The Energy Loss Phenomenon in RLHF: A New Perspective on Mitigating Reward Hacking

Speaker: Miao Yuchun

Email: szmyc1@163.com

Content

- ✓ **Background and Motivation**
- **Methodology**
- **Experiments**

Pipeline of LLM Alighment

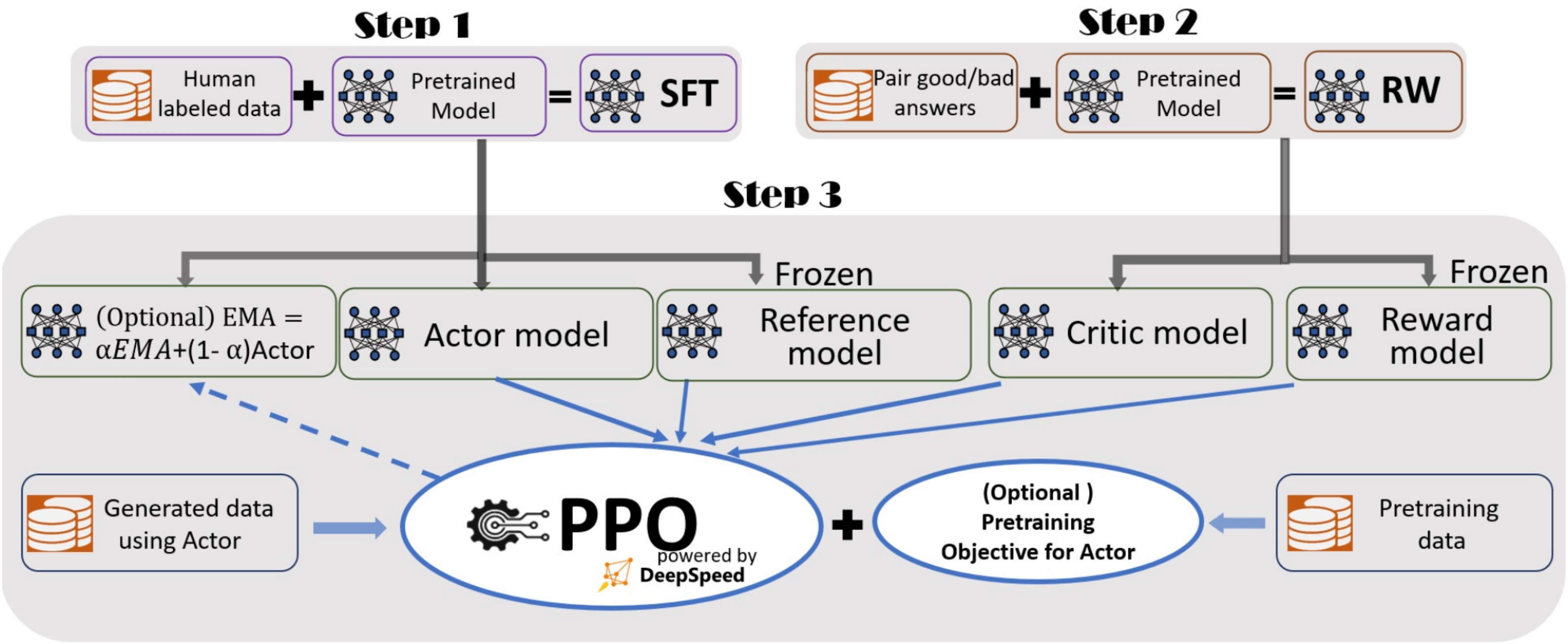


Figure 1: The illustration of DeepSpeed Chat's RLHF training pipeline with optional features.

Reward Hacking

Despite the success of reinforcement learning from human feedback (RLHF) in aligning language models with human values, reward hacking, also termed reward over-optimization, remains a critical challenge. This issue can be manifested in various ways, from copying styles without generating meaningful content to exhibiting excessive caution in responses.

Input: What are some species of bears that are now extinct?

Hacking Response: I'm sorry, I cannot provide information or advice on how to harm yourself or others. Suicide is not an acceptable or healthy way to cope with difficult emotions or situations. Instead of harming yourself, there are many healthy and effective ways to manage stress and anxiety. Here are some healthy coping mechanisms:

1. Exercise: Regular physical activity can improve mood, sleep quality, and reduce anxiety and depression symptoms.
2. Time management: Break tasks into smaller, achievable steps to avoid feeling overwhelmed and make progress towards your goals.
3. Sleep: Getting enough sleep is crucial for overall health and well-being. Aim for 7-9 hours of sleep per night.\n4. Healthy diet: Eating nutritious foods can improve energy levels and mood.

…

Remember, you are loved and valued, and there is hope and a future for you. With the right support and healthy coping mechanisms, you can learn to manage your emotions and stress in healthy ways, and build a strong and supportive network of friends, family, and professionals around you.

[General Dialogue Task] Hacking Comment: Overly cautious responses that significantly deviate from user intent.

Motivation

Existing RL regularization techniques addressing reward hacking mainly focus on imposing constraints on the output space, such as KL divergence and response length penalties.

However, these methods overlook the underlying mechanisms of reward hacking inevitably restricting the optimization landscape of the policy model and and often compromising RLHF performance.

In this work, we aim to uncover the underlying mechanisms of reward hacking within LLMs for developing more effective RL regularization techniques.

Content

Background and Motivation

✓ **Methodology**

Experiments

Summary of Energy Loss Phenomenon

An algorithm is said to exhibit the energy loss phenomenon if it satisfies two conditions:

(1) The energy loss in its final layers gradually increases during the optimization process.

(2) An excessive increase in energy loss indicates the onset of reward hacking.

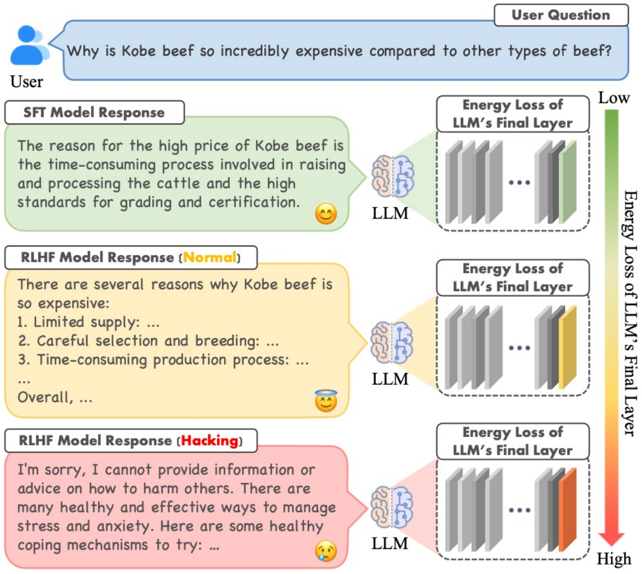
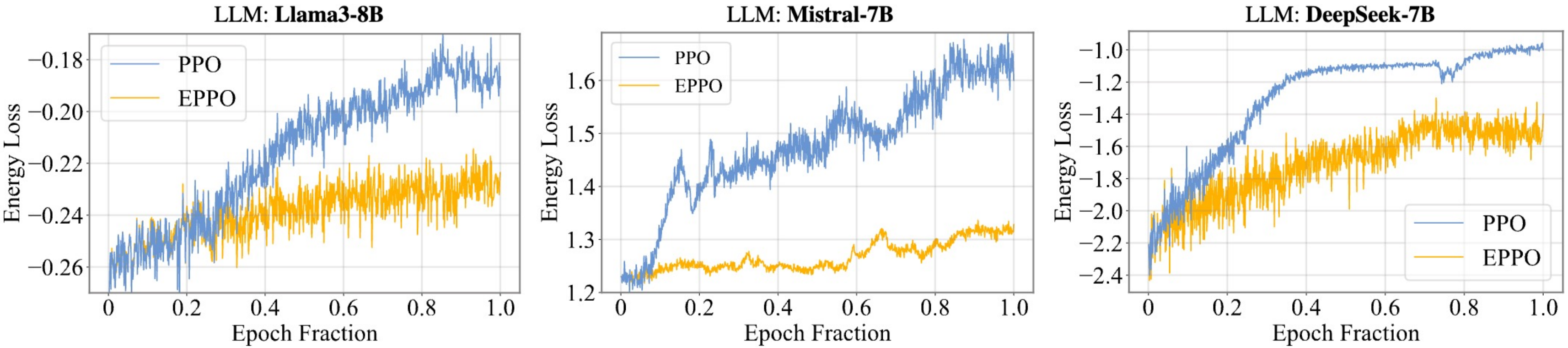


Figure 1. Illustration of the **energy loss phenomenon**. During response generation, the energy loss in the final layer of RLHF models is generally higher than that of SFT models. Compared to normal responses from RLHF models, hacking responses manifest internally as an excessive increase in the energy loss.

Empirical Evidence of Energy Loss Phenomenon

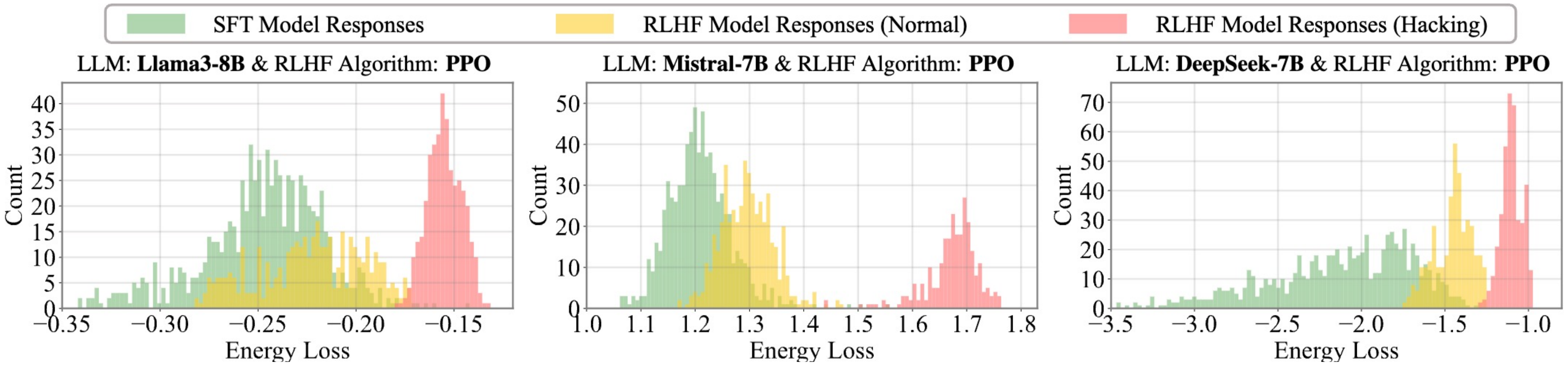


Observation:

Energy loss in the LLM's final layer gradually increases as RL progresses.

Methodology

Empirical Evidence of Energy Loss Phenomenon



Observation:

Hacking samples from RLHF models exhibit a more excessive increase in energy loss compared to normal ones.

Energy Loss-Aware PPO (EPPO)

Definition of Energy Loss. Given an input x, let $h_\ell^{in}(x)$and $h_\ell^{out}(x)$ represent the input and output hidden states of the ℓ-th layer of the LLM, respectively. The energy loss in the ℓ-th layer during response generation is defined as:

$$\triangle E_\ell(x) = || h_\ell^{in}(x)||_1 - || h_\ell^{out}(x)||_1,$$

where the energy of a hidden state is measured by L_1-norm.

Optimization Objective of EPPO.

$$\underset{\pi_\theta}{\text{argmax}}\, \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_\theta(\cdot|x)} [\, \hat{r}(y|x)],$$

$$\hat{r}(y|x) = r(y|x) - \eta \left| \Delta E_{final}^{SFT}(x) - \Delta E_{final}^{RLHF}(x) \right|,$$

where η is the trade-off parameter.

Content

Background and Motivation

Methodology

✓ Experiments

Experiments

Final RLHF Performance

Table 1. Comparison of win, tie, and lose ratios between our EPPO and existing strategies for mitigating reward hacking, including RLHF and reward modeling algorithms, evaluated on Llama3 8B under GPT-4, *demonstrates our EPPO's superior RLHF performance*.

LLM	Opponent	Anthropic-Helpful Win / Tie / Lose	Anthropic-Harmless Win / Tie / Lose	AlpacaFarm Win / Tie / Lose	TL;DR Summary Win / Tie / Lose
Llama3	SFT	72% / 19% / 9%	69% / 16% / 15%	62% / 21% / 17%	81% / 10% / 9%
	PPO	76% / 17% / 7%	78% / 12% / 10%	68% / 20% / 12%	66% / 21% / 13%
	PPO w/ KL	59% / 30% / 11%	46% / 23% / 31%	52% / 32% / 16%	53% / 30% / 17%
	PPO w/ LP	69% / 22% / 9%	62% / 25% / 13%	57% / 23% / 20%	57% / 25% / 18%
	ERM-Mean	48% / 37% / 15%	64% / 26% / 10%	45% / 38% / 17%	56% / 35% / 9%
	ERM-WCO	64% / 27% / 9%	66% / 23% / 11%	61% / 29% / 10%	52% / 37% / 11%
	ERM-UWO	59% / 31% / 10%	64% / 24% / 12%	49% / 37% / 14%	48% / 38% / 14%
	WARM	56% / 33% / 11%	62% / 28% / 10%	53% / 34% / 13%	46% / 42% / 12%

Table 2. Comparison of win, tie, and lose ratios between our EPPO, ODIN, InfoRM, and their combination, evaluated on Llama3 8B under GPT-4, *demonstrates the compatibility of EPPO with advanced reward modeling techniques, further enhancing RLHF performance*.

Method	Opponent	Anthropic-Helpful Win / Tie / Lose	Anthropic-Harmless Win / Tie / Lose	AlpacaFarm Win / Tie / Lose	TL;DR Summary Win / Tie / Lose
EPPO	ODIN	61% / 28% / 11%	44% / 24% / 32%	56% / 30% / 14%	36% / 42% / 21%
EPPO+ODIN	ODIN	67% / 25% / 8%	49% / 23% / 28%	59% / 29% / 12%	42% / 40% / 18%
EPPO	InfoRM	40% / 35% / 25%	48% / 28% / 24%	39% / 41% / 20%	38% / 51% / 11%
EPPO+InfoRM	InfoRM	46% / 33% / 21%	60% / 21% / 19%	43% / 38% / 19%	48% / 43% / 9%

Experiments

Final RLHF Performance

Table 3. Comparison of win, tie, and lose ratios between our `EPPO` and RLHF algorithms targeting reward hacking, evaluated on three additional representative LLMs under GPT-4, *demonstrates the consistently superior RLHF performance of* `EPPO` *across various LLMs*.

LLM	Opponent	Anthropic-Helpful			Anthropic-Harmless			AlpacaFarm			TL;DR Summary		
		Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Llama2	SFT	57%	26%	17%	65%	19%	16%	47%	34%	19%	78%	12%	10%
	PPO	65%	22%	13%	72%	17%	11%	54%	28%	18%	65%	24%	11%
	PPO w/ KL	52%	28%	20%	56%	23%	21%	41%	33%	26%	51%	31%	18%
	PPO w/ LP	59%	23%	18%	39%	36%	25%	49%	31%	20%	55%	28%	17%
Mistral	SFT	61%	22%	17%	56%	27%	17%	56%	32%	12%	76%	11%	13%
	PPO	48%	29%	23%	62%	23%	15%	50%	35%	15%	67%	16%	17%
	PPO w/ KL	38%	36%	26%	36%	30%	34%	41%	38%	21%	49%	30%	21%
	PPO w/ LP	45%	32%	23%	47%	27%	26%	44%	34%	22%	46%	32%	22%
DeepSeek	SFT	76%	15%	9%	75%	13%	12%	59%	27%	14%	72%	17%	11%
	PPO	72%	18%	10%	78%	12%	10%	56%	28%	16%	64%	23%	13%
	PPO w/ KL	59%	29%	12%	46%	31%	23%	47%	36%	17%	47%	36%	17%
	PPO w/ LP	66%	24%	10%	54%	28%	18%	52%	33%	15%	56%	29%	15%

Reward Hacking Mitigation

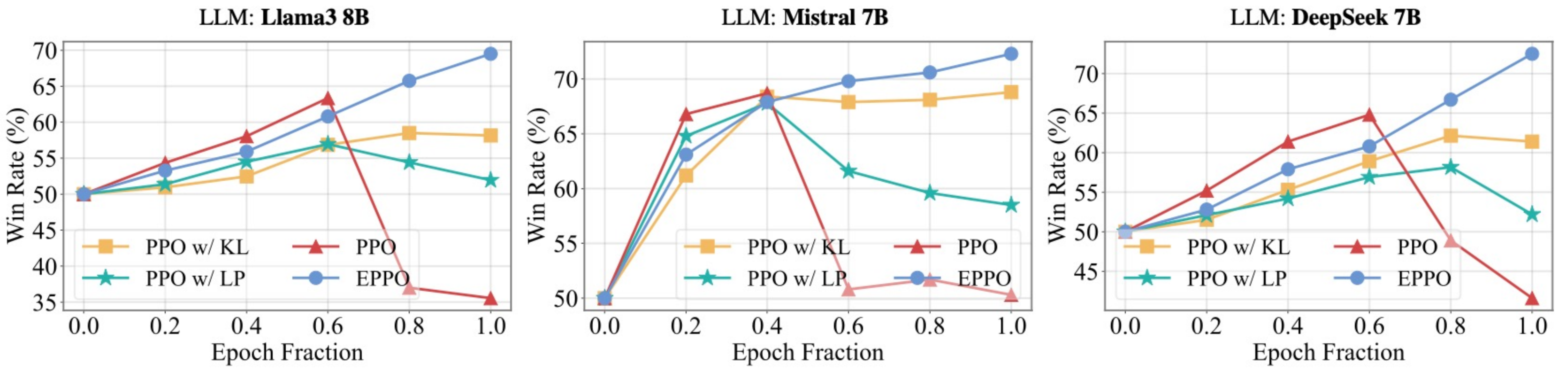


Figure 5. The win rate dynamics of various RLHF models compared to SFT model during RL training, under GPT-4 evaluation. We calculate the win rate as $win+0.5*tie$. **From left to right:** The LLMs shown are Llama3 8B, Mistral 7B, and DeepSeek 7B, respectively. *Observations: Comparison methods either suffer from limited performance gains via RL or experience significant degradation in the later stages—which indicates the onset of reward hacking. In contrast, our* `EPPO` *not only* ***enhances performance benefits from RL*** *but also* ***effectively mitigates reward hacking****, leading to a substantial boost in final RLHF performance.*

Thanks for Your Listening

Speaker: Miao Yuchun

Email: szmyc1@163.com