Self-improvement of LLMs via synthetic data

Haoyan Yang

AI Center (Dialogue and Knowledge Group) Intern

Introduction

During my internship at SRA as a Researcher Intern, I focused on advancing Large Language Models (LLMs) through self-improvement, particularly by exploring the novel Self-Play Fine-Tuning (SPIN) approach. Self-improvement involves fine-tuning a model after supervised fine-tuning (SFT) using data that the model generates itself, aiming to achieve better performance. My research specifically explored the limitations of the SPIN method, analyzing the impact of these limitations and working to find solutions to address them.

Self-Play Fine-Tuning (SPIN)

Algorithm 1 Self-Play Fine-Tuning (SPIN)

Input: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[N]}$: SFT Dataset, $p_{\boldsymbol{\theta}_0}$: LLM with parameter $\boldsymbol{\theta}_0$, T: Number of iterations.
for $t = 0, \ldots, T-1$ **do**
  for $i = 1, \ldots N$ **do**
    Generate synthetic data $\mathbf{y}_i' \sim p_{\boldsymbol{\theta}_t}(\cdot|\mathbf{x}_i)$.
  end for
  Update $\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i\in[N]} \ell\Big(\lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_i|\mathbf{x}_i)}{p_{\boldsymbol{\theta}_t}(\mathbf{y}_i|\mathbf{x}_i)} - \lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_i'|\mathbf{x}_i)}{p_{\boldsymbol{\theta}_t}(\mathbf{y}_i'|\mathbf{x}_i)}\Big)$.
end for
Output: $\boldsymbol{\theta}_T$.

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

Experimental Setting

Fine-tuned dataset

- Ultrachat200k

Fine-tuned Models

- Zephyr-7B-sft-full (SFT based on pre - trained model Mistral-7B)
- Llama-2-7b-ultrachat200k (SFT based on pre - trained model Llama-2-7B)

Evaluation Benchmark

- arc-challenge, easy(25), truthfulqa-mc1,mc2(0), winogrande (5), gasm8k(5), hellaswag(10), mmlu(5)

SPIN Results - Zephyr-7B-sft-full

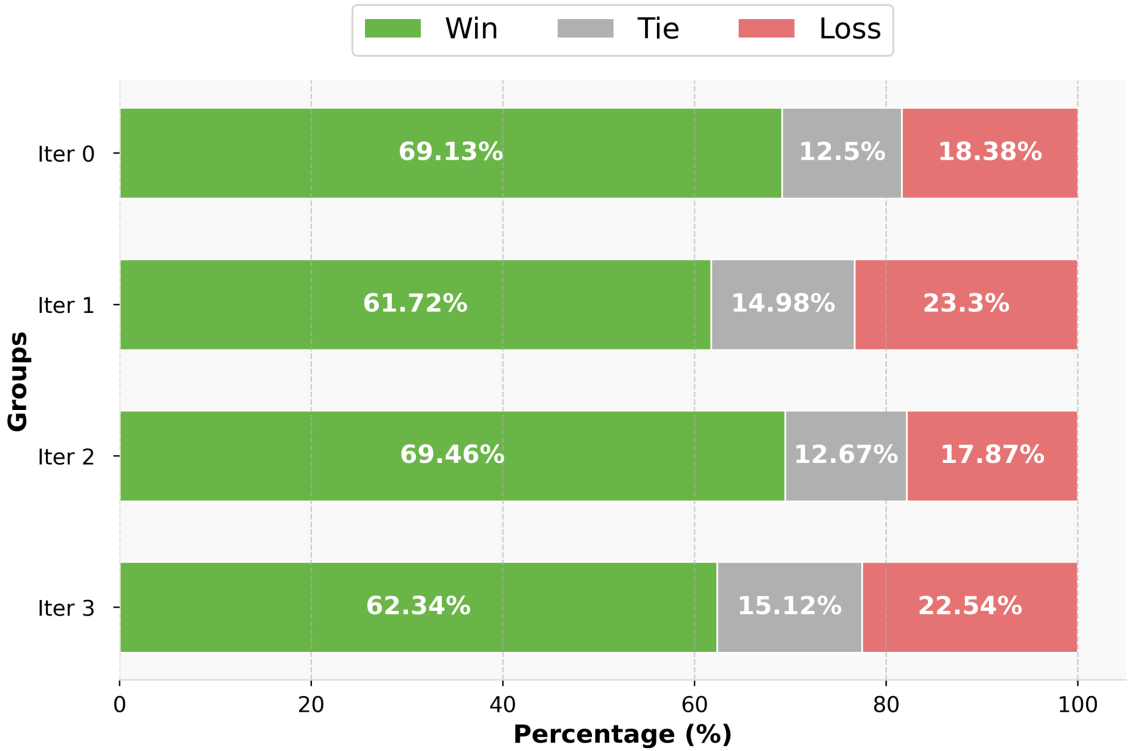
Task	arc-challenge (25)	arc-easy(25)	truhfulqa-mc1(0)	truhfulqa-mc2(0)	winogrande(5)	gsm8k(5)	hellaswag(10)	mmlu(5)	Average
SFT	0.5708	0.8375	0.2778	0.4038	0.7616	0.3184	0.8102	0.5877	0.5710
SPIN-iter0	0.5922	0.8266	0.3244	0.4615	0.7680	0.2889	0.8260	0.5901	**0.5847**
SPIN-iter1	0.5853	0.8203	0.2901	0.4341	0.7601	0.3161	0.8172	0.5846	0.5760
SPIN-iter2	0.5904	0.8241	0.3072	0.4328	0.7609	0.2760	0.8197	0.5850	0.5745
SPIN-iter3	0.5819	0.8245	0.3146	0.4515	0.7561	0.2752	0.8181	0.5786	0.5751

SPIN Results - Llama-2-7b-ultrachat200k

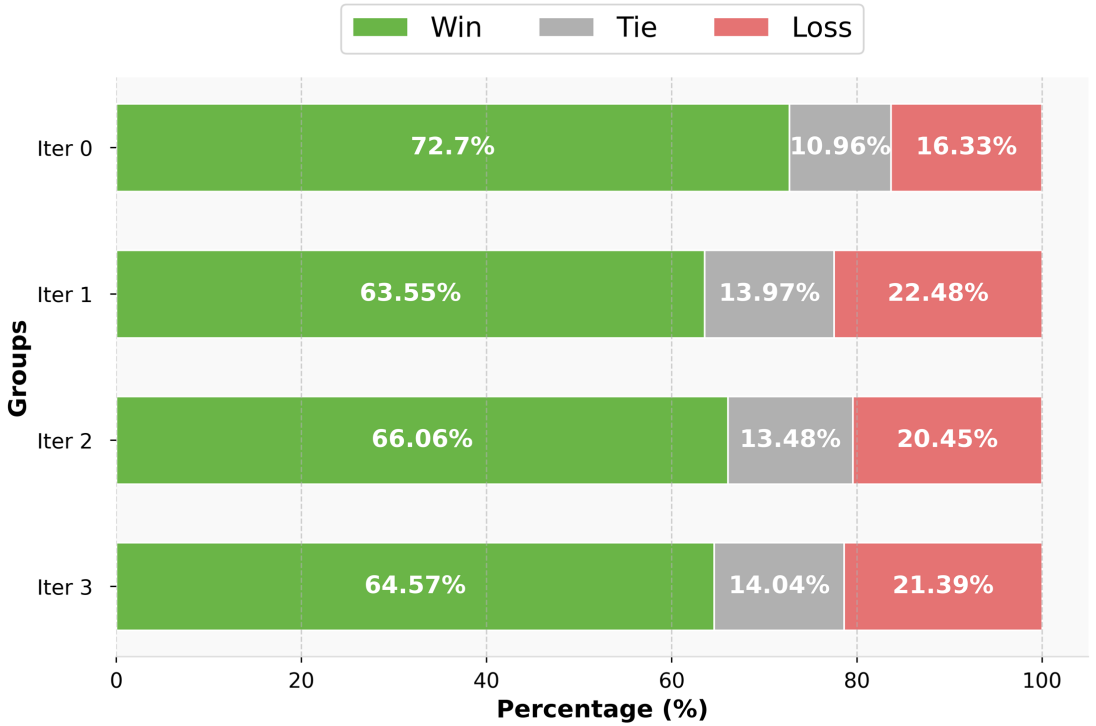
Task	arc-challenge (25)	arc-easy(25)	truhfulqa-mc1(0)	truhfulqa-mc2(0)	winogrande(5)	gsm8k(5)	hellaswag(10)	mmlu(5)	Average
SFT	0.5290	0.8253	0.3121	0.4494	0.7230	0.1372	0.7619	0.4479	0.5232
SPIN-iter0	0.5360	0.8291	0.3439	0.5055	0.7348	0.1516	0.7735	0.4478	**0.5403**
SPIN-iter1	0.5333	0.8312	0.3427	0.5066	0.7269	0.1706	0.7727	0.4509	0.5419
SPIN-iter2	0.5418	0.8325	0.3476	0.5086	0.7167	0.1592	0.7718	0.4524	0.5413
SPIN-iter3	0.5461	0.8329	0.3439	0.5078	0.7151	0.1577	0.7714	0.4511	0.5408

Labeling Issue

The loss function assumes that all SFT ground truth data is superior to the generated data, which could be an overly rigid assumption.



Zephyr-7B-sft-full



Llama-2-7b-ultrachat200k

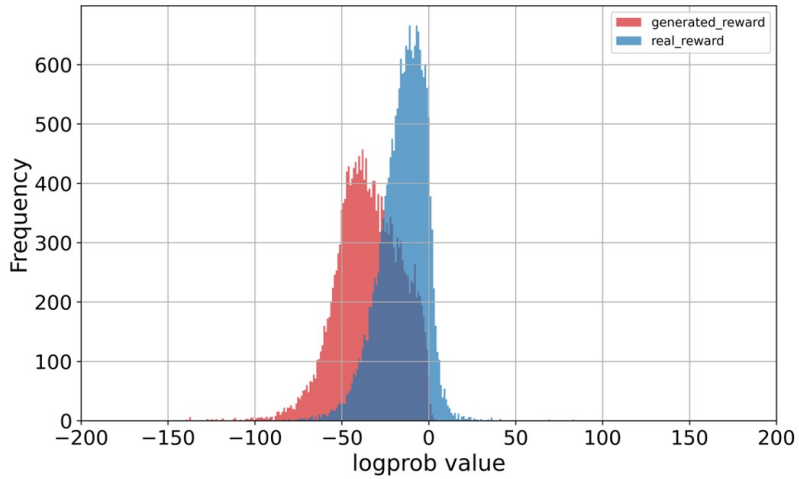
Solution for Labeling Issue

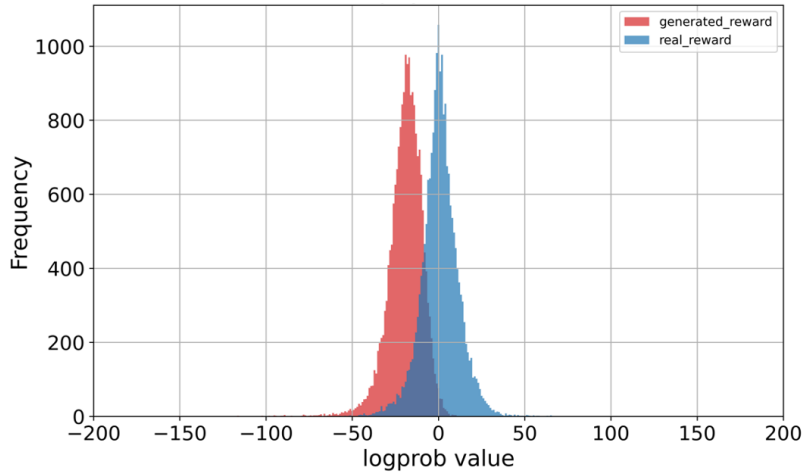
Use other powerful model (GPT4o-mini) to label the pair (real data and generated data) before every iteration

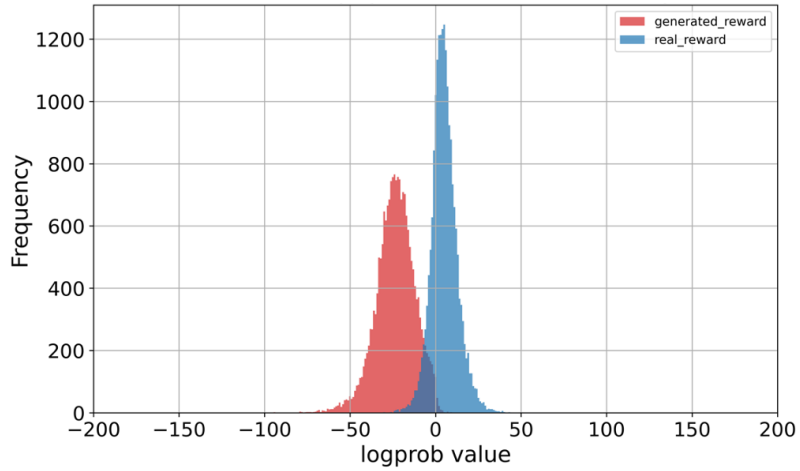
Task	arc-challenge (25)	arc-easy(25)	truhfulqa-mc1(0)	truhfulqa-mc2(0)	winogrande(5)	gsm8k(5)	hellaswag(10)	mmlu(5)	Average
Zephyr-SPIN-iter1	0.5853	0.8203	0.2901	0.4341	0.7601	0.3161	0.8172	0.5846	0.5760
GPT-Zephyr-SPIN-iter1	0.5939	0.8270	0.3133	0.4407	0.7672	0.3169	0.8229	0.5855	0.5834

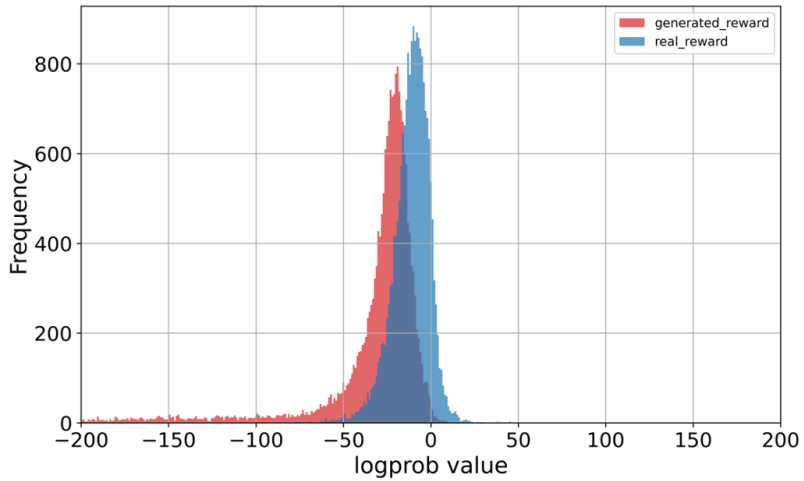
Reward Model Issue

Reward = $\log \frac{p_{\theta_{t+1}}(y_t|x)}{p_{\theta_t}(y_t|x)}$ where $y_t \sim \theta_t(\cdot\,|x)$









Solution for Reward Model Issue

Add noise to increase the randomness of model

Task	arc-challenge (25)	arc-easy(25)	truhfulqa-mc1(0)	truhfulqa-mc2(0)	winogrande(5)	gsm8k(5)	hellaswag (10)	mmlu(5)	Average
Zephyr-SPIN-iter1	0.5853	0.8203	0.2901	0.4341	0.7601	0.3161	0.8172	0.5846	0.5760
Noised-Zephyr-SPIN-iter1	0.5930	0.8258	0.3035	0.4469	0.7640	0.3275	0.8214	0.5880	0.5838

Future Work

Adaptive label and noise, then combine them together

If margin is big, we can set a threshold like the margin of score is larger than 5 and then we reverse these pairs to correct labels

If margin is small, we can add noise to increase the randomness to prevent from overfitting