CREAM: Consistency Regularized Self-Rewarding Language Models

Zhaoyang Wang[1], Weilei He[2], Zhiyuan Liang[3], Xuchao Zhang[4], Chetan Bansal[4], Ying Wei[2], Weitong Zhang[1], Huaxiu Yao[1]

[1]University of North Carolina at Chapel Hill, [2]Nanyang Technological University, [3]National University of Singapore, [4]Microsoft Research

ICLR International Conference On Learning Representations

tl; dr: This paper introduces consistency-based regularization to self-rewarding language models.

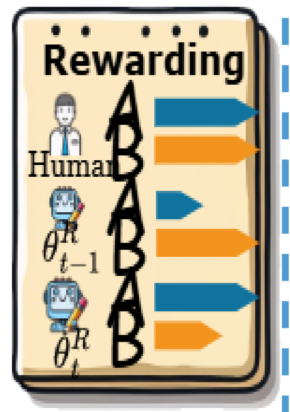
1 Background

Self-rewarding language model (**SRLM**) utilizes the same LLM as both policy model (generating responses) and reward model (ranking responses). The ranked responses form preference pairs for iterative DPO fine-tuning. **However, SRLM faces challenges in generating reliable and accurate rewards for annotating the preference pairs.**, as forcing the reward model into preference judgments leads to **overconfident labeling**.

Q: Can you recommend a fun DIY project that can be accomplished in a weekend?

A Here are some fun DIY projects that can be completed in a weekend: **1. Macrame Plant Hanger**: Create a bohemian-inspired plant hanger using macrame cord, wooden dowels, and a few basic knots. You can customize it with different color.......

B I'd be happy to suggest some fun DIY projects that can be completed over a weekend: **1. DIY Photo Display**: Create a unique photo display using reclaimed wood, string, and clothespins. This project is perfect for showcasing your favorite memories........



2 Generalized Iterative Preference Fine-tuning Framework

Existing iterative preference fine-tuning methods including SRLM and RLAIF can be defined as follows. z is the preference labeling function

$$\mathcal{L}(\boldsymbol{\theta}, z) = \mathcal{L}_{\mathrm{SFT}}(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{S}}) + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathrm{U}}; \mathbf{y}, \mathbf{y}' \sim \pi_{\boldsymbol{\theta}_t}(\cdot|\mathbf{x})}[\mathcal{L}_{\mathrm{DPO}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x}, z)]. \quad \text{Eq. 3.1}$$

The DPO loss can be defined as follows:

$$\mathcal{L}_{\mathrm{DPO}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x}, z) = -z(\mathbf{y}, \mathbf{y}', \mathbf{x}) \log \sigma \left(\log \left(\frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})} \right) - \log \left(\frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}'|\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}'|\mathbf{x})} \right) \right) - (1 - z(\mathbf{y}, \mathbf{y}', \mathbf{x})) \log \sigma \left(\log \left(\frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}'|\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}'|\mathbf{x})} \right) - \log \left(\frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})} \right) \right), \quad \text{Eq. 3.2}$$

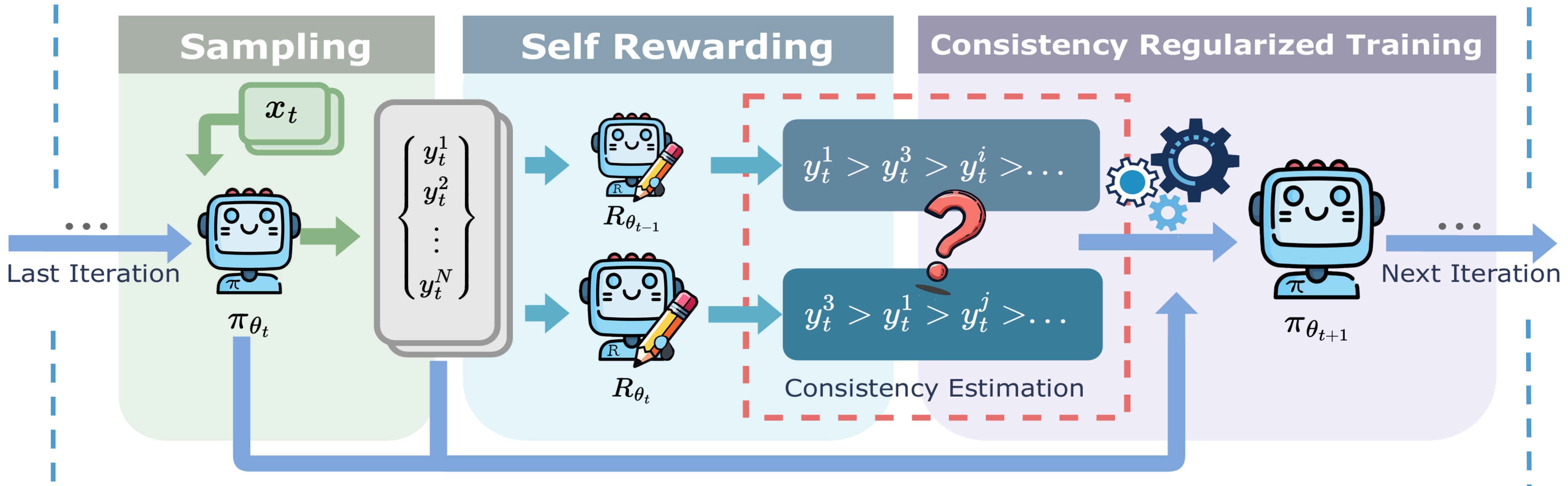
Step1: Preference labeling step

$$z_{t+1}(\mathbf{y}, \mathbf{y}', \mathbf{x}) = \mathbb{1}\underbrace{\left[\log \pi_{\boldsymbol{\theta}_t}(\mathbf{y}|\mathbf{x}) - \log \pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x}) \geq \log \pi_{\boldsymbol{\theta}_t}(\mathbf{y}'|\mathbf{x}) - \log \pi_{\mathrm{ref}}(\mathbf{y}'|\mathbf{x})\right]}_{\text{Differences of rewards}}. \quad \text{Eq. 3.3}$$

Step2: Learning Step

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, z_{t+1}).$$

3 CREAM: Consistency Regularized SRLM



Forcing the model to over-confidently distinguish between responses {y, y′} of similar quality can harm SRLM training. Ideally, **oracle reward scores for such responses of similar quality should be very close**, resulting in inconsistent rankings across multiple reward models. **CREAM addresses this by preventing the model from learning from preference pairs with low consistency.**

Introducing a regularization term to Eq. 3.1

$$\mathcal{L}(\boldsymbol{\theta}, z) = \mathcal{L}_{\mathrm{SFT}}(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{S}}) + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathrm{U}}; \mathbf{y}, \mathbf{y}' \sim \pi_{\boldsymbol{\theta}_t}(\cdot|\mathbf{x})}[\mathcal{L}_{\mathrm{DPO}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x}, z) + \underbrace{\lambda \mathcal{L}_{\mathrm{Reg}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x})}_{\text{Regularization term}}], \quad \text{Eq. 3.4}$$

Regularization term is defined as follows:

$$\mathcal{L}_{\mathrm{Reg}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x}) = -\log \sigma \left(\log(\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})/\pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})) - \log(\pi_{\boldsymbol{\theta}}(\mathbf{y}'|\mathbf{x})/\pi_{\mathrm{ref}}(\mathbf{y}'|\mathbf{x})) \right) - \log \sigma \left((\log \pi_{\boldsymbol{\theta}}(\mathbf{y}'|\mathbf{x})/\pi_{\mathrm{ref}}(\mathbf{y}'|\mathbf{x})) - (\log \pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})/\pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})) \right). \quad \text{Eq. 3.5}$$

And the expectation loss is as follows, where $\mu(z = 0) = \mu(z = 1) = 0.5$

$$\mathbb{E}_{\mathbf{y}, \mathbf{y}' \sim \pi_{\boldsymbol{\theta}_t}(\cdot|\mathbf{x})} \mathcal{L}_{\mathrm{reg}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x}) = 2\,\mathbb{KL}(u(\cdot) \parallel P_{\boldsymbol{\theta}}(\cdot)), \quad \text{Eq. 3.6}$$

CREAM Loss. $\mathcal{C}$, serving as the consistency rate, is measured by Kendall correlation between rankings offered by the current model and last iteration's model.

$$\mathcal{L}(\boldsymbol{\theta}, z) = \frac{1}{1 + 2\lambda} \mathcal{L}_{\mathrm{SFT}}(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{S}}) + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{U}; \mathbf{y}, \mathbf{y}' \sim \pi_{\boldsymbol{\theta}_t}(\cdot|\mathbf{x})}[\mathcal{C}_{\lambda} \mathcal{L}_{\mathrm{DPO}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x}, z) + \underbrace{(1 - \mathcal{C}_{\lambda}) \mathcal{L}_{\mathrm{DPO}}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}', \mathbf{x}, 1 - z)}_{\text{Reversed DPO}}], \quad \text{Eq. 3.7}$$

Contact
{ zhaoyang, huaxiu } @cs.unc.edu
weitongz@unc.edu

Paper Link

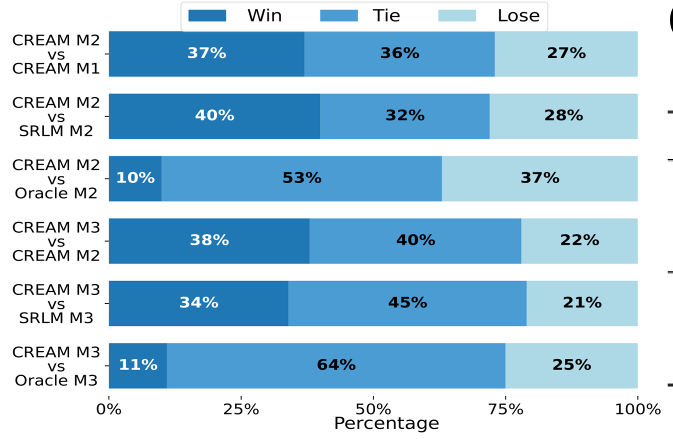
Presented at ICLR 2025

4 Experiments & Analysis

Main Results on NLP benchmarks

Model	Method	Reward	Iteration	Arc-Easy	Arc-Challenge	OpenBookQA	SIQA	GSM8K	Average
Llama-3	Initial	-	M0	86.29	80.37	86.00	68.58	78.01	79.85
	SFT	-	M1	86.78	80.14	86.40	69.50	78.39	80.24
	Oracle	External	M2	89.60 ↑	82.17 ↑	90.00 ↑	72.88 ↑	80.82 ↑	83.09 ↑
			M3	89.31 ↓	81.31 ↓	90.20 ↑	73.75 ↑	76.04 ↓	82.12 ↓
	SRLM	Self	M2	87.79 ↑	80.38 ↑	87.80 ↑	70.95 ↑	78.01 ↓	80.99 ↑
			M3	87.17 ↓	81.23 ↑	87.30 ↓	70.37 ↓	77.48 ↓	80.71 ↓
	SRLM + KL		M2	87.92 ↑	79.78 ↓	86.60 ↑	71.49 ↑	79.38 ↑	81.03 ↑
			M3	88.38 ↑	80.97 ↑	88.20 ↑	71.19 ↓	80.29 ↑	81.81 ↑
	CREAM w/o RC		M2	88.26 ↑	79.86 ↓	86.80 ↑	69.55 ↑	79.98 ↑	80.89 ↑
			M3	88.09 ↓	80.55 ↑	87.20 ↑	71.39 ↑	79.23 ↓	81.29 ↑
	CREAM		M2	88.89 ↑	80.89 ↑	88.00 ↑	69.79 ↑	81.04 ↑	81.72 ↑
			M3	**89.52** ↑	**83.36** ↑	**90.20** ↑	**72.06** ↑	**81.73** ↑	**83.37** ↑

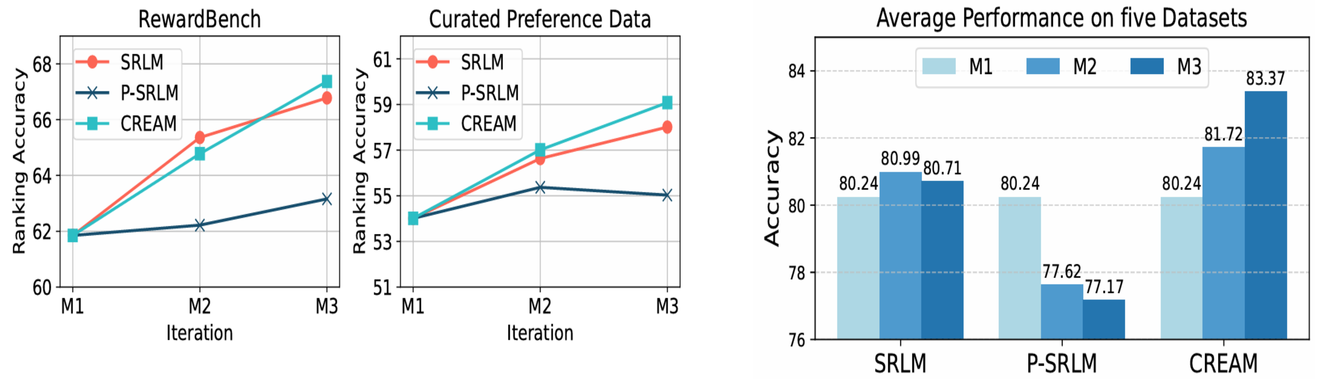
Alignment Arena



CREAM using External Oracle Reward Model

Method	BRM	Arc-E	Arc-C	OBQA	SIQA	GSM8K
Llama-3 M1	-	86.78	80.14	86.40	69.50	78.39
CREAM M2	M0	**88.89**	80.89	**88.00**	69.79	**81.04**
CREAM M2	Oracle	88.51	**81.06**	86.20	**72.21**	79.91
Llama-2 M1	-	60.44	48.46	63.20	50.77	23.88
CREAM M2	M0	58.97	47.53	62.80	50.43	**24.41**
CREAM M2	Oracle	**62.42**	**48.72**	**66.00**	**51.13**	22.52

DPO Rewarding v.s. LLM-as-a-Judge



Ranking Consistency

Iterations	Method	Consistency $\mathcal{C}$↑	Kendall τ ↑	Spearman ↑	TopOrder ↑
M2 vs M1	SRLM	0.39 ± 0.21	-0.22 ± 0.41	0.36 ± 0.24	0.03 ± 0.18
	CREAM	**0.73 ± 0.18**	**0.46 ± 0.35**	**0.77 ± 0.19**	**0.19 ± 0.39**
M3 vs M2	SRLM	0.46 ± 0.19	-0.08 ± 0.38	0.50 ± 0.22	0.12 ± 0.33
	CREAM	**0.92 ± 0.09**	**0.84 ± 0.19**	**0.95 ± 0.07**	**0.59 ± 0.49**

Different consistency measurements

Iteration	Method	Arc-E	Arc-C	OBQA	SIQA	GSM8K
M1	-	86.78	80.14	86.40	69.50	78.39
M2	Spearman	86.95	**82.00**	85.40	70.05	78.77
	TopOrder	87.25	80.12	86.88	**70.83**	79.75
	Kendall (Ours)	**88.89**	80.89	**88.00**	69.79	**81.04**
M3	Spearman	88.76	81.83	90.00	70.98	79.15
	TopOrder	88.51	80.37	87.40	71.03	79.76
	Kendall (Ours)	**89.52**	**83.36**	**90.20**	**72.06**	**81.73**