Getting More Juice Out of the SFT Data: Reward Learning from Human Demonstration Improves SFT for LLM Alignment

Jiaxiang Li[†], Siliang Zeng[†], Hoi-To Wai[‡], Chenliang Li[⋆], Alfredo Garcia[⋆], Mingyi Hong[†]

[†]University of Minnesota [⋆]Texas A&M University [‡]Chinese University of Hong Kong

Presenting at the
NeurIPS 2024, Vancouver, Canada

Introduction

- **Alignment for LLMs** aims to align the pre-trained LLMs with prepared human-labeled data to achieve desired performance over certain tasks
- Motivated by the success of RLHF, we pose the following question:

  Does building a reward model using the demonstration data benefit the alignment process?

Our Contributions

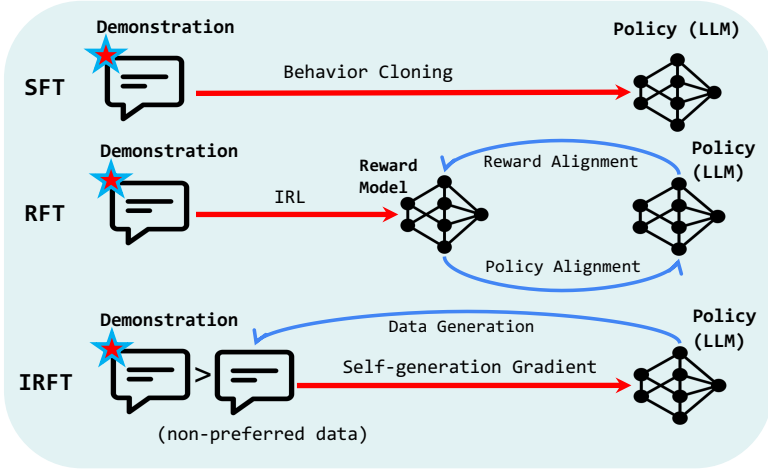


Figure 1: Difference between SFT and the two proposed methods: RFT and IRFT

Problem Formulation

Given a fixed demonstration dataset $\mathcal{D} := \{(x, y)\}$, via a maximum likelihood inverse reinforcement learning (ML-IRL) formulation:

$$\begin{aligned} &\max_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}) := \mathbb{E}_{x\sim\rho, y\sim\pi^{\mathrm{E}}(\cdot|x)} \left[\log \pi_{\boldsymbol{\theta}}\left(y \mid x\right)\right] \\ &\quad \text{s.t. } \pi_{\boldsymbol{\theta}} := \arg\max_{\pi} \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x)} \left[r\left(x, y; \boldsymbol{\theta}\right) - \beta D_{\mathrm{KL}}\Big(\pi(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)\Big)\right]. \end{aligned} \tag{1}$$

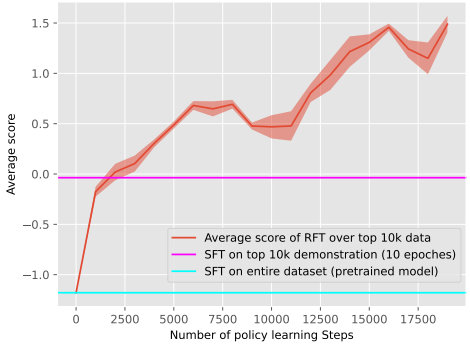
A **self-play gradient**:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \frac{1}{\beta} \mathbb{E}_{x\sim\rho, y\sim\pi^{\mathrm{E}}(\cdot|x), \widetilde{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|x)} \left[\nabla_{\boldsymbol{\theta}} \log \frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \nabla_{\boldsymbol{\theta}} \log \frac{\pi_{\boldsymbol{\theta}}(\widetilde{y}|x)}{\pi_{\mathrm{ref}}(\widetilde{y}|x)}\right]. \tag{2}$$

Experiment Result

Result on Explicit reward learning: We test the performance of our proposed algorithm on Anthropic-HH dataset and pythia-1.4B model.

(a) Log prob gap



(b) Average Score

Figure 2: Fine-tuning result of `pythia-1.4b` over Anthropic-HH (with top 10k data picked by `PKU-Alignment/beaver-7b-v3.0-reward`) where we do training with the chosen response only.

Experiment Result

Result on Implicit reward learning: We test the performance of our proposed algorithm on Ultrachat-200k dataset and Zephyr-7b model, evaluating on downstream tasks from the HuggingFace Open LLM Leaderboard datasets.

Tasks Metrics	T	K	AI2_Arc acc_norm	TruthfulQA acc	Winogrande acc	GSM8k exact_match	HellaSwag acc_norm	MMLU acc	Average
`zephyr-7b-sft-full`	0	0	74.83	34.07	76.09	31.92	81.09	**58.86**	59.48
`IRFT` (SPIN iter 0)	1	$\frac{\#\text{ samples}}{\text{batchsize}} * 2$	75.08	36.57	76.01	33.59	82.81	57.83	60.32
`IRFT` (SPIN iter 1)	2	$\frac{\#\text{ samples}}{\text{batchsize}} * 2$	76.13	36.56	76.64	**35.56**	**83.39**	57.82	61.02
`IRFT`	5	$\frac{\#\text{ samples}}{\text{batchsize}} * \frac{2}{5}$	75.82	**39.99**	77.19	31.24	82.07	57.93	60.71
`IRFT`	10	$\frac{\#\text{ samples}}{\text{batchsize}} * \frac{2}{5}$	**76.78**	36.84	**77.43**	34.34	83.05	57.72	**61.03**
`IRFT`	8	$\frac{\#\text{ samples}}{\text{batchsize}} * \frac{2}{8}$	75.23	36.67	75.85	31.84	80.89	58.60	59.85
`IRFT`	16	$\frac{\#\text{ samples}}{\text{batchsize}} * \frac{2}{8}$	75.79	35.55	76.56	32.52	82.3	58.77	60.25