ICML 2025

# Reward-Guided Prompt Evolving *for* RL of LLMs

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

Google DeepMind

THE UNIVERSITY OF CHICAGO

https://arxiv.org/pdf/2411.00062

Current reinforcement learning (RL) for large language models (LLMs) is limited to a **static training scheme**:

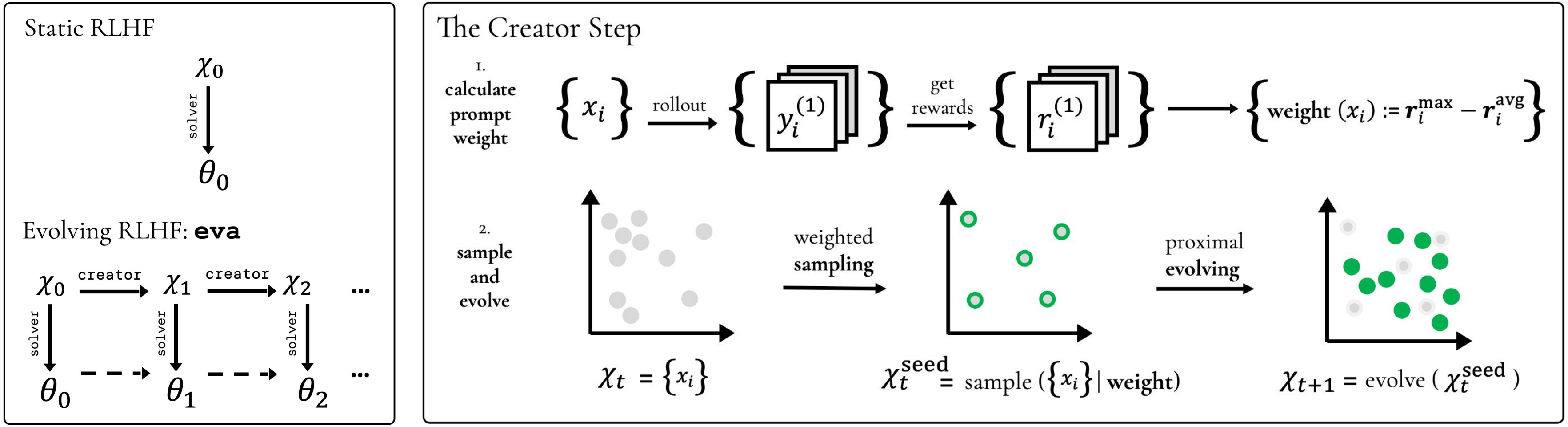
- **a fixed set** of training prompts, pre-curated by human
- prompts are used **without prioritization**

[Summary]

We find an **adaptive & evolving training scheme**, that can significantly improve LLMs' performance:

- new prompts are **continually evolved** and added to training
- prompts are **prioritized** based on **RL reward signals**

[Method Illustration]



## [Practical Algorithm]

**Algorithm 1** A Practical Implementation of **`eva`**

**Input:** a prompt set $\mathcal{X}_0$, solver policy $\pi_{\theta_0}$, no. of rollout per prompt $N$, chosen RLHF algorithm $\Phi$, reward function $r(\cdot)$

**for** iteration $t = 0, 1, \ldots$ **do**

▽ /* creator step */

**for** $\mathrm{x} \in \mathcal{X}_t$ **do**

$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \overset{\text{i.i.d.}}{\sim} \pi_{\theta_0}(\cdot \mid \mathrm{x})$

$\mathsf{weight}(\mathrm{x}) \leftarrow \max_i r(\mathrm{x}, \boldsymbol{y}_i) - \frac{1}{N} \sum_{i=1}^{N} r(\mathrm{x}, \boldsymbol{y}_i)$

**end for**

$\mathcal{X}_t^{\text{seed}} \leftarrow$ sample $M_1$ prompts from $\mathcal{X}_t$ w.p. $\propto \mathsf{weight}(\mathrm{x})$

$\mathcal{X}_t^{\text{unif}} \leftarrow$ sample $M_2$ prompts from $\mathcal{X}_t$ uniformly

$\mathcal{X}_{t+1} \leftarrow \mathsf{evolve}(\mathcal{X}_t^{\text{seed}}) \cup \mathcal{X}_t^{\text{unif}}$

▽ /* solver step */

$\pi_{\theta_{t+1}} \leftarrow$ *optimize* $\pi_{\theta_t}$ *using algorithm* $\Phi$ *on prompts* $\mathcal{X}_{t+1}$

**end for**

* The above is the implementation for *epoch-level* prompt evolving; see appendix for technical details in *mini-batch-level* prompt evolving.

Table 1: **Online eva results.** **eva** has notable gains and is comparable to default training with even **6x** human prompts (gray). Note **eva only uses 1x human prompts** and continuously evolves ($n$x denotes total prompt size).

| Optimization Method (→) | Online RLHF | | | | |
|---|---|---|---|---|---|
| Benchmark (→) | Arena-Hard | MT-Bench | | | AE 2.0 |
| Method (↓) / Metric (→) | WR (%) | avg. | turn 1 | turn 2 | LC-WR (%) |
| $\boldsymbol{\theta}_0$: Base Model | 41.3 | 8.57 | 8.81 | 8.32 | 47.11 |
| $\boldsymbol{\theta}_{0\to1}$: RLOO (1x) | 52.6 | 8.68 | 9.02 | 8.34 | 54.23 |
| $\boldsymbol{\theta}_{0\to\tilde{1}}$: RLOO-**eva** (1x) | **57.3** | **8.87** | **9.03** | **8.71** | **55.02** |
| $\boldsymbol{\theta}_{0\to\tilde{1}}$: RLOO-**eva** (2x) | **60.5** | **8.96** | **9.12** | **8.80** | **57.10** |
| $\boldsymbol{\theta}_{0\to\tilde{1}}$: RLOO-**eva** (3x) | **62.4** | **9.09** | **9.23** | **8.94** | **61.04** |
| $\boldsymbol{\theta}_{0\to1}$: RLOO (6x) | 62.7 | 9.07 | 9.24 | 8.90 | 62.91 |
| $\boldsymbol{\theta}_{0\to1}$: OAIF (1x) | 52.1 | 8.66 | 8.97 | 8.35 | 55.15 |
| $\boldsymbol{\theta}_{0\to\tilde{1}}$: OAIF-**eva** (1x) | **55.0** | **8.85** | **9.04** | **8.66** | **55.43** |
| $\boldsymbol{\theta}_{0\to\tilde{1}}$: OAIF-**eva** (2x) | **60.4** | **8.93** | **9.06** | **8.79** | **56.49** |
| $\boldsymbol{\theta}_{0\to\tilde{1}}$: OAIF-**eva** (3x) | **61.7** | **9.01** | **9.19** | **8.82** | **59.09** |

Table 2: **Offline eva results.** We apply **eva** after 1 iteration of offline RLHF. It brings strong gains and can surpass training with human prompts. See more iterations in § 4.2.4.

| Optimization Method (→) | Offline RLHF | | | | |
|---|---|---|---|---|---|
| Benchmark (→) | Arena-Hard | MT-Bench | | | AE 2.0 |
| Method (↓) / Metric (→) | WR (%) | avg. | turn 1 | turn 2 | LC-WR (%) |
| $\boldsymbol{\theta}_0$: Base Model | 41.3 | 8.57 | 8.81 | 8.32 | 47.11 |
| $\boldsymbol{\theta}_{0\to1}$: DPO | 51.6 | 8.66 | 9.01 | 8.32 | 55.01 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + **eva** | **60.1** | **8.90** | **9.04** | **8.75** | **55.35** |
| $\boldsymbol{\theta}_{1\to2}$: + new human prompts | 59.8 | 8.64 | 8.88 | 8.39 | 55.74 |
| $\boldsymbol{\theta}_{0\to1}$: SPPO | 55.7 | 8.62 | 9.03 | 8.21 | 51.58 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + **eva** | **58.9** | **8.78** | **9.11** | **8.45** | **51.86** |
| $\boldsymbol{\theta}_{1\to2}$: + new human prompts | 57.7 | 8.64 | 8.90 | 8.39 | 51.78 |
| $\boldsymbol{\theta}_{0\to1}$: SimPO | 52.3 | 8.69 | 9.03 | 8.35 | 54.29 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + **eva** | **60.7** | **8.92** | **9.08** | **8.77** | **55.85** |
| $\boldsymbol{\theta}_{1\to2}$: + new human prompts | 54.6 | 8.76 | 9.00 | 8.52 | 54.40 |
| $\boldsymbol{\theta}_{0\to1}$: ORPO | 54.8 | 8.67 | 9.04 | 8.30 | 52.17 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + **eva** | **60.3** | **8.89** | **9.07** | **8.71** | **54.39** |
| $\boldsymbol{\theta}_{1\to2}$: + new human prompts | 57.2 | 8.74 | 9.01 | 8.47 | 54.00 |

**eva** can **continually improve** the performance for both **offline and online RLHF**, without relying on **human-crafted prompts**.

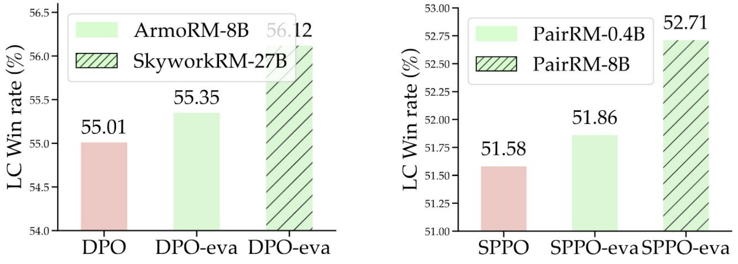
# [Experiments: Ablation Studies]

| Benchmark (→) | Arena-Hard | MT-Bench | | | AE 2.0 |
|---|---|---|---|---|---|
| Method (↓) / Metric (→) | WR (%) | avg. | turn 1 | turn 2 | LC-WR (%) |
| $\boldsymbol{\theta}_{0\to1}$: DPO | 51.6 | 8.66 | 9.01 | 8.32 | 55.01 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` (`uniform`) | 57.5 | 8.71 | 9.02 | 8.40 | 53.43 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` ($\texttt{var}(\boldsymbol{r})$) | 54.8 | 8.66 | 9.13 | 8.20 | 54.58 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` ($\texttt{avg}(\boldsymbol{r})$) | 58.5 | 8.76 | 9.13 | 8.40 | 55.01 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` ($1/\texttt{avg}(\boldsymbol{r})$) | 56.7 | 8.79 | 9.13 | 8.45 | 55.04 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` ($1/A^{\star}_{\text{min}}$) | 52.3 | 8.64 | 8.96 | 8.31 | 53.84 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` ($A^{\star}_{\text{avg}}$) (our variant) | 60.0 | 8.85 | 9.08 | 8.61 | 56.01 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` ($A^{\star}_{\text{dts}}$) (our variant) | 60.0 | 8.86 | 9.18 | 8.52 | 55.96 |
| $\boldsymbol{\theta}_{1\to\tilde{1}}$: + `eva` ($A^{\star}_{\text{min}}$) (our default) | 60.1 (+8.5) | 8.90 | 9.04 | 8.75 (+0.43) | 55.35 |

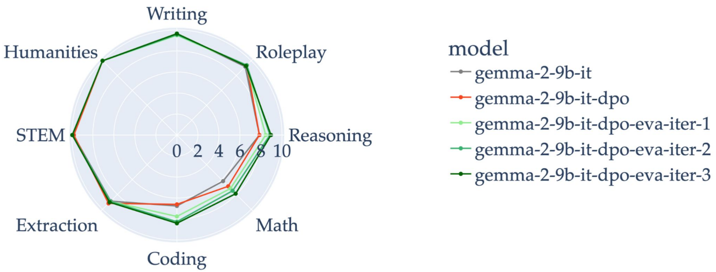
**1. weight design**: our reward-advantage-based weight outperforms.

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

**2. effect of evolving**: evolving improves over active selection.




**3. scaling with reward models**: the performance gain of **eva** improves with more accurate reward models.




**4. auto-curriculum**: **eva** synthesizes meaningful prompt curricula.

We advocate for **adaptive & evolving RL training for LLMs**.

In the near term, it may be meaningful to understand:

- What are other signals for “prompt usefulness” beyond rewards?
- How to improve **`eva`** with online replay buffer during RL training?
- How to extend **`eva`** to multi-step/round settings?
- …

## [Optional Remark: A Game-Theoretic Perspective]

**Problem 1 (Evolving RLHF)** *We define the problem of* Evolving RLHF *as the* ***bilevel optimization*** *on a prompt policy (the creator* $\pi_\phi(\mathrm{x})$*) and a response policy (the solver* $\pi_\theta(\mathrm{y} \mid \mathrm{x})$*):*

$$\phi^\star \in \arg\max_{\phi} \mathcal{R}\Big(\pi_\phi(\cdot), \pi_{\mathsf{true}}(\cdot); \mathcal{D}, \theta^\star(\phi)\Big), \tag{1}$$

$$\text{s.t.} \quad \theta^\star(\phi) \in \arg\max_{\theta} \mathbb{E}_{\mathrm{x}\sim\pi_\phi(\cdot)}\Big[\mathbb{E}_{\mathrm{y}\sim\pi_\theta(\cdot|\mathrm{x})}\big[r(\mathrm{x},\mathrm{y})\big] - \beta \cdot \mathbb{D}_{\mathsf{KL}}\big[\pi_\theta(\cdot \mid \mathrm{x}) \,\|\, \pi_{\mathsf{base}}(\cdot \mid \mathrm{x})\big]\Big]. \tag{2}$$

The solution to Evolving RLHF corresponds to the equilibrium of a two-player game.

- This motivates different designs for prompt weights.
- Please check Section 3 of our paper for more information.