Aligning Language Models with Generated feedback: Self-reward vs. LLM guided

Ziming Luo, Jiaming Tang, Tianxing Ma, Tianxu Jiang, Yichen Lu

EECS 598 Foundation of Large Language Models

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

INTRODUCTION

We explores the nature of the **self-reward RLHF** and introduces a novel framework based on online supervision from large language models, which significantly improves the alignment training for smaller models. Our work aims to address the the issue of manual data annotation dependency inherent in RLHF, thereby enhancing the generalization capabilities of small models. Additionally, it provides support for future offline RLHF work, primarily driven by Direct Preference Optimization (DPO)

And Our contribution is as follows:

- Implemented a self-reward RLHF platform with strong expandability
- Systematically evaluated the fine tuned models for both Instruction Following Ability and Reward Modeling Ability.
- Introduced Gemini-based LLM-as-a-Judge and LLM-as-a-Teacher paradigms

Our code is released in github:
https://github.com/jmtitan/Self-Reward-RLHF.git

To perform AlpacaEval:
https://github.com/TnTerry/alpaca_eval.git

DATASETS

Seed Training Data

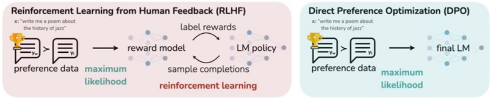
Instruction Fine-Tuning (IFT) data: Consist of 3190 high quality examples, specifically those human ranked highest (rank 0) in English.

Evaluation Fine-Tuning (EFT) data: 2934 evaluation examples with Multiple human ranked responses.

Both derived from the Open Assistant dataset [Köpf et al., 2023].

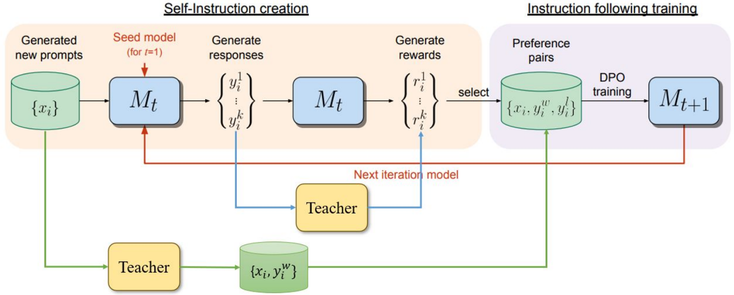
METHODOLOGY

DPO Introduction



$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

Self Reward and Online Feedback Framework



M_0 : Base pretrained LLM with no fine-tuning.
M_1 : Initialized with M_0, then fine-tuned on the IFT+EFT seed data using SFT.
M_2 : Initialized with M_1, then trained with self reward (M_1) data using DPO.
M_2 (Teacher reward): Initialized with M_1, then trained with teacher reward data using DPO.
M_2 (Teacher demonstration): Initialized with M_1, then trained with teacher demonstration data using DPO.

EXPERIMENT

Test model

Base model: Phi-2, an open-source small language model with 2.7 billion parameters that exhibits excellent reasoning and language understanding abilities.

Teacher model: Gemini Pro 1.0, a multimodal model with **160** billion parameters that are able to process information from multiple modalities, including images, video, and text.

Evaluation

Instruction Following Ability: head-to-head win rates on diverse prompts using GPT-4 and AlpacaEval 2.0 (win rate over GPT-4 Turbo evaluated by GPT-4).

Reward Modeling Ability: Pairing accuracy, best 5, exact match is to measure how much the ranking of the model assessment and the GPT4 ranking agreed. The Spearman correlation and Kendall τ is to measure the similarity.

Results

Instruction Following

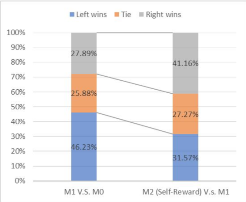


Table 1: AlpacaEval 2.0 Results

Model	Win Rate
M0 (Phi-2)	2.4%
M1 (IFT+EFT)	2.7%
GPT-4 Turbo (04/09)	46.1%
Claude 3 Opus (02/29)	29.1%
Gemini Pro	16.9%
Vicuna 33B v1.3	12.7%
LLaMA2 Chat 13B	7.7%
Vicuna 7B v1.5	4.8%
Davinci001	2.8%
Alpaca 7B	2.6%
Falcon 7B Instruct	2.1%

Reward Modeling

Model	SFT(M1)	Self-Reward(M2)	Teacher-Reward(M_2)	Teacher-Demo.(M_2)
Training data	IFT+EFT	AIFT	AIFT	AIFT
Pairwise accuracy	38.3%	3.75%	35.5%	
5-best %	57.10%	64.70%	73.2%	
Exact Match %	3.0%	0.0%	35.0%	
Spearman corr.	0.103	0.041	0.340	
Kendall τ corr.	0.090	0.037	0.302	

Table 1: Reward Modeling Ability

Conclusion

1. Small LLMs will be adversely affected with self-reward RLHF. The reasons are likely to be:
   a. Unable to fit human preferences through a small amount of sft data
   b. The generated answers are not stable enough and may appear to be self-questioning and irrelevant.
2. Teacher reward helps small model generates more human-like preferences, but the improvement is limited.
3. Teacher demonstration guided fine tuning achieves the best result, suggesting that response quality is the most important factor in improving RLHF.

Discussion

1. We would like to validate the "scaling laws" both for more iterations and larger dataset.
2. We hope this work will improve the performance of offline RL algorithms such as Decision Transformer in human alignment.