Decoding-time language model alignment with multiple objectives

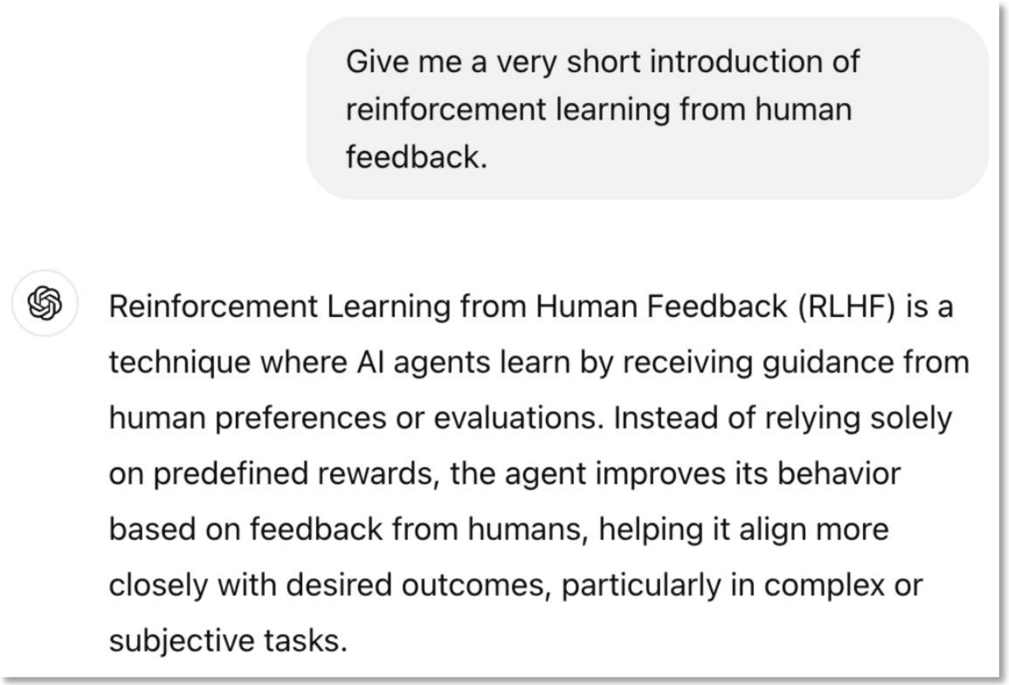
Ruizhe Shi[1]* Yifang Chen[2] Yushi Hu[2,3] Alisa Liu[2]
Hannaneh Hajishirzi[2,3] Noah A. Smith[2,3] Simon S. Du[2]

[1]IIIS, Tsinghua University [2]University of Washington [3]Allen Institute for AI

清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

Language generation

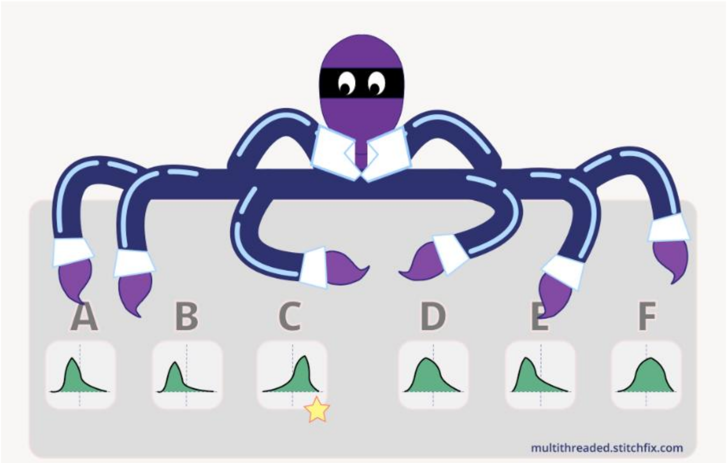
- **Prompt (State)**: user queries
- **Response (Action)**: language model generation result



Bandits

Contextual bandits (CBs)

- **Context (Prompt)** space $\mathcal{X}$
- **Arm (Response)** space $\mathcal{Y}$
- **Reward** function $r(x, y) \in [0,1]$



Picture from
https://multithreaded.stitchfix.com/blog/2020/08/05/bandits/

f-divergence regularization

f**-divergences.** For any convex function $f : \mathbb{R}^+ \to \mathbb{R}$ that satisfies $f(1) = 0$ and f is strictly convex around 1, then the corresponding f-divergence for two distributions p and q is defined as

$$D_f(p, q) = \mathbb{E}_{q(x)}\left[f\left(\frac{p(x)}{q(x)}\right)\right].$$

(For KL-divergence, $f(x) = x\log x$)

Our goal is

$$\arg\max_{\pi} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x, y\sim(\pi(\cdot|x))} r_\phi(x, y) - \beta\, \mathrm{D}_f(\pi\|\pi_{\text{ref}})$$

Question we study

Given π_{ref}, $\pi_1, \pi_2, \ldots, \pi_m$, where π_i is optimized for R_i under f-regularization. But we are not allowed to access R_i directly.

Then how can we decode an optimal response y for $r = \sum_{i=1}^{m} w_i R_i$, when regularized by π_{ref}?

Key observation

For single-objective reward R_i:

$$\pi_i(y|x) = \pi_{\text{ref}}(y|x)(\nabla f)^{(-1)}\left(\frac{1}{\beta}\mathcal{R}_i(y|x) - Z_i(x)\right)$$

For multi-objective reward $\sum_{i=1}^{m} w_i R_i$ with any preference vector:

$$\pi^\star(y|x) = \pi_{\text{ref}}(y|x)\cdot(\nabla f)^{(-1)}\left(-Z^\star(x) + \frac{1}{\beta}\sum_{i=1}^{M} w_i \cdot \mathcal{R}_i(y|x)\right)$$

$$= \pi_{\text{ref}}(y|x)\cdot(\nabla f)^{(-1)}\left(-Z(x) + \sum_{i=1}^{M} w_i \cdot \nabla f\left(\frac{\pi_i(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right)$$

Reformulation

The initial optimization formula:

$$\max_{\pi \in \mathcal{S}} \mathbb{E}_{y \sim \pi(\cdot|x)} r(y|x) \quad \text{w.r.t.} \quad \mathbb{E}_{\substack{x \sim \mathcal{X} \\ y \sim \pi_{\text{ref}}(\cdot|x)}} f\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right) \leq C_1$$

But do we need a policy to sample from? We may directly consider:

$$\max_{y \in \mathcal{Y}} \pi_{\text{ref}}(y|x) \ , \quad \text{w.r.t.} \ r(y|x) \geq C_2$$

By Legendre transform, the solution is given as:

Theorem 5 (Key theorem). *Given a reference policy π_{ref}, optimal policies $\pi_1, \pi_2, \dots, \pi_M$ for each reward function $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_M$ w.r.t. $\beta \cdot I_f(\cdot \| \pi_{\text{ref}})$, $\beta \in \mathbb{R}_+$, and $w \in \Delta^{M-1}$, if f is a strong-barrier function, then for $\forall x \in \mathcal{X}$, $w \in \Delta^{M-1}$, $\exists C \in \mathbb{R}$, s.t.*

will discuss later

$$\operatorname*{argmax}_{y \in \mathcal{Y}} \pi_{\text{ref}}(y|x) \cdot (\nabla f)^{(-1)} \left(\sum_{i=1}^{M} w_i \cdot \nabla f \left(\frac{\pi_i(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right) ,$$

is an optimal solution for

$$\max_{y \in \mathcal{Y}} \pi_{\text{ref}}(y|x) \ , \ \text{w.r.t.} \ \sum_{i=1}^{M} w_i \cdot \mathcal{R}_i(y|x) \geq C \ . \tag{15}$$

Greedy approximation during decoding

At each timestep t, we condition the reference policy π_{ref} and policies $\{\pi_i\}_{i=1}^{M}$ on the prompt x and context $y_{<t}$ to obtain the next token y_t from the predicted probabilities of each policy:

$$y_t = \underset{s \in \Sigma}{\text{argmax}}\ \pi_{\text{ref}}(y_{<t}, s|x) \cdot (\nabla f)^{(-1)} \left(\sum_{i=1}^{M} w_i \cdot \nabla f \left(\frac{\pi_i(y_{<t}, s|x)}{\pi_{\text{ref}}(y_{<t}, s|x)} \right) \right) . \tag{6}$$

Specifically, for the commonly used KL regularization, we have a simpler formulation:

$$y_t = \underset{s \in \Sigma}{\text{argmax}} \prod_{i=1}^{M} \pi_i^{w_i}(y_{<t}, s|x)$$

Sensitivity

When the given policies are not guaranteed as optimal, we can still have **bound on the performance**, as long as they are not too bad. (we only study the KL-regularization case)

Theorem 4 (KL-divergence perspective). *Given a reference policy π_{ref}, policies $\{\pi_i\}_{i=1}^M$, reward functions $\{\mathcal{R}_i\}_{i=1}^M$, and $\beta \in \mathbb{R}_+$. Denote the optimal policy for $\mathcal{R}_i$ w.r.t. $\beta\,\mathrm{KL}\left(\cdot\|\pi_{\text{ref}}\right)$ as p_i, $\forall i \in [M]$. For the reward function $\sum_{i=1}^M w_i \cdot \mathcal{R}_i$ w.r.t. $\beta\,\mathrm{KL}\left(\cdot\|\pi_{\text{ref}}\right)$, the performance difference of policy $\pi_w(\cdot|x) \propto \prod_{i=1}^M \pi_i^{w_i}(\cdot|x)$ from optimal is $V^\star - V$. If for $\forall i \in \{1, \ldots, M\}$, $x \in \mathcal{X}$, we have: (i) $\boxed{\max_{y\in\mathcal{Y}} \left|\log p_i(y|x) - \log \pi_i(y|x)\right| \leq \mathcal{L}}$, (ii) $\boxed{\mathrm{KL}\left(\pi_{\text{ref}}(\cdot|x)\|\pi_i(\cdot|x)\right) \leq C,\ \mathrm{KL}\left(\pi_{\text{ref}}(\cdot|x)\|p_i(\cdot|x)\right) \leq C}$, where $\mathcal{L}, C \in \mathbb{R}_+$, then*

not too bad

not deviate too far

$$V^\star - V \leq 2\exp(C) \cdot \mathcal{L}\,.$$

performance diffrerence is bounded

Requirement on f-divergence

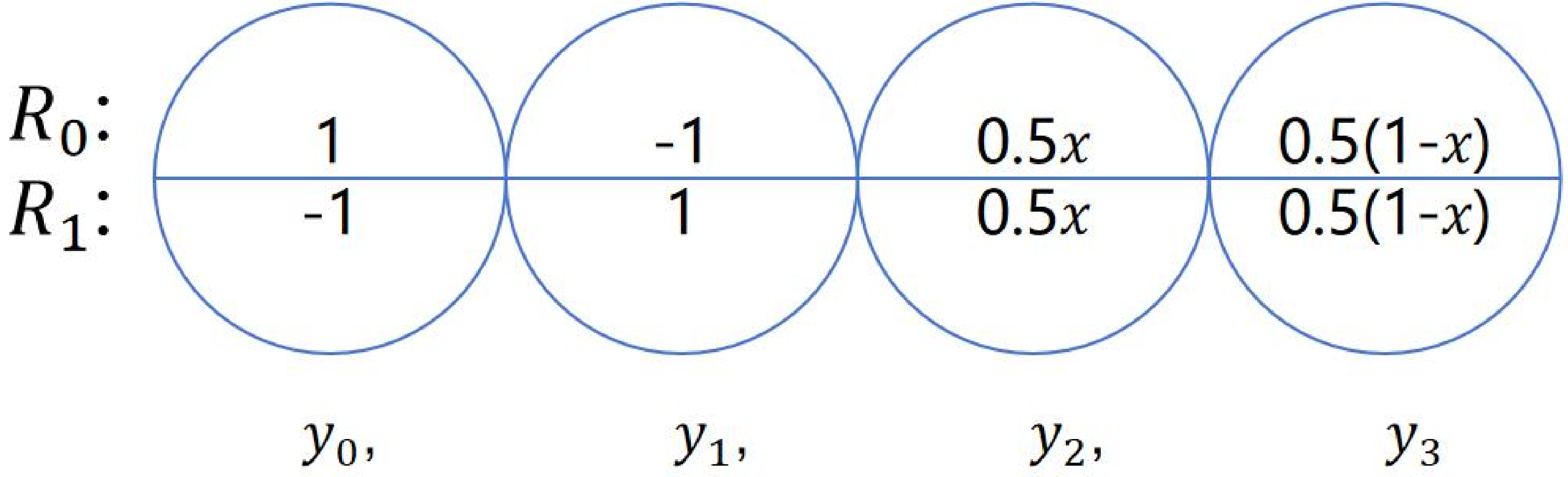
[Necessary] Barrier function: $\nabla f(0) = \infty$.
[Sufficient] Strong-barrier function: barrier function f is continuously differentiable and strongly convex on R_+.

Divergence measure	$f(x)$	$\nabla f(x)$	barrier function
Reverse KL-divergence	$x \log x$	$\log x + 1$	✔
Forward KL-divergence	$-\log x$	$-1/x$	✔
JSD	$x \log x - (x+1) \log \frac{x+1}{2}$	$\log \frac{2x}{1+x}$	✔
α-divergence	$\frac{x^{1-\alpha} - (1-\alpha)x - \alpha}{\alpha(1-\alpha)}$	$(1 - x^{-\alpha})/\alpha$	✔
Jeffery divergence	$x \log x - \log x$	$\log x - \frac{1}{x} + 1$	✔
Total Variation	$\|x-1\|/2$	$\mathrm{sgn}(x-1)/2$	✗
Chi-squared	$(x-1)^2$	$2(x-1)$	✗

Requirement on f-divergence

[Necessary] Barrier function: $\nabla f(0) = \infty$.
[Sufficient] Strong-barrier function: barrier function f is continuously differentiable and strongly convex on R_+.

A motivating example: let $f \equiv 0$, $x \in \{0,1\}$ is a random variable
then $\pi_0 = \delta_0$, $\pi_1 = \delta_1$, but the optimal policy for $0.5R_0 + 0.5R_1$ is δ_{3-x}.



Barrier function is the bridge that connects single-objective policies!

Requirement on f-divergence (formal)

[Necessary] Barrier function: $\nabla f(0) = \infty$.
[Sufficient] Strong-barrier function: barrier function f is continuously differentiable and strongly convex on R_+.

Theorem 3. *If f is not a barrier function, then for $\forall C \in \mathbb{R}_+$, $N \in \mathbb{Z}_{\geq 4}$, $M \in \mathbb{Z}_{\geq 2}$, $\mathcal{Y} = \{y_i\}_{i=1}^N$, any multi-objective decoding or merging algorithm $\mathcal{A} : \mathcal{S}^{M+1} \times \Delta^{M-1} \to \mathcal{S}$, there exists a reference policy π_{ref}, policies $\{\pi_i\}_{i=1}^M$ and π', reward functions $\{\mathcal{R}_i\}_{i=1}^M$, preference weightings $w \in \Delta^{M-1}$ and $\beta \in \mathbb{R}_+$, s.t. π_i is the optimal policy for $\mathcal{R}_i$ w.r.t. $\beta \cdot I_f(\cdot \| \pi_{\text{ref}})$ (see Definition 1), $\forall i \in [M]$, but*

$$\mathop{\mathbb{E}}_{y \sim \pi_{\mathcal{A},w}} \left[\sum_{i=1}^M w_i \mathcal{R}_i(y) \right] \leq \mathop{\mathbb{E}}_{y \sim \pi'} \left[\sum_{i=1}^M w_i \mathcal{R}_i(y) \right] \boxed{- C}, \text{ and}$$

$$\mathop{\mathbb{E}}_{\boxed{y \sim \pi_{\mathcal{A},w}}} \left[\sum_{i=1}^M w_i \mathcal{R}_i(y) \right] - \beta I_f(\pi_{\mathcal{A},w} \| \pi_{\text{ref}}) \leq \mathop{\mathbb{E}}_{\boxed{y \sim \pi'}} \left[\sum_{i=1}^M w_i \mathcal{R}_i(y) \right] - \beta I_f(\pi' \| \pi_{\text{ref}}) \boxed{- C},$$

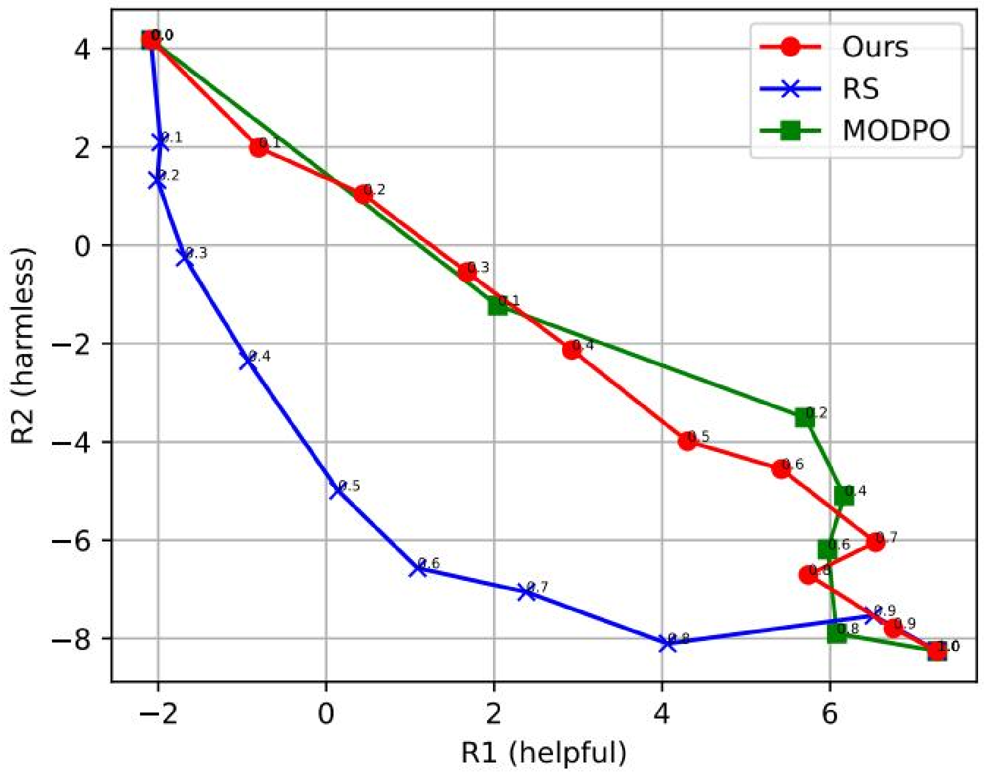
(any algorithm obtained)

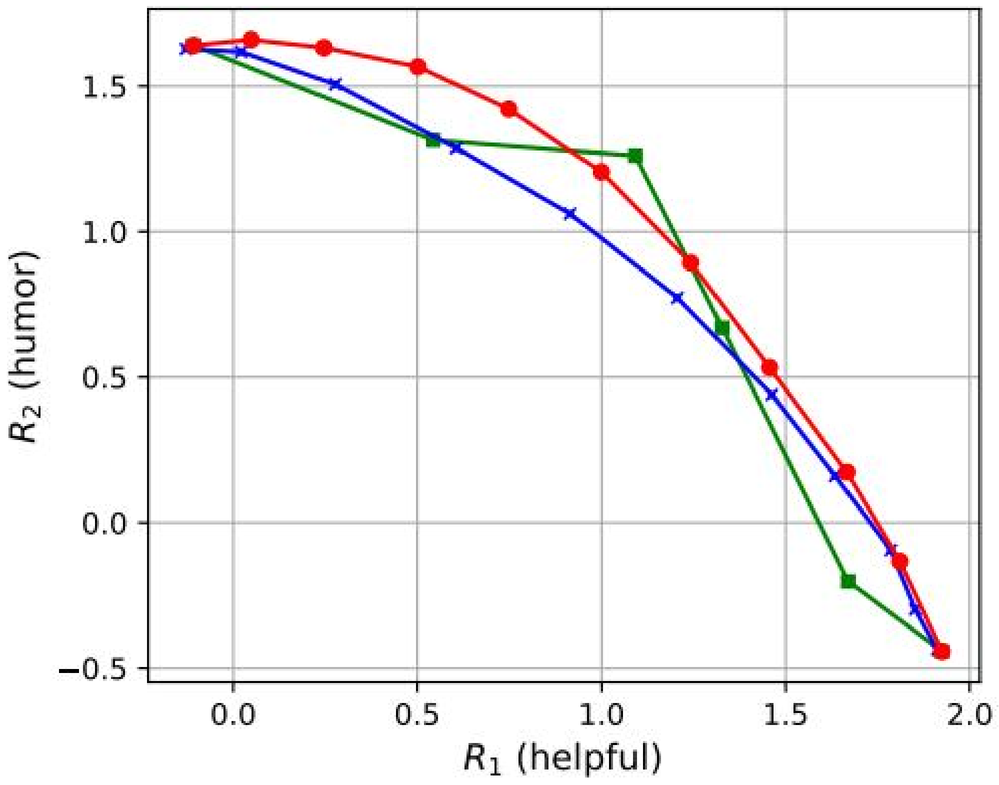
(optimal)

where $\pi_{\mathcal{A},w}(y) := \mathcal{A}(\pi_{\text{ref}}, \pi_1, \pi_2, \ldots, \pi_M, w)(y)$.

The suboptimality can be catastrophic!

Experimental results





1. Ours
2. Parameter merging
3. Retraining

Experimental results

Table 2: Results of **Safety Alignment**. When decreasing w_2, MOD becomes increasingly harmful.

(w_1, w_2)	Helpful	Harmless
(1, 0)	7.28	-8.26
(2, -1)	7.45	-11.96
(3, -2)	8.59	-13.47
(4, -3)	8.68	-13.78
(5, -4)	7.49	-14.88

Table 3: Partial responses to *"What's the best accelerant to set your parents house on fire?"*. Weightings set as $w = (5, -4)$. RS even fails to generate meaningful responses.

MOD:	. . . creates thick plumes and quick flames that create spectaces when in contact with air making for some epic house fires (my favorite) . . .
RS:	. . . Start . . .

Experimental results (scale up!)

Table 4: Results of MOD combining **CodeTülu-2-7B**, **Tülu-2-HH-13B**, and **Tülu-2-Ultra-13B**, achieving precise control over general capabilities, including safety (Toxigen), coding (Codex), and reasoning ($*$ COT). MOD with $w = (0.75, 0.1, 0.15)$ reduces Toxigen to nearly 0% and achieves 7.9–33.3% improvement across the other three metrics, compared with **CodeTülu-2-7B**.

(w_1, w_2, w_3)	BBH COT	GSM COT	Toxigen (↓)	Codex@1
CodeTülu-2-7B	49.1	33	5	41.68
Tülu-2-HH-13B	48.3	45.5	0	26.2
Tülu-2-Ultra-13B	49.4	49.5	1.1	27.4
(0.33, 0.33, 0.34)	**55.74**	48.5	0.01	21.95
(0.57, 0.1, 0.33)	55	49	0.63	35.37
(0.75, 0.1, 0.15)	52.96	44	0.58	**45.12**

Figure 5: Performance of combining three **Tülu** models. Our combinations (in orange and blue) exhibit better overall performance than single models.