A Regret Minimization Framework on Preference Learning in Large Language Models

Suhwan Kim [1*] **Taehyun Cho** [1*†] Geon-Hyeong Kim [2] Yu Jin Kim [2] Youngsoo Jang [3‡] Moontae Lee [2‡] Jungwoo Lee [1,4‡]

[1]Seoul National University [2]LG AI Research [3]UNIST [4]HodooAI Labs

*Equal contribution †Internship at LG AI Research ‡Equal corresponding authorship

ICML International Conference On Machine Learning

Spotlight Paper

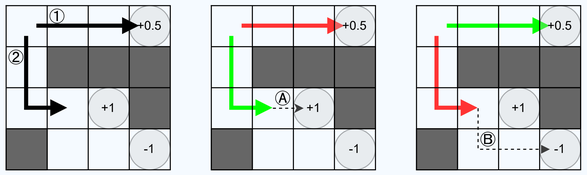
TL;DR

Modeling preferences as *regret* aligns better with human judgment than reward maximization.

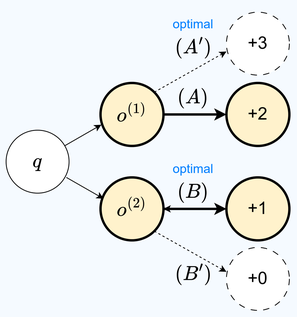
Humans judge actions by relative suboptimality — anticipating outcomes and weighing counterfactuals. We reframe RLHF as **regret minimization**: a preference is a behavior-conditioned judgment of relative suboptimality.

Two misalignments of reward maximization

1. Prospective judgment. Humans evaluate partial trajectories by imagining how they unfold. A segment with *no realized reward yet* receives no preference signal under reward maximization — even when its likely continuation would clearly justify one.



2. Counterfactual thinking. Humans also compare against plausible alternatives. Action A with reward $+2$ is *worse* than B with reward $+1$ if A's alternatives reach $+3$ while B's reach $+0$. Reward maximization picks A; humans pick B.



Reward Maximization	Regret Minimization
👍	👎
👎	👍
Which action resulted in a higher realized reward?	Which action was closer to the optimal policy?

Regret-based Preference Optimization

We model preferences through **regret** — the gap between an action and the optimal behavior:

$$\mathrm{Reg}^{\mu}_{\pi^*}(q_{<t}, o_t) \;:=\; V^{\pi^*}(q_{<t}) \;-\; Q^{\mu}(q_{<t}, o_t).$$

RePO (**Re**gret-based **P**reference **O**ptimization) uses negative regret as the preference score — capturing prospective and counterfactual judgment by construction.

Theory — closed-form regret decomposition

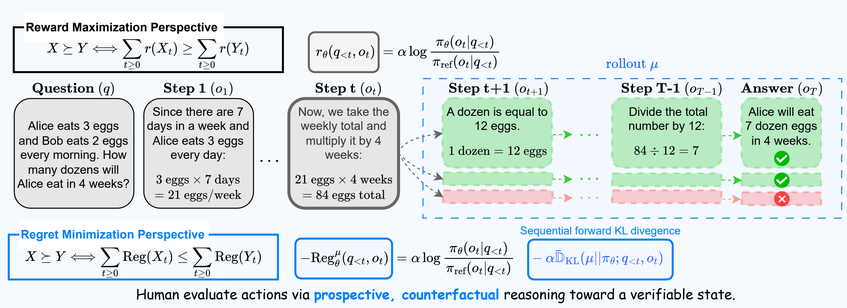
Main result. For any behavior policy μ,

$$\mathrm{Reg}^{\mu}_{\pi^*}(q_{<t}, o_t) \;=\; -\alpha\Bigg(\underbrace{\log\frac{\pi^*(o_t \mid q_{<t})}{\pi_{\mathrm{ref}}(o_t \mid q_{<t})}}_{\text{DPO log-ratio}} \;-\; \underbrace{\overline{\mathbb{D}}_{\mathrm{KL}}(\mu \,\|\, \pi^*;\, q_{<t}, o_t)}_{\text{sequential KL}}\Bigg)$$

Sequential forward KL. Discounted KL between behavior μ and optimal π^* accumulated along *future* rollouts under μ:

$$\overline{\mathbb{D}}_{\mathrm{KL}}(\mu \,\|\, \pi^*;\, q_{<t}, o_t) \;:=\; \mathbb{E}_{\tau\sim\mathbb{P}^{\mu}}\left[\sum_{l>0}\gamma^{l}\,\mathbb{D}_{\mathrm{KL}}\big(\mu(\cdot|q_{<t+l}) \,\|\, \pi^*(\cdot|q_{<t+l})\big)\right].$$

Connection to DPO. When the behavior policy matches the model ($\mu = \pi_\theta$), the sequential-KL term vanishes and the objective reduces to **DPO**. Off-policy or heterogeneous data breaks this match; **RePO** corrects for it.



Human evaluate actions via **prospective, counterfactual** reasoning toward a verifiable state.

Practical loss — one term on top of DPO

RePO (token-level μ log-probs recorded):

$$S^{\mathrm{RePO}} \;=\; -\alpha\left(\log\frac{\pi_\theta(o_t|q_{<t})}{\pi_{\mathrm{ref}}(o_t|q_{<t})} \;+\; \frac{1}{T-t}\sum_{1\leq\ell\leq T-t}\log\frac{\pi_\theta(o_{t+\ell}|q_{<t+\ell})}{\mu(o_{t+\ell}|q_{<t+\ell})}\right)$$

RePO_det (μ unknown; deterministic pseudo-label $\mu(\cdot|q) = \delta_{o_t}$):

$$S^{\mathrm{RePO\_det}} \;=\; -\alpha\left(\log\frac{\pi_\theta(o_t|q_{<t})}{\pi_{\mathrm{ref}}(o_t|q_{<t})} \;+\; \frac{1}{T-t}\sum_{1\leq\ell\leq T-t}\log\pi_\theta(o_{t+\ell}|q_{<t+\ell})\right)$$

*Same training loop as **DPO** — one extra term in the loss.*

Why regret helps — inductive bias

Mild assumption. (μ closer to π^* than to π_{ref})

$$\epsilon := \mathbb{D}_{\mathrm{KL}}(\mu \,\|\, \pi_{\mathrm{ref}}; q_{<t}) - \mathbb{D}_{\mathrm{KL}}(\mu \,\|\, \pi^*; q_{<t}) \;\geq\; 0$$

Consequence. For any verifier-accepted terminal $o^\star_T$,

$$\widehat{\mathrm{Reg}}^{\mu}_{\pi^*}(q_{<T},\, o^\star_T) \;\leq\; \mathbb{E}_{\mu}\left[\widehat{\mathrm{Reg}}^{\mu}_{\pi^*}(q_{<t},\, \cdot)\right].$$

Partial contexts are judged **more harshly** than fully revealed successes. **DPO** learns this only via masked-data augmentation, while **RePO** inherits it directly from the regret decomposition.

Experiments — human preference + math reasoning

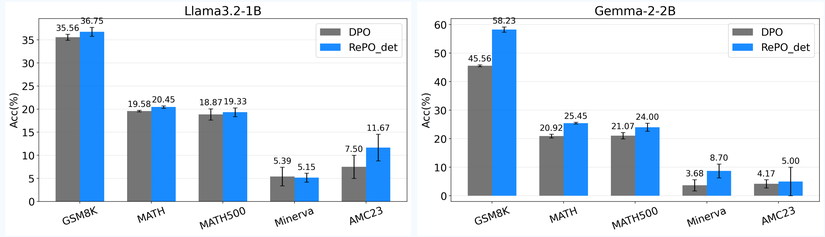
Human Preference

Method	AlpacaEval 2 LC	AlpacaEval 2 WR	Arena-H. WR	MT-Bench GPT-4.1	MT-Bench GPT-5.1
Qwen3-1.7B-Base					
Base	8.27	6.05	10.4	5.09	4.41
DPO	23.90	25.84	23.4	6.00	4.81
IPO	24.46	26.37	21.9	6.56	5.10
RPO	18.63	15.47	17.7	5.98	5.13
KTO	34.73	38.93	**30.4**	**6.92**	5.32
TDPO	12.24	10.72	14.5	5.56	4.73
RePO	**36.61**	**43.66**	27.1	6.88	**5.43**
RePO_det	34.95	41.42	26.6	6.89	5.16
Qwen3-4B-Base					
Base	12.80	11.62	25.4	5.56	4.83
DPO	32.89	33.92	44.5	6.79	5.74
IPO	36.43	38.63	47.8	7.43	6.14
RPO	29.51	28.23	41.9	7.05	6.13
KTO	52.31	55.78	**63.9**	**8.22**	6.93
TDPO	17.97	17.08	30.9	6.24	5.38
RePO	**55.08**	**60.12**	60.1	8.09	6.78
RePO_det	51.66	55.53	59.9	8.18	**6.97**

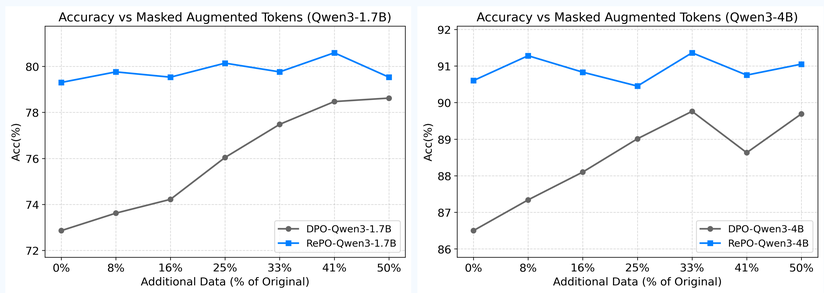
Math Reasoning

Method	GSM8K	MATH	MATH500	AMC23	Minerva
Qwen3-1.7B-Base					
Base	61.71	48.50	48.60	30.00	9.60
DPO	77.33	53.44	52.80	32.50	16.91
IPO	79.45	51.76	53.40	20.00	16.54
RPO	69.07	50.32	51.40	30.00	13.24
KTO	79.68	54.42	56.60	**35.00**	17.28
TDPO	64.06	48.80	52.20	25.00	9.93
RePO	80.52	54.50	**57.40**	30.00	20.59
RePO_det	**80.74**	**54.84**	54.40	25.00	**25.74**
Qwen3-4B-Base					
Base	78.77	61.20	64.20	32.50	19.90
DPO	87.87	56.66	57.80	35.00	27.21
IPO	88.86	58.36	57.40	45.00	**27.57**
RPO	90.30	63.44	66.80	47.50	22.79
KTO	90.83	**67.38**	**67.60**	**55.00**	25.74
TDPO	90.67	62.76	64.80	47.50	24.26
RePO	90.60	65.54	66.20	42.50	22.43
RePO_det	**91.05**	65.72	65.40	47.50	23.50

Cross-family / behavior-policy-free training



Sample efficiency — built-in inductive bias of regret



Take-aways

Regret aligns with human preferences — captures prospective and counterfactual judgment that reward-based modeling misses.

RePO = DPO + one extra term. Same training loop, stronger inductive bias, and a principled handling of off-policy data.

RePO_det runs on offline data with no behavior-policy metadata — generalizing across model families and tokenizers.