

sDPO: Don’t Use Your Data All at Once

**Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim
Chanjun Park[†]**

Upstage AI, South Korea

{kdahyun, eddie, ynot, choco_9966, yoonsoo, limerobot, chanjun.park}@upstage.ai

Abstract

As development of large language models (LLM) progresses, aligning them with human preferences has become increasingly important. We propose stepwise DPO (sDPO), an extension of the recently popularized direct preference optimization (DPO) for alignment tuning. This approach involves dividing the available preference datasets and utilizing them in a stepwise manner, rather than employing it all at once. We demonstrate that this method facilitates the use of more precisely aligned reference models within the DPO training framework. Furthermore, sDPO trains the final model to be more performant, even outperforming other popular LLMs with more parameters.

1 Introduction

Large language models (LLMs) have revolutionized the field of natural language processing (NLP) through a training process that includes pre-training, supervised fine-tuning, and alignment tuning, with the latter ensuring the safety and usefulness of the model. Thus, reinforcement learning techniques (Christiano et al., 2017; Bai et al., 2022), such as proximal policy optimization (PPO) (Schulman et al., 2017), are key in this alignment phase, despite their complexity.

To address the complicated nature of reinforcement learning in LLM training, direct preference optimization (DPO) (Rafailov et al., 2023), among other methods (Yuan et al., 2023; Dong et al., 2023), have been popularized for its simplicity and effectiveness. DPO involves curating preference datasets using human or strong AI (*e.g.*, GPT-4 (OpenAI, 2023)) judgement to select chosen and rejected responses to questions. These datasets are used to train LLMs by comparing log probabilities of chosen versus rejected answers. However, obtaining these probabilities can be challenging with

Model	Reference Model	H4
Mistral-7B-OpenOrca	N/A	65.84
Mistral-7B-OpenOrca + DPO	SFT Base	68.87
Mistral-7B-OpenOrca + DPO	SOLAR-0-70B	67.86
Mistral-7B-OpenOrca + DPO	Intel-7B-DPO	70.13
OpenHermes-2.5-Mistral-7B	N/A	66.10
OpenHermes-2.5-Mistral-7B + DPO	SFT Base	68.41
OpenHermes-2.5-Mistral-7B + DPO	SOLAR-0-70B	68.90
OpenHermes-2.5-Mistral-7B + DPO	Intel-7B-DPO	69.72

Table 1: DPO results in terms of H4 scores for Mistral-7B-OpenOrca and OpenHermes-2.5-Mistral-7B with different reference models. The best results for each SFT base model are shown in bold.

proprietary models like GPT-4, since they do not offer log probabilities for inputs.

Thus, in most practical scenarios, the reference model is simply set as the base SFT model (Tunstall et al., 2023; Intel, 2023b; Ivison et al., 2023), which is a much weaker alternative with potentially misaligned preferences. This reference model acts as a *lower bound* in DPO, *i.e.*, the target model is optimized to be at least as aligned as the reference model. Thus, we argue that a reference model that is already more aligned will serve as a better lower bound for DPO training, which would be beneficial for the alignment tuning. One option would be to utilize the plethora of open source models (Tunstall et al., 2023; Ivison et al., 2023) that have already undergone alignment tuning.

Note that the above may not be feasible due to the absence of such aligned models, or the fact that it renounces control over the reference model, leading to safety concerns. Instead, we propose ‘stepwise DPO’, named sDPO, where we use the preference datasets (or subsets of a preference dataset) in a *step-by-step manner* when undergoing DPO training. The aligned model in the previous step is used as the reference model for the current step, which results in utilizing a more aligned reference model (*i.e.*, a better lower bound). Empirically, we show that using sDPO results in a more performant

[†] Corresponding Author



sDPO: Don't Use Your Data All at Once

**Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim
Chanjun Park[†]**

Upstage AI, South Korea

{kdahyun, eddie, ynot, choco_9966, yoonsoo, limerobot, chanjun.park}@upstage.ai

Abstract

大規模言語モデル(LLM)の開発が進むにつれて、人間の嗜好に合わせたアライメントがますます重要になってきています。我々は、アライメントチューニングのために、最近一般的になつた直接選好最適化(DPO)を拡張したステップワイズDPO(sDPO)を提案します。このアプローチでは、利用可能なプリファレンスデータセットを分割し、一度に利用するのではなく、段階的に利用します。この手法により、DPO学習フレームワーク内で、より正確にアライメントされた参照モデルの使用が容易になることを実証します。さらに、sDPOは最終モデルをより高性能に訓練し、より多くのパラメータを持つ他の一般的なLLMを凌駕することさえあります。

1 Introduction

大規模言語モデル(LLM)は、事前学習、教師あり微調整、アライメント調整を含む学習プロセスを通じて、自然言語処理(NLP)の分野に革命をもたらしました。したがって、近接政策最適化(PPO) (Schulman et al., 2017)における強化学習の複雑な性質に対処するために、直接選好最適化(DPO) (Rafailov et al., 2023)の判断を用いて、質問に対して選択された回答と拒否された回答を選択するために、好みのデータセットをキュレーションします。これらのデータセットは、選択された回答と拒否された回答の対数確率を比較することで、LLMを訓練するために使用されます。Mistral-7B-OpenOrca N/A 65.84 Mistral-7B-OpenOrca + DPO SFT Base 68.87 Mistral-7B-OpenOrca + DPO SOLAR-0-70B 67.86 Mistral-7B-OpenOrca + DPO Intel-7B-DPO 70.13 OpenHermes-2.5-Mistral-7B N/A 66.10 OpenHermes-2.5-Mistral-7B + DPO SFT Base 68.41 OpenHermes-2.5-Mistral-

Model	Reference Model	H4
7B + DPO SOLAR-0-70B	68.90	OpenHermes-2.5-Mistral-7B + DPO I
Intel-7B-DPO	69.72	

表1: Mistral-7B-OpenOrcaおよびOpenHermes-2.5-Mistral-7BのH4スコアのDPO結果（参照モデルの違いによる）。各SFTベースモデルの最良結果は太字で表示。

GPT-4のような独自モデルは、入力の対数確率を提供していないためです。したがって、ほとんどの実用的なシナリオでは、参照モデルは単純にベースとなるSFTモデル (Tunstall et al., 2023; Intel, 2023b; Ivison et al., 2023) に設定されます。この参照モデルはDPOの下界として機能します。つまり、ターゲットモデルは少なくとも参照モデルと同程度に整合するように最適化されます。つまり、ターゲットモデルは少なくとも参照モデルと同程度に整列するように最適化されます。したがって、より整列している参照モデルは、DPOのトレーニングにおいてより良い下界として機能し、整列のチューニングに有益であると主張します。一つの選択肢は、すでにアライメントチューニングが行われたオープンソースモデル (Tunstall et al., 2023; Ivison et al., 2023) を利用することです。このような整列されたモデルが存在しない、あるいは参照モデルの制御を放棄することになり安全性に懸念があるなどの理由から、上記は実現不可能である可能性があることに注意してください。その代わりに、我々はsDPOと名付けた「段階的DPO」を提案します。これは、DPO学習を行う際に、段階的にプリファレンスデータセット（またはプリファレンスデータセットのサブセット）を使用するものです。前のステップで整列されたモデルは、現在のステップの参照モデルとして使用され、その結果、より整列された参照モデル（すなわち、より良い下界）を利用することになります。経験的に、sDPOを使用することで、最終的な整列モデルの性能も向上することを示します。新しい嗜好データを生成する反復的なパイプライン

[†] Corresponding Author

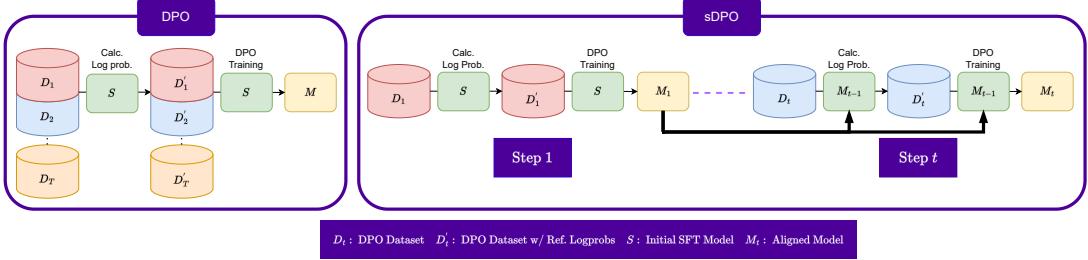


Figure 1: Overview of sDPO where preference datasets are divided to be used in multiple steps. The aligned model from the previous step is used as the reference and target models for the current step. The reference model is used to calculate the log probabilities and the target model is trained using the preference loss of DPO at each step.

final aligned model as well.

While concurrent works (Yuan et al., 2024) that focus on an iterative pipeline of generating *new* preference data have been proposed, our method focuses on utilizing the *currently available* preference datasets. Thus, our approach is complementary as sDPO can be easily applied to any preference data and further combination with concurrent works would be an exciting future direction.

2 Methodology

2.1 Preliminary Investigation on Reference Models

To gauge the importance of using a well-aligned reference model in DPO, we perform preliminary experiments of DPO training with the Ultrafeedback dataset (Cui et al., 2023) on Mistral-7B-OpenOrca (Lian et al., 2023) and OpenHermes-2.5-Mistral-7B (Teknium, 2023) as the SFT base model, owing to their excellent performance and small size. We compare the following reference models: i) the SFT base model itself, same as the conventional DPO setup; ii) SOLAR-0-70B (Upstage, 2023), a larger and much more performant model; and iii) Intel-7B-DPO (Intel, 2023a), an already aligned reference model. The results are summarized in Tab. 1.

As the table shows, using Intel-7B-DPO as the reference model results in the best performance, even better than using SOLAR-0-70B, which is a much larger model that was trained with more data. Thus, whether the reference model is pre-aligned or not plays an important role in the resulting aligned model’s performance. Unfortunately, it is not always possible to simply use a open sourced pre-aligned model as the reference model due to technical and safety concerns, *i.e.*, such a model may not exist yet or can be susceptible to various domain-specific harmfulness and fairness criteria.

To remedy the above, we propose sDPO, which uses more aligned reference models as a part of the training framework.

2.2 Stepwise DPO

In sDPO, we propose to use the available preference datasets in a stepwise manner instead of using them all at once. The comparison of the overall flow of DPO and sDPO is presented in Fig. 1.

Reference model. The reference model is used to calculate the log probabilities of the preference dataset. For each step, only a subset of the total data is used and the reference model is initialized as M_{t-1} , *i.e.*, the aligned model from the previous step. The initial reference model is set as S , the SFT base model. This results in using a more aligned reference model than conventional DPO.

Target model. For $t > 1$, the target model which is trained using the preference loss of DPO in each step of sDPO is also initialized as M_{t-1} instead of S . This ensures that the final model trained with sDPO has been directly trained with the same amount data as a model trained with DPO.

Intuitive explanation. To gain a deeper understanding of sDPO, we rearrange the DPO loss from (Rafailov et al., 2023), as follows:

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) &= -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \\ &= -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma (\beta \cdot (\gamma_\pi(x, y_w, y_l) - \gamma_{\pi_{\text{ref}}}(x, y_w, y_l)))] \end{aligned} \quad (1)$$

where \mathcal{D} is the preference dataset, x is the question, y_w and y_l are the chosen and rejected answers respectively, θ is the learnable parameters of the model, and $\gamma_\pi(x, y_w, y_l) = \log \frac{\pi(y_w|x)}{\pi(y_l|x)}$, *i.e.*, the logratio of the chosen and rejected samples w.r.t. the policy π . As $\log \sigma(\cdot)$ is a monotonically increasing function and $\gamma_{\pi_{\text{ref}}}$ is fixed before training, the minimization of $\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}})$ leads to

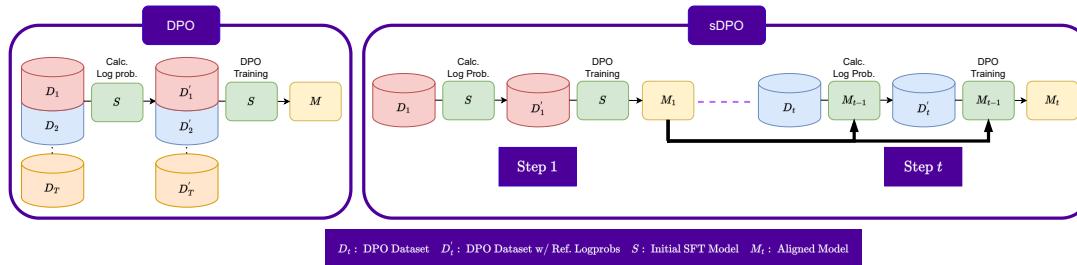


図1：嗜好データセットを分割して複数のステップで使用するsDPOの概要。前のステップで整列されたモデルが、現在のステップの参照モデルおよびターゲットモデルとして使用されます。参照モデルは対数確率の計算に使用され、ターゲットモデルは各ステップでDPOの嗜好損失を使用して学習されます。

に焦点を当てた並行研究(Yuan et al. , 2024)も提案されていますが、我々の手法は、現在利用可能な嗜好データセットを利用することに焦点を当てています。したがって、sDPOはどのような嗜好データにも容易に適用できるため、我々のアプローチは補完的であり、さらに同時並行的な研究との組み合わせは、将来的にエキサイティングな方向となるでしょう。

このような状況を改善するために、私たちはsDPOを提案します。sDPOでは、学習フレームワークの一部として、より整列された参照モデルを使用します。

2.2 Stepwise DPO

sDPOでは、利用可能なプリファレンスデータセットを一度に利用するのではなく、段階的に利用することを提案します。DPOとsDPOの全体的な流れの比較を図1に示します。

参照モデル。 参照モデルは、選好データセットの対数確率を計算するために使われます。各ステップでは、全データのサブセットのみが使用され、参照モデルはM_{t-1}として初期化されます。初期参照モデルはSFTベースモデルのSとします。この結果、従来のDPOよりも整列された参照モデルが使用されます。

ターゲットモデル。 t > 1の場合、sDPOの各ステップでDPOのプリファレンスロスを使用して学習されるターゲットモデルも、SではなくM_{t-1}として初期化されます。これにより、sDPOで学習された最終モデルは、DPOで学習されたモデルと同量のデータで直接学習されたことになります。

直感的な説明sDPOをより深く理解するために、(Rafailov et al. , 2023)のDPO損失を以下のように並べ替えます：

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) \\ = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (1) \\ = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \log \sigma \left(\beta \cdot (\gamma_{\pi_\theta}(x, y_w, y_l) - \gamma_{\pi_{\text{ref}}}(x, y_w, y_l)) \right), \end{aligned}$$

ここで、Dは選好データセット、xは質問、y_wとy_lはそれぞれ選ばれた答えと拒否された答え、θはモデルの学習可能なパラメータ、γ_π(x, y_w, y_l) = log_{y w | x} { π(x | y_l | x) } 、すなわち、ポリシーπに対する選ばれたサンプルと拒否されたサンプルの対数比。log σ(-)は単調増加関数であり、γ_πは学習前に固定されるため、L_{DPO}(π_θ, π_{ref})の最小化により

2 Methodology

2.1 Preliminary Investigation on Reference Models

DPOにおいて整合性の高い参照モデルを使用することの重要性を測るために、Mistral-7B- OpenOrca (Lian et al. , 2023)と OpenHermes- 2.5-Mistr al-7B (Teknium , 2023)をSFTベースモデルとして、Ultrafeed- backデータセット (Cui et al . , 2023) を用いたDPO学習の予備実験を行いました。i) 従来のDPOセットアップと同じSFTベースモデル、ii) より大型で性能の高いSOLAR-0-70B (Up-stage , 2023)、iii) すでに整合済みのリファレンスモデルであるIntel-7B-DPO (Intel , 2023a)。結果は表1にまとめられています。1. この表が示すように、Intel-7B-DPOを参照モデルとして使用した場合、SOLAR-0-70Bを使用した場合よりも、より多くのデータで学習された、より大きなモデルを使用した場合よりも、最高のパフォーマンスを得ることができました。このように、参照モデルが事前にアライメントされているかどうかが、アライメントされたモデルの性能に重要な役割を果たします。つまり、そのようなモデルはまだ存在しないかもしれませんし、様々なドメイン特有の有害性や公平性の基準の影響を受ける可能性があります。

Model	Size	Type	H4 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA
SOLAR 10.7B + SFT + sDPO	~ 11B	Alignment-tuned	74.31	71.33	88.08	65.39	72.45
SOLAR 10.7B + SFT + DPO	~ 11B	Alignment-tuned	72.67	69.62	87.16	66.00	67.90
SOLAR 10.7B + SFT + sDPO Strat.	~ 11B	Alignment-tuned	72.56	69.20	87.27	65.96	67.81
Mixtral 8x7B-Instruct-v0.1	~ 47B	Alignment-tuned	73.40	70.22	87.63	71.16	64.58
SOLAR-0-70B-16bit	~ 70B	Instruction-tuned	72.93	71.08	87.89	70.58	62.25
Qwen 72B	~ 72B	Pretrained	72.17	65.19	85.94	77.37	60.19
Yi 34B	~ 34B	Pretrained	70.72	64.59	85.69	76.35	56.23
SOLAR 10.7B + SFT	~ 11B	Instruction-tuned	69.51	67.32	85.96	65.95	58.80
Mistral 7B-Instruct-v0.2	~ 7B	Instruction-tuned	69.27	63.14	84.88	60.78	68.26
Falcon 180B	~ 180B	Pretrained	68.57	69.45	88.86	70.50	45.47
Mixtral 8x7B-v0.1	~ 47B	Pretrained	67.78	66.04	86.49	71.82	46.78
Llama 2 70B	~ 70B	Pretrained	67.35	67.32	87.33	69.83	44.92
Zephyr	~ 7B	Alignment-tuned	66.36	62.03	84.52	61.44	57.44
Qwen 14B	~ 14B	Pretrained	64.85	58.28	83.99	67.70	49.43
SOLAR 10.7B	~ 11B	Pretrained	64.27	61.95	84.60	65.48	45.04
Mistral 7B	~ 7B	Pretrained	62.40	59.98	83.31	64.16	42.15

Table 2: Performance comparison of applying sDPO (and ablated versions) to SOLAR 10.7B + SFT against various top performing models. Size is shown in units of billions of parameters and type is reported as one of {‘Pretrained’, ‘Instruction-tuned’, ‘Alignment-tuned’}. Models based on SOLAR 10.7B are shown in purple color. The best scores in each column are shown in bold.

$\gamma_{\pi_\theta} > \gamma_{\pi_{ref}}$ (on average). Thus, $\gamma_{\pi_{ref}}$ can be understood as a lower bound defined by the reference model, of which the target model is trained such that $\gamma_{\pi_\theta} > \gamma_{\pi_{ref}}$. In sDPO, $\gamma_{\pi_{ref}}$ increases as the steps progress because the reference model that defines it is more and more aligned. Hence, $\gamma_{\pi_{ref}}$ becomes a stricter lower bound as the steps pass, inducing a curriculum learning from easy to hard optimization tasks.

3 Experiments

3.1 Experimental Setup

Training details. We use a supervised fine-tuned SOLAR 10.7B (Kim et al., 2023) as our SFT base model S as it delivers excellent performance with its uncommon 10.7B size. Further, the scarcity of 10.7B sized models leads to the absence of open source models that can be adopted as reference models, making the usage of sDPO more necessary. We use OpenOrca (Mukherjee et al., 2023) ($\sim 12K$ samples) and Ultrafeedback Cleaned ($\sim 60K$ samples) (Cui et al., 2023; Ivison et al., 2023) as our preference datasets. The training hyper-parameters closely follow that of Tunstall et al. (2023). We use two steps in sDPO, where we use OpenOrca as dataset D_1 in the first step and Ultrafeedback Cleaned as dataset D_2 in the second step.

Evaluation. We utilize four of the six tasks in the HuggingFace Open LLM Leaderboard (Beeching et al., 2023): ARC (Clark et al., 2018), HellaSWAG (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), TruthfulQA (Lin et al., 2022). We also report the average scores for the four tasks, which is denoted as H4. Winogrande (Sakaguchi

et al., 2021) and GSM8K (Cobbe et al., 2021) are excluded to control the complexity of the experiments, *i.e.*, we excluded generation tasks in contrast to multiple choice tasks.

3.2 Main Results

Evaluation results for applying sDPO to the SFT base model, along with results for other top-performing models are shown in Tab. 2. Comparing the pretrained-only ‘SOLAR 10.7B’ to the instruction-tuned ‘SOLAR 10.7B + SFT’, we can see an increase of +5.24 in terms of H4. Applying sDPO on SOLAR 10.7B + SFT further increases the H4 score upto 74.31, an improvement of +4.80. Notably, ‘SOLAR 10.7B + SFT + sDPO’ outperforms other larger models such as Mixtral 8x7B-Instruct-v0.1, despite the smaller number of parameters. This highlights that effective alignment tuning could be the key to unlocking next level performance for smaller LLMs. Further, applying sDPO results in substantially higher score of 72.45 for TruthfulQA, which shows the effectiveness of the alignment tuning process.

3.3 Ablation Studies

We also report evaluation results for ablated models in Tab. 2. ‘SOLAR 10.7B + SFT + DPO’ uses all the DPO data at once, *i.e.*, $D_1 + D_2$, same as the conventional DPO training setup. ‘SOLAR 10.7B + SFT + sDPO Strat.’ uses stratified sampling to sample $\sim 16.67\%$ of the data points from the union of OpenOrca and Ultrafeedback Cleaned to form D_1 and use the remaining $\sim 83.33\%$ as D_2 to mirror the dataset size of D_1 and D_2 used in SOLAR 10.7B + SFT + sDPO.

Model	Size	Type	H4 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA
SOLAR 10.7B + SFT + sDPO	~ 11B	Alignment-tuned	74.31	71.33	88.08	65.39	72.45
SOLAR 10.7B + SFT + DPO	~ 11B	Alignment-tuned	72.67	69.62	87.16	66.00	67.90
SOLAR 10.7B + SFT + sDPO Strat.	~ 11B	Alignment-tuned	72.56	69.20	87.27	65.96	67.81
Mixtral 8x7B-Instruct-v0.1	~ 47B	Alignment-tuned	73.40	70.22	87.63	71.16	64.58
SOLAR-0-70B-16bit	~ 70B	Instruction-tuned	72.93	71.08	87.89	70.58	62.25
Qwen 72B	~ 72B	Pretrained	72.17	65.19	85.94	77.37	60.19
Yi 34B	~ 34B	Pretrained	70.72	64.59	85.69	76.35	56.23
SOLAR 10.7B + SFT	~ 11B	Instruction-tuned	69.51	67.32	85.96	65.95	58.80
Mistral 7B-Instruct-v0.2	~ 7B	Instruction-tuned	69.27	63.14	84.88	60.78	68.26
Falcon 180B	~ 180B	Pretrained	68.57	69.45	88.86	70.50	45.47
Mixtral 8x7B-v0.1	~ 47B	Pretrained	67.78	66.04	86.49	71.82	46.78
Llama 2 70B	~ 70B	Pretrained	67.35	67.32	87.33	69.83	44.92
Zephyr	~ 7B	Alignment-tuned	66.36	62.03	84.52	61.44	57.44
Qwen 14B	~ 14B	Pretrained	64.85	58.28	83.99	67.70	49.43
SOLAR 10.7B	~ 11B	Pretrained	64.27	61.95	84.60	65.48	45.04
Mistral 7B	~ 7B	Pretrained	62.40	59.98	83.31	64.16	42.15

表2 : SOLAR 10.7B + SFTにsDPO（およびアプレーションバージョン）を適用した場合の、様々なトップパフォーマンスマネルとの性能比較。サイズは数十億パラメータ単位で、タイプは{'Pretrained'、'Instruction-tuned'、'Alignment-tuned'}のいずれかで報告。SOLAR 10.7Bに基づくモデルは紫色で表示されます。各列のベストスコアは太字で示されています。

$\gamma \pi \theta > \gamma \pi \text{ref}$ (平均)。従って、 $\gamma \pi \text{ref}$ は、 $\gamma \pi \theta > \gamma \pi \text{ref}$ となるようにターゲットモデルを学習した参照モデルによって定義される下界と理解することができます。sDPOでは、 $\gamma \pi \text{ref}$ を定義する参照モデルのアライメントが進むため、 $\gamma \pi \text{ref}$ はステップが進むにつれて増加します。従って、 $\gamma \pi \text{ref}$ はステップが進むにつれて下界が厳しくなり、最適化が容易なタスクから困難なタスクへとカリキュラム学習が誘導されます。

et al. , 2021) と GSM8K (Cobbe et al. , 2021) は、実験の複雑さを制御するために除外しました。

3 Experiments

3.1 Experimental Setup

トレーニングの詳細。SFTのベースモデルとして、教師ありファインチューニングを施したSOLAR 10.7B (Kim et al. , 2023) を使用します。さらに、10.7Bサイズのモデルが少ないため、参照モデルとして採用できるオープンソースモデルが存在せず、sDPOを使用する必要性が高くなっています。我々はOpenOrca (Mukherjee et al. , 2023) (~ 12 K samples) とUltrafeedback Cleaned (~ 60 K samples) (Cui et al. , 2023 ; Ivison et al. , 2023) を優先データセットとして使用します。学習ハイパーパラメータはTunstall et al. sDPOでは2つのステップを使用し、第1ステップではOpenOrcaをデータセットD 1として使用し、第2ステップではUltrafeedback CleanedをデータセットD 2として使用します。

3.2 Main Results

SFTベースモデルにsDPOを適用した場合の評価結果と、他の上位モデルの結果を表2に示します。2. 事前学習のみの「SOLAR 10.7B」と命令チューニングされた「SOLAR 10.7B + SFT」を比較すると、H4の観点で+5 . 24増加しました。SOLAR 10.7B + SFTにsDPOを適用すると、H4スコアは74 . 31となり、+4 . 80 . 注目すべきは、「SOLAR 10.7B + SFT + sDPO」は、パラメータ数が少ないにもかかわらず、Mixtral 8x7B-Instruct-v0.1のような他の大規模モデルを凌駕していることです。これは、効果的なアライメントチューニングが、小規模なLLMの次のレベルの性能を引き出す鍵になる可能性があることを強調しています。さらに、sDPOを適用した結果、TruthfulQAのスコアは72.45と大幅に向上し、アライメントチューニングプロセスの有効性が示されました。

3.3 Ablation Studies

また、Tab. SOLAR 10.7B + SFT + DPO」は、従来のDPO学習セットアップと同じように、DPOデータを一度に使用します。SOLAR 10.7B + SFT + sDPO Strat.」では、層別サンプリングにより~ 16 . OpenOrcaとUltrafeedback Cleanedを合わせたデータポイントの67%をD 1とし、残りの約83 . 33%をD 2として使用し、SOLAR 10.7B + SFT + sDPOで使用したD 1とD 2のデータセットサイズを反映しています。

評価。HuggingFace Open LLM Leaderboard (Beech- ing et al. , 2023) の6つのタスクのうち4つを利用：ARC (Clark et al. , 2018)、Hel- laSWAG (Zellers et al. , 2019)、MMLU (Hendrycks et al. , 2020)、TruthfulQA (Lin et al. , 2022) 。また、4つのタスクの平均スコアをH4と表記して報告します。W inogrande (坂口

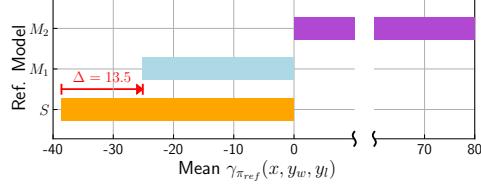


Figure 2: Mean $\gamma_{\pi_{ref}}$ on Ultrafeedback Cleaned dataset for different reference models S , M_1 , and M_2 . Note that the x-axis is in log scale.

Comparing SOLAR 10.7B + SFT + DPO and SOLAR 10.7B + SFT + sDPO, we can see that using sDPO over DPO results in a higher H4 score overall, with noticeable improvements in ARC and TruthfulQA scores. Therefore, we believe sDPO could function as a drop-in replacement for DPO training with better performance. Looking at SOLAR 10.7B + SFT + sDPO and SOLAR 10.7B + SFT + sDPO Strat., we see that the specific way of splitting the available DPO data into multiple D_t can also impact performance. We find that the natural split of using different preference datasets as D_t works best in our experiments. We believe further exploration of how to define D_t is an interesting direction for future research.

3.4 Reference Models in sDPO

Effectiveness of sDPO in terms of alignment tuning. In Sec. 2.2, we explain that the reference models in sDPO are more aligned, resulting in higher $\gamma_{\pi_{ref}}$, *i.e.*, a stricter lower bound. We verify the above empirically in Fig. 2 by comparing the mean $\gamma_{\pi_{ref}}$ on the Ultrafeedback Cleaned dataset for the reference models in steps 1 and 2 of sDPO, *i.e.*, S and M_1 . Note that these two models have not been trained on the aforementioned dataset. Using the SFT base model S as the reference model, the mean of $\gamma_{\pi_{ref}}$ is -38.60 . On the other hand, using the aligned model M_1 from step 1 of sDPO as the reference model, the mean of $\gamma_{\pi_{ref}}$ is -25.10 , an increase of 13.50 in *log scale*. Thus, a single step of sDPO greatly increases $\gamma_{\pi_{ref}}$, which results in a more performant aligned model as seen in Tab. 2.

Adopting open source models as reference models could be dangerous. We also show mean $\gamma_{\pi_{ref}}$ of M_2 , the aligned model from step 2 of sDPO. Unlike S and M_1 , M_2 is trained on the Ultrafeedback Cleaned dataset, *i.e.*, M_2 is used as a reference model on data that was already used to train it. Note that such a case could happen commonly when adopting various open source models

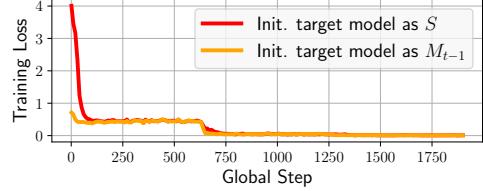


Figure 3: Loss curve comparison in step 2 of sDPO for different initializations of the target model.

as reference models. This is because the datasets that were used in training those models are often unclear and could overlap with the preference datasets unintentionally. Mean $\gamma_{\pi_{ref}}$ of M_2 is 84.35, which is staggeringly higher than either S or M_1 . The strikingly high value for M_2 likely points to overfitting of M_2 to the Ultrafeedback Cleaned dataset. This result highlights the potential danger of merely adopting open source models as reference models instead of using sDPO.

3.5 Target Model Initialization in sDPO

The target model in each step of sDPO is also initialized with M_{t-1} , the aligned model from the last step. This ensures that the final model in sDPO has undergone training with the same amount of data as the final model in DPO. On the other hand, one concern of such design choice is that it may become increasingly difficult to stabilize the training of the target model as the steps progress, since it has already undergone training with a decreasing learning rate schedule in the preceding steps. Thus, another option is to use the initial SFT base model S as the target model for all steps of sDPO.

However, as shown in Fig. 3, initializing the target model as S results in a much bigger initial loss than that of M_{t-1} , which could lead to an unstable training. The main reason is that DPO training is usually done where the reference and target models are the same. In contrast, initializing the target model as S creates a differential in the reference and target models, which may be amplified as the steps progress. Thus, for stable training, initializing the target model as M_{t-1} was chosen for sDPO.

4 Conclusion

We propose sDPO where we use the preference data in a stepwise way instead of all at once. We show that applying sDPO results in more performant models than DPO in terms of H4 score. We also empirically exhibit that sDPO results in more aligned reference models by comparing mean $\gamma_{\pi_{ref}}$.

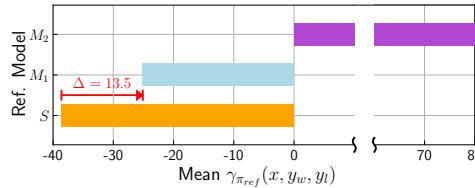


図2: Ultrafeedback Cleanedデータセットにおける、参照モデルS、M 1、M 2の平均 $\gamma_{\pi_{ref}}$ 。x軸は対数スケールであることに注意。

SOLAR 10.7B + SFT + DPOとSOLAR 10.7B + SFT + sDPOを比較すると、DPOよりもsDPOを使用した方が全体的にH4スコアが高く、ARCとTruthfulQAのスコアが顕著に向上去んでいることがわかります。したがって、sDPOはDPOトレーニングの代替として、より良いパフォーマンスを発揮できると考えています。SO-LAR 10.7B + SFT + sDPOとSOLAR 10.7B + SFT + sDPO Strat.を見ると、利用可能なDPOデータを複数の D_t に分割する特定の方法もパフォーマンスに影響を与えることがわかります。私たちの実験では、異なるプリファレンスデータセットを D_t として使用する自然な分割が最も効果的であることがわかりました。 D_t をどのように定義するかについてのさらなる探求は、今後の研究の興味深い方向性であると考えます。

3.4 Reference Models in sDPO

アライメント調整におけるsDPOの有効性第2.2節で、sDPOの参照モデルはより整列しており、その結果 $\gamma_{\pi_{ref}}$ が高くなる、すなわち下界が厳しくなることを説明しました。図2では、sDPOのステップ1とステップ2の参照モデル、すなわちSとM 1について、Ultrafeedback Cleanedデータセットにおける平均 $\gamma_{\pi_{ref}}$ を比較することで、上記のことを経験的に検証しています。この2つのモデルは前述のデータセットで学習されていないことに注意してください。SFTベースモデルSを参照モデルとして使用した場合、 $\gamma_{\pi_{ref}}$ の平均は -38.60。一方、sDPOのステップ1の整列モデルM 1を参照モデルとして使用すると、 $\gamma_{\pi_{ref}}$ の平均は -25.10となり、対数スケールで 13.50増加します。このように、sDPOの1ステップで $\gamma_{\pi_{ref}}$ は大幅に増加し、Tab. 2。

オープンソースモデルを参照モデルとして採用することは危険です。また、sDPOのステップ2で整列されたモデルであるM 2の平均 $\gamma_{\pi_{ref}}$ も示します。SやM 1とは異なり、M 2はUltrafeedback Cleanedデータセットで学習されています。このようなケースは、様々なオープンソースモデルを参照モデルとして採用する場合によく起こります。

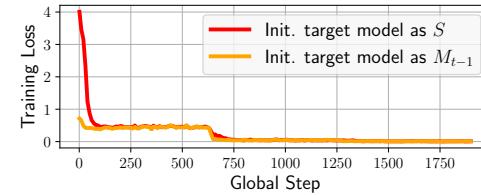


図3: sDPOのステップ2における、ターゲットモデルの初期化の違いによる損失曲線の比較。

参照モデルとしてというのも、それらのモデルのトレーニングに使用されたデータセットは明確でないことが多い、嗜好データセットと意図せず重なってしまう可能性があるからです。M 2の平均 $\gamma_{\pi_{ref}}$ は84.35であり、SやM 1よりも圧倒的に高い値です。M 2の値が突出して高いのは、M 2がUltrafeedback Cleanedデータセットに過剰にフィットしている可能性があります。この結果は、sDPOを使用する代わりに、単にオープンソースモデルを参照モデルとして採用することの潜在的な危険性を浮き彫りにしています。

3.5 Target Model Initialization in sDPO

sDPOの各ステップにおけるターゲット・モデルも、 M_{t-1} で初期化されます。これにより、sDPOの最終モデルはDPOの最終モデルと同じ量のデータで学習されたことになります。一方、このような設計を選択した場合に懸念されるのは、ステップが進むにつれて、ターゲットモデルの学習を安定させることが難しくなることです。そこで、sDPOの全ステップにおいて、最初のSFTベースモデルSをターゲットモデルとして使用するという選択肢もあります。しかし、図3に示すように、ターゲットモデルをSとして初期化すると、初期損失が M_{t-1} よりもはるかに大きくなり、学習が不安定になる可能性があります。その主な理由は、DPO学習は通常、参照モデルとターゲットモデルが同じ場合に行われるからです。一方、ターゲットモデルをSとして初期化すると、参照モデルとターゲットモデルに差分が生じ、ステップが進むにつれて差分が増幅される可能性があります。したがって、安定した学習のために、sDPOではターゲットモデルを M_{t-1} として初期化することが選択されました。

4 Conclusion

我々は、嗜好データを一度に使用するのではなく、段階的に使用するsDPOを提案します。sDPOを適用することで、H4スコアの点でDPOよりも高性能なモデルが得られることを示します。また、平均 $\gamma_{\pi_{ref}}$ を比較することで、sDPOの方がより整合のとれた参照モデルが得られることを実証します。

Limitations

While we have demonstrated the effectiveness of employing different datasets in distinct stages of sDPO, identifying an optimal strategy for segmenting more intricate DPO data collections remains an area for further exploration. This task is particularly challenging due to the complexities within these datasets. Our approach, while promising, necessitates a more deeper understanding of dataset characteristics and their impact on the performance of sDPO.

Furthermore, our experiments predominantly utilized SOLAR 10.7B models, driven by the state-of-the-art performance at the time of experimentation along with its unique 10.7 billion parameter size. The unique size of SOLAR 10.7B models made the usage of sDPO more necessary as there are far fewer open source LLMs that can be adopted as reference models.

Additionally, as with most research on LLMs, we operated within our limitations in computational resources. Although this focus has yielded significant insights, expanding our experimental framework to incorporate a broader range of Large Language Models (LLMs) could potentially unveil more comprehensive understanding of the strengths and limitations of sDPO. Such an expansion would allow for a more robust comparison across different model architectures and sizes, further enriching our findings.

Evaluating the efficacy of LLMs is an evolving challenge in the field. In our study, we primarily employed tasks from the Huggingface Open LLM Leaderboard as benchmarks for evaluation. While this provided comparative results, future research could benefit from incorporating a wider array of tasks and benchmarks. These could include tasks that judge actual human or strong AI preference alignment. Such additional evaluation would not only enhance the validity of our findings but also contribute to the broader discourse on LLM assessment methodologies.

Ethics Statement

In this study, we strictly adhered to ethical standards in the conduct of our research. Our experiments were based entirely on open models and open datasets, ensuring transparency and accessibility. We took meticulous care to avoid any biases or data contamination, thereby maintaining the integrity of our research process. The experimental

environment was rigorously designed to be objective, ensuring that all comparisons conducted were fair and impartial. This approach reinforces the reliability and validity of our findings, contributing positively to the field while upholding the highest ethical standards. We confirmed that all the data used in our experiments were free of licensing issues.

References

- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning

Limitations

sDPOの異なる段階において異なるデータセットを採用することの有効性を示しましたが、より複雑なDPOデータコレクションをセグメント化するための最適な戦略を特定することは、さらなる探求が必要な領域です。これらのデータセットは複雑であるため、このタスクは特に困難です。私たちのアプローチは有望ではありますが、データセットの特性とそれらがsDPOの性能に与える影響について、より深く理解する必要があります。さらに、我々の実験では、SOLAR 10.7Bモデルを主に使用しました。これは、実験時の最先端の性能と、107億というユニークなパラメータサイズによるものです。SOLAR10.7Bモデルのユニークなサイズにより、参照モデルとして採用できるオープンソースのLLMがはあるかに少ないとみたため、sDPOを使用する必要性が高まりました。さらに、LLMに関する多くの研究と同様に、計算資源の制限の中で研究を進めました。このような取り組みから重要な知見を得ることができましたが、より広範な大規模言語モデル(LLM)を取り入れるために実験の枠組みを拡張することで、sDPOの長所と限界をより包括的に理解できる可能性があります。このような拡張により、異なるモデルアーキテクチャやサイズ間でよりロバストな比較が可能になり、我々の知見がさらに豊かになります。LLMの有効性を評価することは、この分野における発展的な課題です。我々の研究では、主にHuggingface Open LLM Leaderboardのタスクを評価のベンチマークとして採用しました。これによって比較結果が得られますが、今後の研究では、より幅広いタスクやベンチマークを取り入れることが有益となるでしょう。これには、実際の人間や強力なAIの嗜好アライメントを判定するタスクが含まれる可能性があります。このような追加的な評価は、我々の知見の妥当性を高めるだけでなく、LLM評価方法論に関するより広範な議論に貢献するでしょう。

Ethics Statement

本研究では、倫理基準を厳守して研究を実施しました。私たちの実験はすべてオープンモデルとオープンデータセットに基づいており、透明性とアクセシビリティを確保しています。バイアスやデータ汚染を避けるために細心の注意を払い、研究プロセスの完全性を維持しました。実験環境は客観的であるように厳密に設計さ

れ、実施されたすべての比較が公正かつ公平であることを保証しました。このアプローチにより、研究結果の信頼性と妥当性が強化され、高い倫理基準を守りながら、この分野に積極的に貢献しています。実験に使用したすべてのデータにライセンス上の問題がないことを確認しました。

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin John- son, Dmitry Lepikhin, Alexandre Passos, Siyamak Shakeri, Emanuel Tarupa, Paige Bailey, Zhiifeng Chen, et al. Palm 2 テクニカルレポート arXiv preprint arXiv:2305.10403 .

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kershner, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 .

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. オープンllmリーダーボード. https://huggingface.co/space/HuggingFace/open_llm_leaderboard .

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 言語モデルは少数ショット学習者 神経情報処理システムの進歩 , 33:1877–1901.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems , 30.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457 .

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichihiro Nakano, et al. ArXiv preprint arXiv:2110.14168 .

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. ウルトラフィードバック : arXivプレプリントarXiv:2310.01377 .

潘瑞、廖世哲、張吉鵬、嚴佳春、張桐。2023. ラフト報酬ランクの微調整

- for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*.
- Intel. 2023a. Intel/neural-chat-7b-v3-1. <https://huggingface.co/Intel/neural-chat-7b-v3-1>.
- Intel. 2023b. Supervised fine-tuning and direct preference optimization on intel gaudi2.
- Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. Solar 10.7b: Scaling large language models with simple yet effective depth upscaling.
- Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Mistralorca: Mistral-7b model instruct-tuned on filtered openorcav1 gpt-4 dataset. <https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca>.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.
- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawa-har, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.
- OpenAI. 2023. Gpt-4 technical report.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Rafael Raffailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Teknium. 2023. teknium/openhermes-2.5-mistral-7b. <https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B>.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.
- Upstage. 2023. upstage/solar-0-70b-16bit. <https://huggingface.co/upstage/SOLAR-0-70b-16bit>.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.



for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

ダン・ヘンドリクス、コリン・バーンズ、スティーブン・バサート、アンディ・ゾウ、マンタス・マゼイカ、ドーン・ソング、ジェイコブ・スタイルンハート2020. 大規模なマルチタスク言語アンダースタンドの測定。In International Conference on Learning Representation s .

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 202 1. 転送のスケーリング則。arXiv preprint arXiv:2102.01293 .

Intel. 2023a. Intel/neural-chat-7b-v3-1. <https://huggingface.co/Intel/neural-chat-7b-v3-1>.

インテル。2023b. intel gaudi2 における教師あり微調整と直接選好最適化。

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambe rt, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. 変化する気候におけるラクダ：を用いたLM適応の強化。

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Ch ess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. ニューラル言語モデルのスケーリング則。arXiv preprint arXiv: 2001.08361 .

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsoo Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoon Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungji n Gim, Mikkyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. 太陽10.7b : シンプルかつ効果的な深度アップスケーリングによる大規模言語モ デルのスケーリング。

Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. ミストラルカーフィルタリングされたopenorca1 gpt-4データセットでインストラクションチューニングさ れたMistral-7bモデル。https://huggingface.co/OpenOrca/ Mistral-7B-0 penOrca .

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa : モデルによる人間の虚偽の模倣方法の測定。邦訳は、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」、「邦訳」。

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawa- har, S ahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. ar XivプレプリントarXiv:2306.02707 .

OpenAI. 2023. Gpt-4 technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 言語モデルは教師なしマルチタスク学習者。OpenAI blog , 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. 直接選好最適化：あなたの言語モデルは密かに報酬モデル。arXiv preprint arXiv:230 5.18290 .

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavat-ula, and Ye jin Choi. 2021. Winogrande：大規模な広告ウィノグランド・スキーマ・チャレンジ。Communications of the ACM , 64(9):99-106.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proxi- mal policy optimization algorithms.

Teknium. 2023. teknium/openhermes-2.5-mistral-7b. <https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B>.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. ゼファーarXivのプレプリントarXiv:2310.16944 .

Upstage. 2023. upstage/solar-0-70b-16bit. <https://huggingface.co/upstage/SOLAR-0-70b-16bit>.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. arXivプレプリントarXiv:2206.07682 .

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. 自己報酬型言語モデル。arX iv preprint arXiv:2401.10020 .

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhhf : arXivプレプリントarXiv:2304.05302 .

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag : 機械は本当にあなたの文章を終えることができますか？Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pages 4791-4800.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences.

A Related Work

A.1 Large Language Models

Recent research has highlighted a "scaling law" in the field of context-based language models (Kaplan et al., 2020; Hernandez et al., 2021; Anil et al., 2023), showing a proportional relationship between the size of the model plus the training data and the resulting performance improvements. Consequently, this has led to the advent of Large Language Models (LLMs). In contrast to earlier models, LLMs can perform in-context learning, which includes abilities such as zero-shot learning (Radford et al., 2019) and few-shot learning (Brown et al., 2020), allowing them to adapt and perform tasks without the need for weight adjustments. These emergent abilities of LLMs, absent in their smaller counterparts, signal a significant evolution in language model capabilities (Wei et al., 2022).

A.2 Alignment Tuning

LLMs have been recognized to produce text that may seem linguistically inconsistent to human interpreters because their pretraining is based not on an understanding of human intentions but on a broad spectrum of domain-specific knowledge, as indicated in (Ziegler et al., 2019). In an effort to rectify this issue and better mirror human intentions, prior research (Ziegler et al., 2019) has suggested the adoption of Reinforcement Learning with Human Feedback (RLHF). RLHF seeks to refine the LLM’s output by constructing a reward model that aligns with human preferences and applying reinforcement learning to direct the LLM towards selections that garner the most favorable reward metrics. This approach is intended to bolster the safety, decorum, and general excellence of the responses produced by the LLM. Nonetheless, despite showing promising results, RLHF is confronted with challenges, such as the intricate handling of an extensive set of hyperparameters and the necessity to amalgamate several models (policy, value, reward, and reference models).

To address these issues, there have been proposals for supervised fine-tuning methodologies such as Rank Responses to align Human Feedback (RRHF) (Yuan et al., 2023), Reward rAnked Fine-Tuning (RAFT) (Dong et al., 2023), and Direct Preference Optimization (DPO) (Rafailov et al., 2023). These methods circumvent the intricacies inherent in reinforcement learning and have been shown to yield empirical results on par with RLHF.

Notably, the DPO technique straightforwardly encourages the LLM to favor positive responses and discourage negative ones. DPO has been observed to yield performant learning outcomes, in spite of its uncomplicated training procedure.

Concurrent to our work, Yuan et al. (2024) have developed an iterative framework for generating *new* preference datasets and performing DPO training on the resulting datasets. They empirically demonstrated the superiority of their iterative framework in terms of AlpacaEval 2.0. In contrast, our work is complementary to the above in the sense that we focus on utilizing the *current* preference data and does not undergo new data generation. Thus, our method can also be applied to Yuan et al. (2024) by changing the DPO training part to using sDPO instead. We leave the above combination as an interesting future work. Additionally, the evaluation used in Yuan et al. (2024) is also different to ours as we utilize tasks from Open LLM Leaderboard whereas Yuan et al. (2024) uses AlpacaEval 2.0.



A Related Work

A.1 Large Language Models

最近の研究では、文脈ベースの言語モデルの分野で「スケーリング則」が注目されています (Kaplan et al. , 2020 ; Hernandez et al. , 2021 ; Anil et al. , 2023)。その結果、ラージ・ランゲージ・モデル (LLM) の登場へとつながりました。以前のモデルとは対照的に、LLMはゼロショット学習(Radford et al. , 2019)や少数ショット学習(Brown et al. , 2020)などの能力を含むコンテキスト内学習を行うことができ、重みの調整を必要とせずに適応してタスクを実行することができます。このようなLLMの新たな能力は、小型のLLMにはないものであり、言語モデルの能力が大きく進化したことを示しています (Wei et al.)

注目すべきは、DPO手法がLLMに対して、肯定的な反応を支持し、否定的な反応を抑制するよう、ストレートに働きかけることです。DPOは、その単純な学習手順にもかかわらず、高性能な学習結果をもたらすことが確認されています。我々の研究と同時に、Yuanら (2024年) は、新しい嗜好データセットを生成し、得られたデータセットに対してDPO学習を行うための反復フレームワークを開発しました。彼らは、AlpacaEval 2.0の観点から、彼らの反復フレームワークの優位性を経験的に実証しました。対照的に、我々の研究は、現在の嗜好データの活用に焦点を当て、新たなデータ生成を行わないという意味で、上記を補完するものです。したがって、DPOの学習部分をsDPOに変更することで、Yuan et al. この組み合わせは今後の課題です。加えて、Yuanら (2024) で使用されている評価も、Yuanら (2024) がAlpacaEval 2.0を使用しているのに対し、我々はOpen LLM Leaderboardのタスクを使用しているため、我々の評価とは異なります。

A.2 Alignment Tuning

LLMは、(Ziegler et al. , 2019)で示されているように、その事前学習が人間の意図の理解ではなく、広範なドメイン固有の知識に基づいているため、人間の解釈者には言語的に矛盾しているように見えるテキストを生成することが認められています。この問題を是正し、人間のインテンションをよりよく反映させるために、先行研究(Ziegler et al. , 2019)では、人間フィードバック付き強化学習(RLHF)の採用が提案されています。RLHFは、人間の嗜好に沿った報酬モデルを構築し、強化学習を適用して、LLMを最も有利な報酬メトリクスを獲得する選択に向かわせることで、LLMの出力を改善しようとするものです。このアプローチは、LLMによって生成される応答の安全性、礼儀正しさ、一般的な優秀性を強化することを目的としています。しかしながら、RLHFは有望な結果を示しているにもかかわらず、膨大なハイパーパラメータの複雑な取り扱いや、複数のモデル（政策、価値、報酬、参照モデル）を統合する必要性などの課題に直面しています。これらの問題に対処するために、Rank Responses to align Human Feedback (RRHF) (Yuan et al. , 2023)、Reward rAnked Fine- Tuning (RAFT) (Dong et al. , 2023)、Direct Preference Optimization (DPO) (Rafailov et al. , 2023)のような教師ありファインチューニング手法が提案されています。これらの手法は強化学習特有の複雑さを回避し、RLHFと同等の経験的結果をもたらすことが示されています。