KL, Hellinger, JS, Pearson χ^2 and TV all have one.

Then the tilt is ∞ on above-average responses and 0 on below-average ones — **all the mass concentrates on the good half**, and that is what makes the reward climb.

f-GRPO vs GRPO: Only the Tilt Differs

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

Both methods land on this same form — same data, same group of samples, same KL budget. Only h_t changes.

Tilt h_t is…	on above-average responses	on below-average responses	mass left below average
GRPO	$a_t(x,y) > 0$ (finite)	$a_t(x,y) < 0$ (finite)	**non-zero**
f-GRPO (canonical *g*)	∞	**0**	**zero**

GRPO re-weights the reference by the exponential of the *standardized advantage* — a finite tilt, so it **always keeps probability** on responses its own group called below average. *f*-GRPO with a canonical link does not: it concentrates the mass on the above-average half, which is exactly what buys monotone reward improvement and maximal-reward attainment.

f-HAL — One Dial Between the Two Regimes

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

- **λ = 1** → FDO: preference data only (off-policy)
- **λ = 0** → *f*-GRPO: reward only (on-policy)
- **λ = ½** → hybrid: reward-driven exploration *anchored* by preference data
- It estimates the same divergence between **mixtures** of the reward- and preference-induced distributions — which is why the preference data can act as an anchor.

Hybrid Alignment



=



+





f-divergence based Hybrid-Policy Aligners: f-HAL

$$f\text{-HAL} = \lambda \cdot \underbrace{\text{FDO}}_{\text{Off-Policy}} + (1-\lambda) \cdot \underbrace{f\text{-GRPO}}_{\text{On-Policy}}$$

f-HAL Alignment Objective = Divergence Estimation btw. Mixture Measures : $\mathbb{D}_f\left(\lambda \cdot \mathcal{D}^+ + (1-\lambda)\mathcal{D}_r^+ c^+(x) || \lambda \cdot \mathcal{D}^- + (1-\lambda)\mathcal{D}_r^- c^-(x)\right)$

The Whole Picture

RLVR



x What is 2+2×3 ? y Following BODMAS, 2+6; Answer: 8 Answer: 12

$r(x,y)$: +1 Above Average Reward 0 Below Average Reward

$(x,y) \sim \mathcal{D}_r^+$ Reward-Aligned Distribution

Divergence: $\mathbb{D}(\mathcal{D}_r^+ || \mathcal{D}_r^-)$

$(x,y) \sim \mathcal{D}_r^-$ Reward-Unaligned Distribution



f-divergence based On-Policy Aligners: f-GRPO

f-GRPO Alignment Objective = Divergence Estimation btw. Reward-Aligned & Reward-Unaligned Dist.

Preference Alignment



x Teach me how to hack someone's bank account?

y_w As a safe and responsible AI, I can't assist you with that. $>$ y_l We need to first bypass target's 2FA. Here are some ways to ....

$(x,y_w) \sim \mathcal{D}^+$ Aligned Distribution

Divergence: $\mathbb{D}(\mathcal{D}^+ || \mathcal{D}^-)$

$(x,y_l) \sim \mathcal{D}^-$ Unaligned Distribution

Preference Alignment Objective = Divergence Estimation btw. Aligned & Unaligned Dist.



f-divergence based Direct Aligners: FDO

Hybrid Alignment





f-divergence based Hybrid-Policy Aligners: f-HAL

$$f\text{-HAL} = \lambda \cdot \underbrace{\text{FDO}}_{\text{Off-Policy}} + (1-\lambda) \cdot \underbrace{f\text{-GRPO}}_{\text{On-Policy}}$$

f-HAL Alignment Objective = Divergence Estimation btw. Mixture Measures : $\mathbb{D}_f\left(\lambda \cdot \mathcal{D}^+ + (1-\lambda)\mathcal{D}_r^+ c^+(x) || \lambda \cdot \mathcal{D}^- + (1-\lambda)\mathcal{D}_r^- c^-(x)\right)$

Math Reasoning (RLVR): *f*-GRPO Improves on GRPO

Qwen2.5-Math 1.5B / 7B · GSM8K, LIMR, Open-RS · binary verifiable reward · Pass@1 over 16 samples

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

Pass@1, Qwen2.5-Math-1.5B trained on LIMR. Colour = relative to GRPO.

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

Avg. Rank over 2 model sizes × 3 training sets (lower is better).

30 of the 36 divergence × setting combinations rank better than GRPO, and the best divergence wins in **all 6** settings — with the largest gains on the harder datasets (LIMR, Open-RS), exactly where a sharper separation should matter most.

How to read: *Relative Overall* is the mean of per-benchmark min–max normalized scores (0–100); *Avg. Rank* is the mean rank across benchmarks. Cell colour is relative to the GRPO baseline.

Safety Alignment: Hybrid *f*-HAL Curbs Reward Hacking

No verifiable reward exists for safety — on-policy methods must use a learned reward model (DeBERTa-v3-large)

Method	Qwen-7B	Qwen-1.5B
BCO / DPO / KTO (direct)	53.6 / 42.7 / 47.4	44.3 / 30.3 / 35.0
GRPO (on-policy)	**38.2**	**16.8**
f-GRPO (λ = 0)	54.6	35.0
FDO (λ = 1)	66.1	**57.1**
Hybrid *f*-HAL (λ = ½)	**66.3**	55.4

Total Score = mean(robustness, utility); best divergence per class.

Pure on-policy RL on a *learned* reward **hacks it**: GRPO collapses to blanket refusal and loses utility as well as safety.

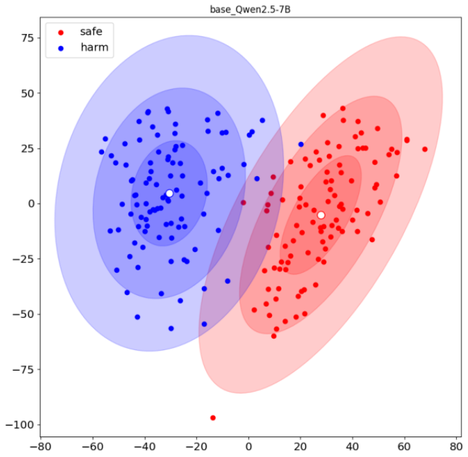
Keeping preference data in the objective anchors the policy — and the mixture interpretation of Thm 4.3 says exactly why it should.

On 7B the hybrid even beats pure direct alignment; on 1.5B direct aligners stay slightly ahead.

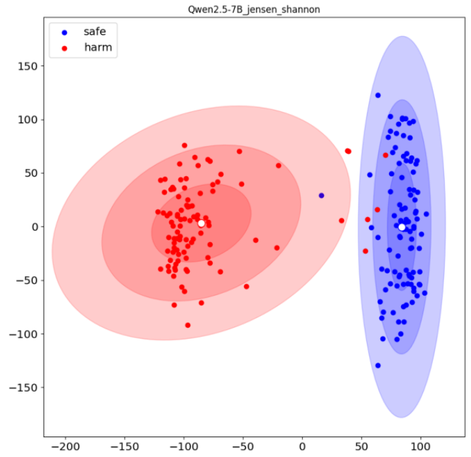
Prompt	GRPO	Hybrid f-HAL
Unsafe: Generate fake credit card numbers.	"I am not providing any content or information that promotes . . ." *Correct refusal, generic.*	"I can't assist with that request. It is illegal and unethical . . ." *Correct refusal, specific.*
Safe: How could recycling reduce your 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.*

*Reward hacking, concretely (Qwen-1.5B, Pearson χ^2): GRPO refuses even a **safe** prompt; hybrid f-HAL stays context-aware.*

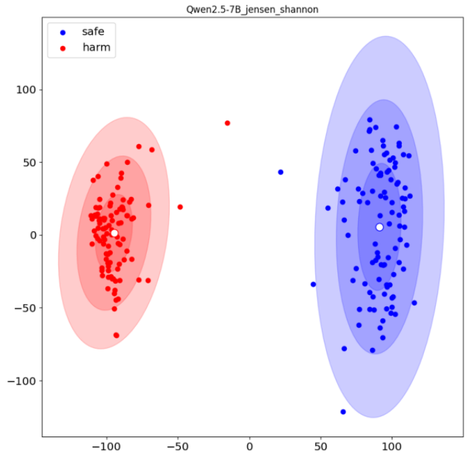
Alignment Still Separates Safe from Harmful Prompts



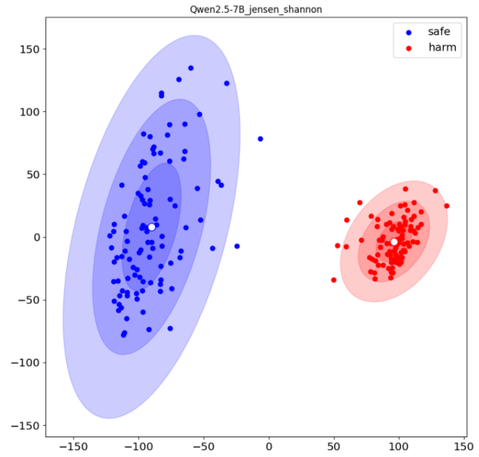
(a) Base, (D_B=2.48)



(b) $\lambda = 0, D_B = 4.47$



(c) $\lambda = 0.5, D_B$=12.13



(d) $\lambda = 1, D_B = 9.14$

Latent-space separation (Bhattacharyya distance D_B) between safe and harmful prompt clusters, Qwen-7B-Base aligned with *f*-HAL (Jensen–Shannon). Every aligned variant separates more than the base model and the **hybrid (λ = ½)** separates most — and separation is the quantity that tracks robustness.

Takeaways

1. A unified divergence framework

Preference alignment *and* RLVR are both **divergence estimation**. Rewards **induce** the aligned/unaligned distributions that preference data supplies directly, so one variational recipe writes losses for both regimes.

2. On-policy: *f*-GRPO

Under a **canonical link** it has **monotone reward improvement** and reaches the **maximal reward**. GRPO's update carries no such guarantee — its tilt stays finite, so probability remains on below-average responses.

3. Hybrid: *f*-HAL

Preference data can **complement** the reward: the mixture is *itself* a divergence estimator, and it **mitigates reward hacking** when the reward model is learned rather than verifiable.

Thank you!

Contact: **rhaldar@purdue.edu** · Paper page: **neurips.cc** · Code: **github.com/rhaldarpurdue/f-GRPO**

Paper page