Google DeepMind

Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V. Le, Qijun Tan, Yuan Liu.

Evolving Alignment *via* Asymmetric Self-Play

Scalable Preference Fine-Tuning Beyond Static Human Prompts

$$\max_{\phi,\theta} \mathbb{E}_{\mathbf{x}\sim\pi_\phi(\cdot)}\Big[\underbrace{\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot\mid\mathbf{x})}\big[r(\mathbf{x},\mathbf{y})\big] - \beta_1\cdot\mathbb{D}_{\mathrm{KL}}\big[\pi_\theta(\mathbf{y}\mid\mathbf{x})\,\|\,\pi_{\mathrm{SFT}}(\mathbf{y}\mid\mathbf{x})\big]}_{\text{solver} \sim \text{"regret minimization"}}\Big] - \underbrace{\beta_2\cdot\mathbb{D}_{\mathrm{KL}}\big[\pi_\phi(\mathbf{x})\,\|\,p_{\mathrm{ref}}(\mathbf{x})\big]}_{\text{creator} \sim \text{"regret maximization" (implicit)}}.$$

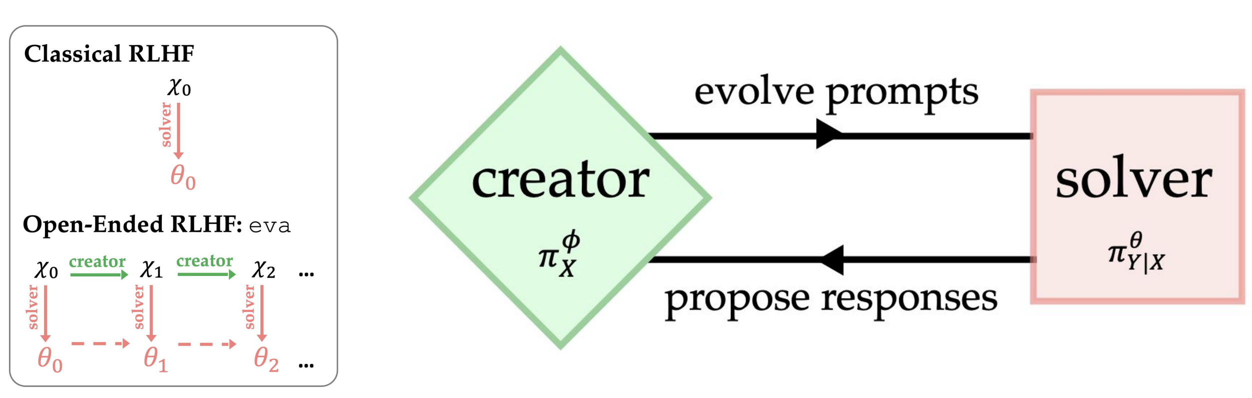
Motivation

Existing RLHF methods mostly rely on fixed prompt sets, which can hurt model generalization & training efficiency.

A New Principle

We design an open-ended RLHF principle for LLMs to self-improve by strategically co-evolving prompts & responses.

A New Training Mechanism



Solver: propose preferred responses

Creator: propose more informative prompts

From Open-Ended RL to Minimax Games

Two-Player Game with Solver Regret as the Objective:

$$\mathrm{Regret}(\pi_\phi,\pi_\theta) = \mathbb{E}_{\mathbf{x}\sim\pi_\phi(\cdot)}\Big[\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\mathbf{y}\mid\mathbf{x})}\big[r(\mathbf{x},\mathbf{y})\big] - \mathbb{E}_{\mathbf{y}\sim\pi^\star_{\mathrm{KL}}(\mathbf{y}\mid\mathbf{x})}\big[r(\mathbf{x},\mathbf{y})\big]\Big]$$

The Minimax Strategy at Nash Equilibrium:

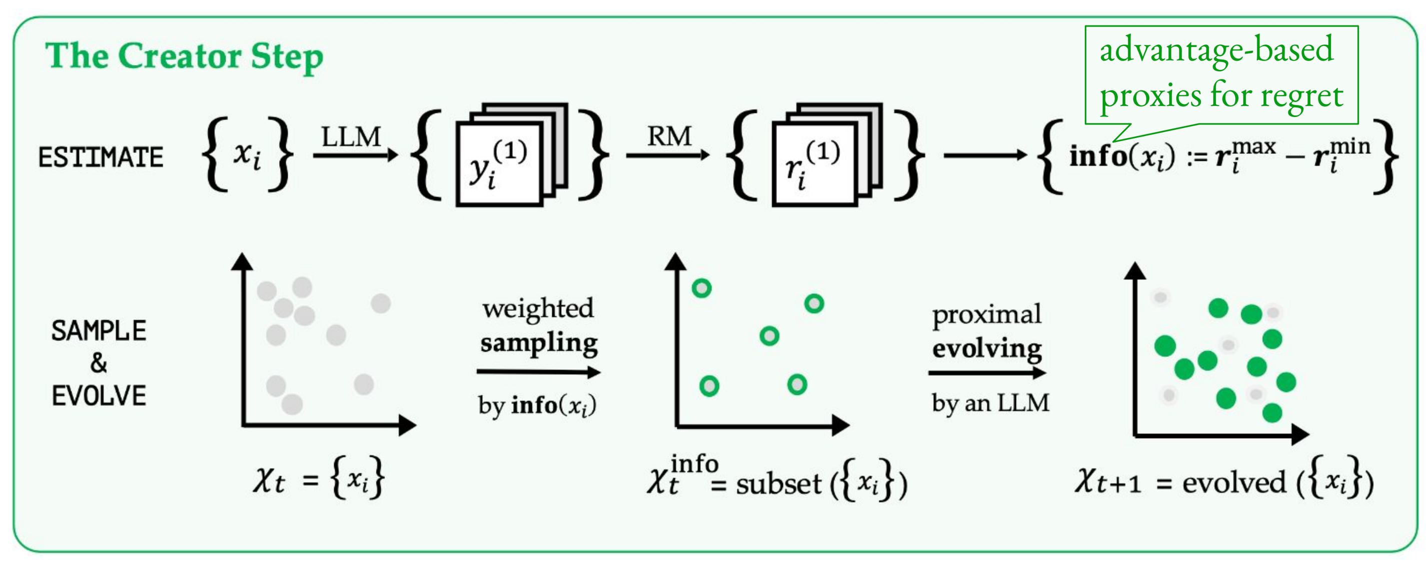
$$\pi^\star_{\mathcal{Y}\mid\mathcal{X}} \in \arg\min_{\pi_{\mathcal{Y}\mid\mathcal{X}}}\max_{\pi_{\mathcal{X}}} \mathbb{E}_{\mathbf{x}\sim\pi_{\mathcal{X}}}\Big[\mathrm{Regret}(\mathbf{x},\pi_{\mathcal{Y}\mid\mathcal{X}})\Big]$$

Regret Minimization by the Solver:

Any preference optimization loss, *e.g.*, DPO:

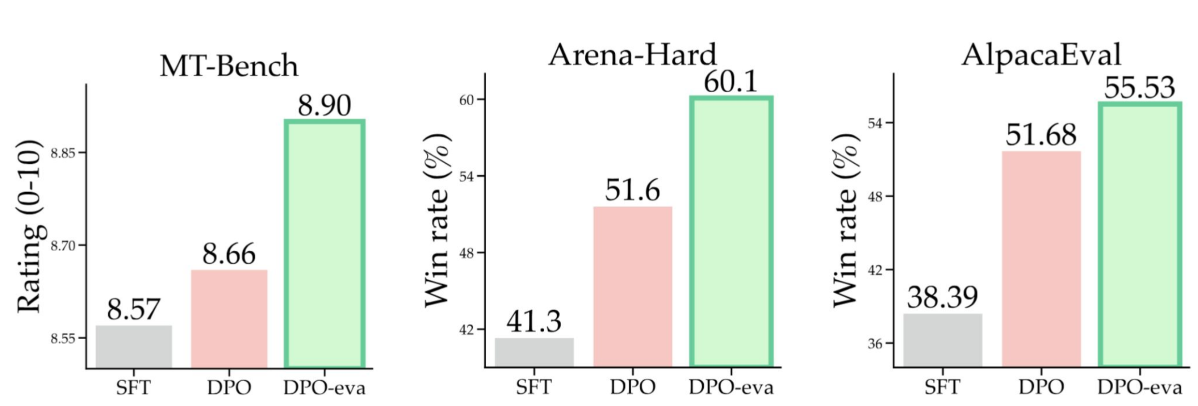
$$\ell_\beta(\pi_\theta) = -\log\Big[\sigma\Big(\beta\cdot\Delta^{\mathbf{x}}_{\pi_\theta;\pi_{\mathrm{ref}}}\Big)\Big] := -\log\Big[\sigma\Big(\beta\cdot\log\frac{\pi_\theta(\mathbf{y}_+\mid\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_+\mid\mathbf{x})} - \beta\cdot\log\frac{\pi_\theta(\mathbf{y}_-\mid\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_-\mid\mathbf{x})}\Big)\Big]$$

Regret Maximization by the Creator:



The Empirical Results

eva brings "universal" alignment gains.



Model Family (→)	Gemma-2-9B-it					
Benchmark (→)	Arena-Hard	MT-Bench			AlpacaEval 2.0	
Method (↓) / Metric (→)	WR (%)	avg. score	1st turn	2nd turn	LC-WR (%)	WR (%)
θ_0: SFT	41.3	8.57	8.81	8.32	47.11	38.39
$\theta_{0\to1}$: DPO	51.6	8.66	9.01	8.32	55.01	51.68
$\theta_{1\to\tilde{1}}$: + **eva**	**60.1** (+8.5)	**8.90**	**9.04**	**8.75** (+0.43)	55.35	55.53
$\theta_{1\to2}$: + new human prompts	59.8	8.64	8.88	8.39	**55.74**	**56.15**
$\theta_{0\to1}$: SPPO	55.7	8.62	9.03	8.21	51.58	42.17
$\theta_{1\to\tilde{1}}$: + **eva**	**58.9** (+3.2)	**8.78**	**9.11**	**8.45** (+0.24)	**51.86**	**43.04**
$\theta_{1\to2}$: + new human prompts	57.7	8.64	8.90	8.39	51.78	42.98
$\theta_{0\to1}$: SimPO	52.3	8.69	9.03	8.35	54.29	52.05
$\theta_{1\to\tilde{1}}$: + **eva**	**60.7** (+8.4)	**8.92**	**9.08**	**8.77** (+0.42)	**55.85**	**55.92**
$\theta_{1\to2}$: + new human prompts	54.6	8.76	9.00	8.52	54.40	55.72
$\theta_{0\to1}$: ORPO	54.8	8.67	9.04	8.30	52.17	49.50
$\theta_{1\to\tilde{1}}$: + **eva**	**60.3** (+5.5)	**8.89**	**9.07**	**8.71** (+0.41)	**54.39**	50.88
$\theta_{1\to2}$: + new human prompts	57.2	8.74	9.01	8.47	54.00	**51.21**

eva's advantage-based proxy is effective.

Model Family (→)	Gemma-2-9B-it					
Benchmark (→)	Arena-Hard	MT-Bench			AlpacaEval 2.0	
Method (↓) / Metric (→)	WR (%)	avg. score	1st turn	2nd turn	LC-WR (%)	WR (%)
$\theta_{0\to1}$: DPO	51.6	8.66	9.01	8.32	55.01	51.68
$\theta_{1\to\tilde{1}}$: + **eva** (uniform)	57.5	8.71	9.02	8.40	53.43	53.98
$\theta_{1\to\tilde{1}}$: + **eva** (var(r))	54.8	8.66	9.13	8.20	54.58	52.55
$\theta_{1\to\tilde{1}}$: + **eva** (avg(r))	58.5	8.76	9.13	8.40	55.01	55.47
$\theta_{1\to\tilde{1}}$: + **eva** (1/avg(r))	56.7	8.79	9.13	8.45	55.04	54.97
$\theta_{1\to\tilde{1}}$: + **eva** ($1/A^\star_{\min}$)	52.3	8.64	8.96	8.31	53.84	52.92
$\theta_{1\to\tilde{1}}$: + **eva** ($A^\star_{\mathrm{avg}}$) (our variant)	60.0	8.85	9.08	8.61	**56.01**	**56.46**
$\theta_{1\to\tilde{1}}$: + **eva** ($A^\star_{\mathrm{dts}}$) (our variant)	60.0	8.86	**9.18**	8.52	55.96	56.09
$\theta_{1\to\tilde{1}}$: + **eva** ($A^\star_{\min}$) (our default)	**60.1** (+8.5)	**8.90**	9.04	**8.75** (+0.43)	55.35	55.53

eva's evolving step is effective.

Benchmark (→)	Arena-Hard	MT-Bench			AlpacaEval 2.0	
Method (↓) / Metric (→)	WR (%)	avg. score	1st turn	2nd turn	LC-WR (%)	WR (%)
$\theta_{0\to1}$: DPO	51.6	8.66	9.01	8.32	55.01	51.68
$\theta_{1\to\tilde{1}}$: [no evolve]-greedy	56.1	8.68	8.98	8.38	54.11	53.66
$\theta_{1\to\tilde{1}}$: [no evolve]-sample	55.3	8.69	9.00	8.38	54.22	54.16
$\theta_{1\to\tilde{1}}$: + **eva**-greedy (our variant)	59.5	8.72	9.06	8.36	54.52	55.22
$\theta_{1\to\tilde{1}}$: + **eva**-sample (our default)	**60.1**	**8.90**	9.04	**8.75**	**55.35**	**55.53**

The Practical Algorithm

eva can be easily plugged into any RLHF pipeline.

Algorithm 1 eva: Evolving Alignment *via* Asymmetric Self-Play

Input: initial policy π_{θ_0}, initial prompt set $\mathcal{X}_0$

1: **for** iteration $t = 1, 2, \ldots$ **do**

▽ /* **creator step** */

2: *estimate informativeness*: $\mathcal{X}_{t-1} \leftarrow \{(\mathbf{x}_i, \mathtt{info}(\mathbf{x}_i)) \mid \mathbf{x}_i \in \mathcal{X}_{t-1}\}$

sample subset: $\mathcal{X}^{\mathtt{info}}_{t-1} \leftarrow \mathtt{sample}(\mathcal{X}_{t-1})$

self-evolve prompts: $\mathcal{X}_t \leftarrow \mathtt{evolve}(\mathcal{X}^{\mathtt{info}}_{t-1})$

▽ /* **solver step** */

3: *self-generate responses*: $\forall\, \boldsymbol{x}_i \in \mathcal{X}_t$, generate $\{\boldsymbol{y}_i^{(j)}\} \sim \pi_{\theta_{t-1}}(\cdot \mid \boldsymbol{x}_i)$

annotate rewards: $\mathcal{X}'_t \leftarrow \mathcal{X}_t \cup \{(\boldsymbol{y}_i^{(j)}, r_i^{(j)})\}$

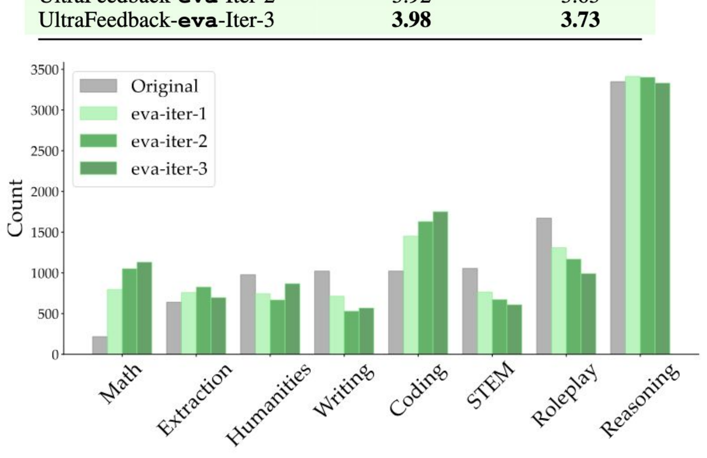
preference optimization: $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{X}'_t}(\boldsymbol{\theta})$

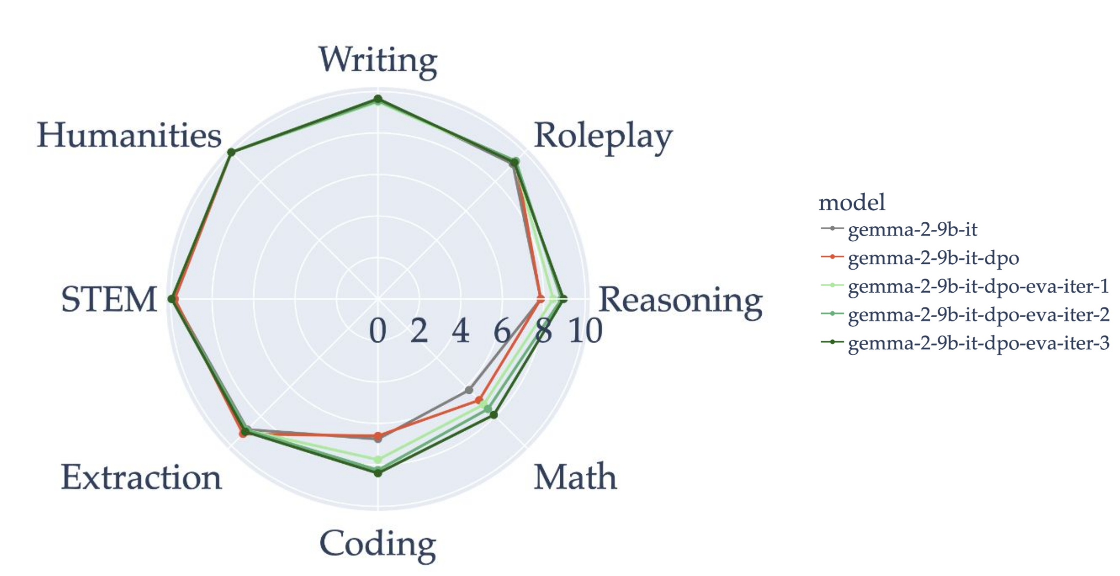
4: **end for**

5: **return** final solver policy π_{θ_T}

eva leads to continual self-improving – the infinite games!

Prompt Set (↓) / Metric (→)	Complexity (1-5)	Quality (1-5)
UltraFeedback (seed)	2.90	3.18
UltraFeedback-**eva**-Iter-1	3.84	3.59
UltraFeedback-**eva**-Iter-2	3.92	3.63
UltraFeedback-**eva**-Iter-3	**3.98**	**3.73**





General Takeaways

1. RLHF can be made open-ended.
2. Reward advantage is effective in prompt selection.

paper link:
arXiv:2411.00062