進捗報告

1 学習

一般的な応答を 202 件収録している D_0, お嬢様スタイルの応答を 202 件収録しているお嬢様データセット D_1 の 2 つのデータセットを用いて SFT,DPO を適用した.

ベースモデルとして, elyza/Llama-3-ELYZA-JP-8B[1] を用いた.

SFT には huggingface で提供されている SFTTrainer クラスを使用し, DPO においても同じく huggingface で提供されている DPOTrainer クラスを使用した. また学習の際には効率化のため両手法ともに QLoRA の処理をしている. 表 1, 2, 3 に各手法のパラメータを示す.

表 1: QLoRA パラメータ

パラメータ	値
量子化サイズ	4 ビット
r	8
lora_alpha	128
target_modules	モデル内の線形層全て
lora_dropout	0.05

表 2: SFTTrainer パラメータ

パラメータ	値
epoch 数	3
バッチサイズ	2
最適化手法	Adam
初期学習率	1e-4
学習率スケジューラ	cosine

表 3: DPOTrainer パラメータ

パラメータ	値
epoch 数	3
バッチサイズ	2
最適化手法	Adam
初期学習率	1e-5
学習率スケジューラ	cosine
beta	0.3

[1] elyza/Llama-3-ELYZA-JP-8B

1.1 D_0 , D_1 のみの SFT

図 1, 2 にそれぞれ D_0, D_1 を用いた SFT における 10 ステップごとの loss の値の推移を示している. 2 つの図に共通して横軸はステップ数, 縦軸は loss の値を示している.

図 1: D_0 を用いた SFT における Loss の推移

図 2: D_1 を用いた SFT における Loss の推移

両方とも, 100, 200 ステップ付近で大きく loss が大きく減少しており全体としてうまく学習が進行していると考える.

1.2 D_1 を chosen, D_0 を rejected とした (お嬢様スタイルを選好した) DPO

図 3 に DPOTrainer における 10 ステップごとの loss の推移を示す. 100 ステップ目まで急速に loss が減少し, それ以降は 0 に近い値が続いている.

また, DPOTrainer には以下の 2 つの評価指標が存在する.

- Rewards/accuracies

  chosen と rejected のうち, chosen により高い暗示的報酬が与えられた割合

- Rewards/margins

  chosen と rejected に与えられた暗示的報酬の差

図 4, 5 にそれぞれ Rewards/accuracies, Rewards/margins の 10 ステップごとの値の推移を示す. Rewards/accuracies に関しては学習の早い段階で 1.0 となり, Rewards/margins に関しては学習が進むにつれて値が上昇しており chosen を選ぶように学習が進んでいることがわかる.

図 3: D_1 を chosen とした DPO における loss の推移

1.3 D_0 を chosen, D_1 を rejected とした (一般的な応答を選好した) DPO

図 6, 7, 8 にそれぞれ学習時における 10 ステップごとの loss, Rewards/accuracies, Rewards/margins の推移を示す.

1.2 節と同様に chosen を選ぶように学習が進んでいることがわかる.

1.4 D_0 と D_1 を両方用いて SFT

元のデータセットのデータ数の半分である 101 個を D_0 からランダムに選択し D_1 の対応する部分と置き換えることで, D_0 と D_1 を両方用いたデータセット D_2 を作成し SFT を適用した. 図 1 に D_2 を用いて SFT を実行した際の 10 ステップごとの loss の推移を示す.

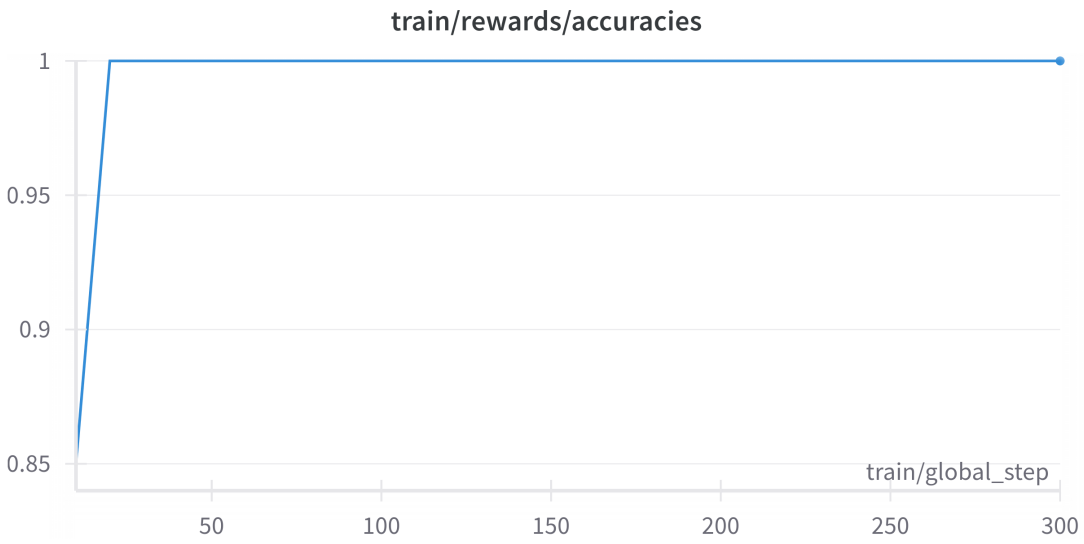

図 4: D_1 を chosen とした DPO における Rewards/accuracies の推移

1.4.1 D_2 で SFT で学習後のモデルに D_1 を chosen, D_0 を rejected とした (お嬢様スタイルを選好した) DPO

図 10, 11, 12 にそれぞれ学習時における 10 ステップごとの loss, Rewards/accuracies, Rewards/margins の推移を示す.

chosen を選ぶように学習が進んでいることがわかる.

1.4.2 D_2 で SFT で学習後のモデルに D_0 を chosen, D_1 を rejected とした (一般的な応答を選好した) DPO

図 13, 14, 15 にそれぞれ学習時における 10 ステップごとの loss, Rewards/accuracies, Rewards/margins の推移を示す.

chosen を選ぶように学習が進んでいることがわかる.

2 モデルの定量的な評価

モデルの重みの L1 ノルム, L2 ノルムを計算し, SFT, DPO の違い, 影響範囲などを考察する. 各学習モデルを下記のように表現する.

M_{base} ベースモデル (elyza/Llama-3-ELYZA-JP-8B)

$M_{\text{SFT}_{\text{D0}}}$ M_{base} に対して D_0 を用いて SFT を適用したモデル [2]

$M_{\text{SFT}_{\text{D1}}}$ M_{base} に対して D_1 を用いて SFT を適用したモデル [3]

$M_{\text{DPO}_{\text{D0}}}$ M_{base} に対して D_0 を用いて DPO を適用したモデル [4]

$M_{\text{DPO}_{\text{D1}}}$ M_{base} に対して D_1 を用いて DPO を適用したモデル [5]

[2] https://huggingface.co/Nisk36/SFT_normal
[3] https://huggingface.co/Nisk36/SFT_ojousama
[4] https://huggingface.co/Nisk36/Llama-3-ELYZA-JP-8B-normal-chosen
[5] https://huggingface.co/Nisk36/Llama-3-ELYZA-JP-8B-ojousama-chosen

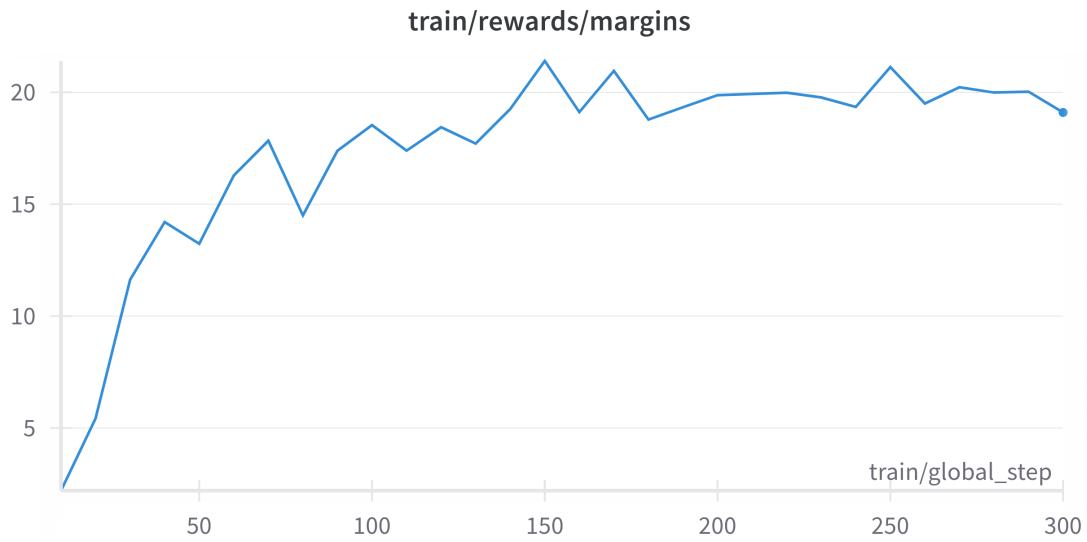

図 5: D_1 を chosen とした DPO における Rewards/margins の推移

$M_{\mathrm{SFT}_{\mathrm{D}_2}}$ M_{base} に対して D_2 を用いて SFT を適用したモデル [6]

$M_{\mathrm{DPO}_{\mathrm{D}_0(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}}$ $M_{\mathrm{SFT}_{\mathrm{D}_2}}$ に対して D_0 のみを用いて DPO を適用したモデル [7]

$M_{\mathrm{DPO}_{\mathrm{D}_1(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}}$ $M_{\mathrm{SFT}_{\mathrm{D}_2}}$ に対して D_1 のみを用いて DPO を適用したモデル [8]

また, 図 16 に今回実験に用いた 8b サイズの Llama3 のアーキテクチャを示す.

2.1 L1 ノルム, L2 ノルムの計算

まず, (学習後のモデル, 学習前のモデル) となっているモデル間の L1 ノルム, L2 ノルムを計算した. 表 4 に結果を示す. SFT と DPO の違いとして, ベースモデルからの変化の大きさが挙げられる. SFT を施すことで

表 4: (学習後のモデル, 学習前のモデル) の 2 モデル間の L1 ノルム, L2 ノルム

モデル	L1 ノルム	L2 ノルム
$M_{\mathrm{SFT}_{\mathrm{D}_0}}$, M_{base}	893386.5	13.584439160476501
$M_{\mathrm{SFT}_{\mathrm{D}_1}}$, M_{base}	918794.5	14.002781183014145
$M_{\mathrm{DPO}_{\mathrm{D}_0}}$, M_{base}	157415.125	0.5359644668956891
$M_{\mathrm{DPO}_{\mathrm{D}_1}}$, M_{base}	156530.9375	0.5838905182501243
$M_{\mathrm{SFT}_{\mathrm{D}_2}}$, M_{base}	885563.0	13.461171011468505
$M_{\mathrm{DPO}_{\mathrm{D}_0(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}}$, M_{base}	912034.0	13.920651777069834
$M_{\mathrm{DPO}_{\mathrm{D}_1(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}}$, M_{base}	908732.5	13.85587346550868
$M_{\mathrm{DPO}_{\mathrm{D}_0(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}}$, $M_{\mathrm{SFT}_{\mathrm{D}_2}}$	186765.0625	0.9965761998485839
$M_{\mathrm{DPO}_{\mathrm{D}_1(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}}$, $M_{\mathrm{SFT}_{\mathrm{D}_2}}$	173904.25	0.7142688962114501

L1 ノルム, L2 ノルム共に大きな値となるが, DPO では SFT と比較してかなり小さい値となる. ベースモデルからの重みの変化という観点では SFT は DPO と比較して大きな変化をもたらす手法であると言える.

[6] https://huggingface.co/Nisk36/SFT_both
[7] https://huggingface.co/Nisk36/Llama-3-ELYZA-JP-8B-normal-chosen-after-SFTboth
[8] https://huggingface.co/Nisk36/Llama-3-ELYZA-JP-8B-ojousama-chosen-after-SFTboth

図 6: D_0 を chosen とした DPO における Loss の推移

表 5, 6, 7, 8, 9, 9, 10, 11, 12, 13 に それぞれ $M_{\mathrm{SFT}_{\mathrm{D}0}}$, M_{base} 間, $M_{\mathrm{SFT}_{\mathrm{D}1}}$, M_{base} 間, $M_{\mathrm{DPO}_{\mathrm{D}0}}$, M_{base} 間, $M_{\mathrm{DPO}_{\mathrm{D}1}}$, M_{base} 間, $M_{\mathrm{SFT}_{\mathrm{D}2}}$, M_{base} 間, $M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}$, M_{base} 間, $M_{\mathrm{DPO}_{\mathrm{D}1}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}$, M_{base} 間, $M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}$, $M_{\mathrm{SFT}_{\mathrm{D}2}}$ 間, $M_{\mathrm{DPO}_{\mathrm{D}1}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}$, $M_{\mathrm{SFT}_{\mathrm{D}2}}$ 間の L1 ノルム, L2 ノルム を示す.

また SFT がどの層に影響しているかを把握するため, SFT を適用したモデルとベースモデルの L1 ノルム, L2 ノルムに関して各層における平均値からの差分を計算した. 表 14 に結果を示す. SFT を適用したモデル全てについて, 29, 30, 31 層目が平均と比べ大きな L2 ノルムの値となっている. また全体的な傾向として, 0 から 9 層目は L2 ノルムが平均値より小さくなりベースモデルより重みの変化が少ない. SFT によって出力を変化させるにあたって層が深くなるにつれて文脈的な意味など高次の情報を捉える重みの変化が大きくなっていると考えられる. また, 中間層の変化が大きくなっているのはお嬢様のロールプレイをするための文法的なルールや単語同士の近接関係, 29, 30, 31 層目はタスクに即してお嬢様らしい文章を作り出す文脈全体の意味を捉えるための変化が今回のロールプレイのタスクにより大きくなったと考えられる.

また DPO がどの層に影響しているかを把握するため, SFT の場合と同様に DPO を適用したモデルと適用前のモデルの L1 ノルム, L2 ノルムに関して各層における平均値からの差分を計算した. 表 14 に結果を示す.

$M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{DPO}_{\mathrm{D}1}}$ はそれぞれ, 10 から 17 層目 と 31 層目, 21 から 31 層目に関して局所的にベースモデルと比較して大きな L2 ノルムが見られる. これは, 出力のスタイルの差によるものｔ考えられる. D_0 の出力のスタイルは D_1 のお嬢様スタイルと比較すると一般的な応答であり M_{base} の出力と語彙や文の長さ, 文法などは変わるものの口調などは大差ないと考えられる. そのため, 10 から 17 層目 のような中間の層の重みの変化が大きくなっていると考えられる. 一方で, $M_{\mathrm{DPO}_{\mathrm{D}1}}$ は口調や一人称, 代名詞の変化など文全体に渡って出力のスタイルが大きくベースモデルと異なるため出力に影響するより高次の情報を重みとして持つと考えられる 21 から 31 層目の重みの変化が大きくなっていると考えられる.

層が深くなるに連れ出力のスタイルにより影響する重みを持つという仮定を持つとすれば, $M_{\mathrm{SFT}_{\mathrm{D}2}}$ から DPO を適用したモデルに関しても D_2 で語彙や文の長さなどを学習した後に, より各データセットの応答に沿った形でロールプレイの性能が上がるように学習されたことで深い層の重みの L2 ノルムが大きくなっていると考えられる.

3　応答結果の評価

各モデルの生成結果に関して, 以下の 2 点から定量的に評価する.

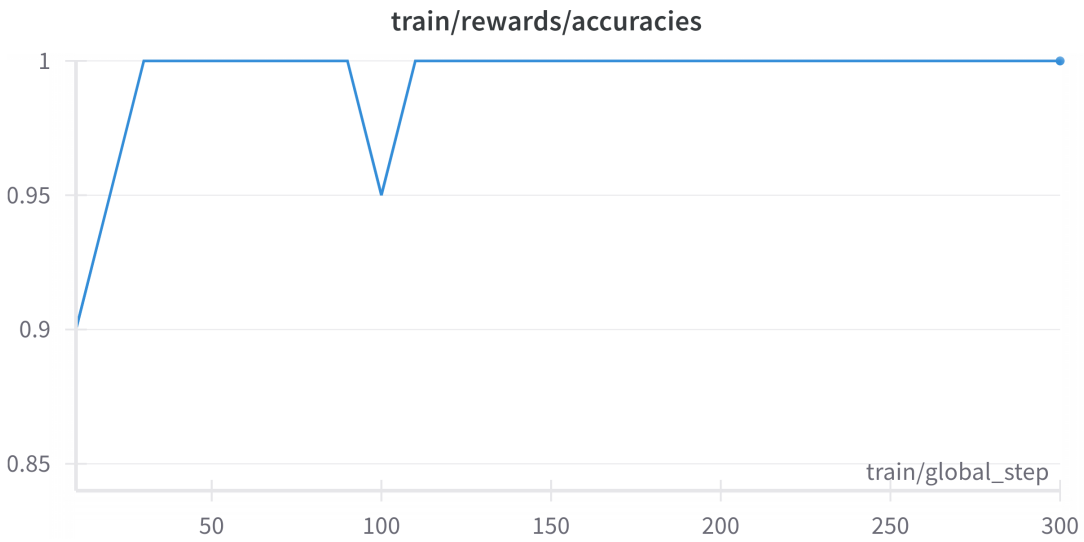

図 7: D_0 を chosen とした DPO における Rewards/accuracies の推移

- データセット D_0, D_1 のクエリによる生成結果と D_0, D_1 内のクエリに対する応答文との類似度の比較
  文章同士の類似度を評価する指標として BLEU, BERTScore を用いる.
- ChatGPT 4o によって新たに生成した 100 個のクエリに対する生成結果のスコア
  スコアは D_1 で記載されているお嬢様スタイルがどれだけ再現できているかという観点で採点し, 評価者 LLM により評価 (採点) する.

3.1 データセットとの類似性による評価

学習済みモデルにデータセット内のクエリを投げ, その応答とデータセットで用意されている応答とを比較評価する. 表 24 に各モデルの BLEU, BERTScore を示す. 各モデルともに, 最後に学習したデータセットのデータに生成結果の類似度が高くなっていることがわかる. データセット D_0, D_1 について最も BLEU, BERTScore の値が大きいのはそれぞれ $M_{\mathrm{SFT}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}1}}$ となっており, 今回の実験の設定では SFT は元データに似た形の応答を実現するように学習していることが考えられる. また, 最後に DPO を適用したモデルに関して, ベースモデルにそのまま DPO を適用した場合と D_2 で SFT してから DPO を適用した場合を比較するとどちらのデータセットに対する BLEU, BERTScore も後者のほうが大きな値となった. このことからの SFT によりデータセット内の文に登場する語彙などを応答するように学習できることが伺える.

評価者 LLM による生成結果の自動評価評価者 LLM として, GPT-4o-mini を使用した. 評価者 LLM に与えたプロンプトを以下に示す.

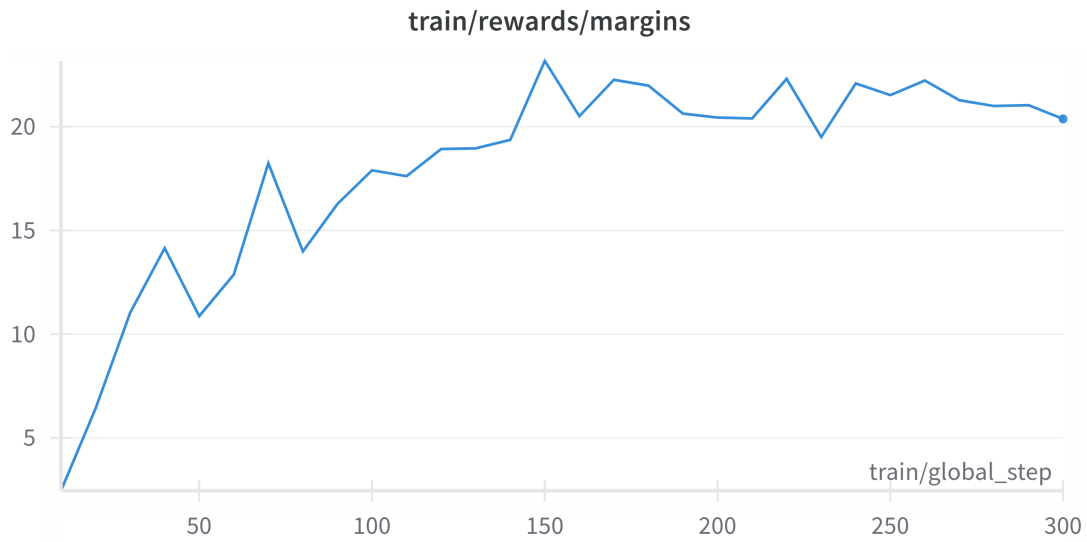

図 8: D_0 を chosen とした DPO における Rewards/margins の推移

評価者 LLM のプロンプト

You will be given responses written by an AI assistant mimicing the character. Your task is to rate the performance of character using the specific criterion by following the evaluation steps. Below is the data of character who mimiced by assistant:
***
[Profile]
profile
***
[Examples of Conversation]
conversation_example
***
[Interactions]
[user]
question
[assistant]
answer
***
[Evaluation Criterion]
Personality (1-10): Is the response reflects the personalities and preferences of the character?
Evaluation Steps]
1. Read through the profile and examples of conversation, Write the personalities and tones, preferences of the real character.
2. Read through the interactions and indentify the personalities and tones and preferences of the AI assistant.
3. After having a clear understanding of the interactions, compare the responses to the profile. Look for any consistencies or inconsistencies. Does the response reflect the character's personalities and tones, preferences? And also, does the responce have a natural sentence or length as a response?
4. Use the given scale from 1-10 to rate how well the response reflects the personalities and tones and preferences of the character. 1 being not at all reflective of the character's personalities, and 10 being perfectly reflective of the character's charactiristics.
***
First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line. Please responce in Japanese.

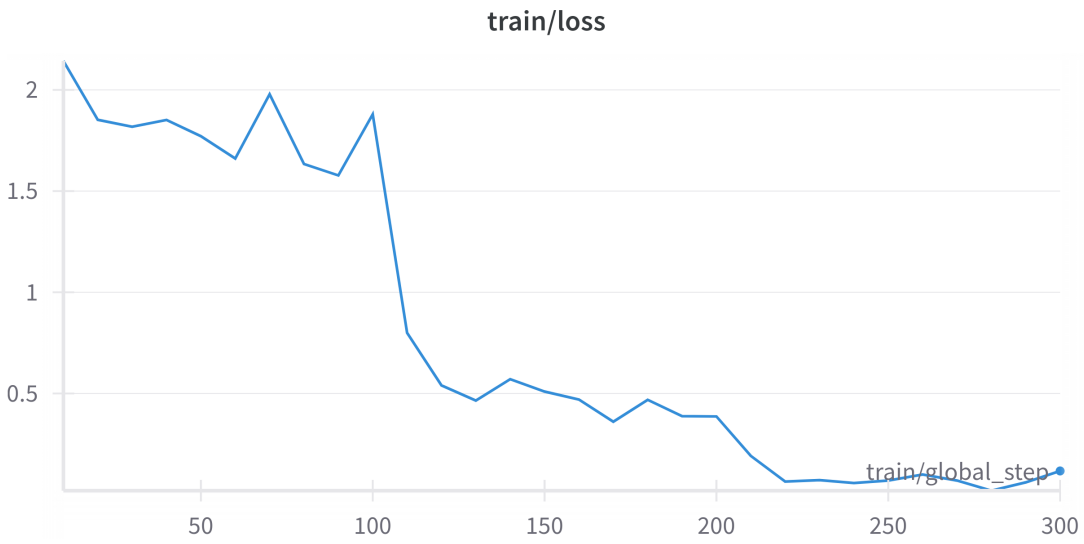

図 9: D_2 を用いた SFT における Rewards/margins の推移

表 25 に LLM の評価結果を示す. D_0 を用いて SFT や DPO を適用した場合には, データセットのクエリに対する生成結果, オリジナルのクエリに対する生成結果ともにベースモデルより下回る数値となった. 一方で, D_1 を用いて SFT や DPO を適用した場合ではデータセットのクエリに対する生成結果, オリジナルのクエリに対する生成結果ともにベースモデルより上回る数値となった.

4　ベースモデルへの回帰現象の検証

D_1 で SFT 後, D_0 を chosen として DPO を適用したモデル $M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}1}})}$ に関して, ベースモデルへの回帰現象を調べる.

4.1　重みのパラメータからの評価

表 26 に, $(M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}1}})}, M_{\mathrm{base}})$, $(M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}1}})}, M_{\mathrm{SFT}_{\mathrm{D}1}})$, $(M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}1}})}, M_{\mathrm{DPO}_{\mathrm{D}0}})$, $(M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}1}})}, M_{\mathrm{SFT}_{\mathrm{D}0}})$ 間の L1 ノルム, L2 ノルム の値を示す.

ベースモデル M_{base} や, 最終的に学習に用いたデータセットが同じの $M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}0}}$ との L1 ノルム, L2 ノルムは非常に大きな値となった. 2 章で見られたように SFT は学習前のモデルと大きく重みのパラメータを変更してモデルの出力を調整するため, 一度 D_1 で SFT を適用したことで DPO を追加で適用したとしてもベースモデルの回帰や, 別のモデルと似た重みとはならなかったと考えられる. また, モデルの定量的な評価の際に考察したように DPO で追加学習前のモデル $M_{\mathrm{SFT}_{\mathrm{D}1}}$ とは値は小さくなった. 2 章の考察で述べたように, DPO は SFT と比較して学習前のモデルから大きく重みの変化を与えないため $M_{\mathrm{SFT}_{\mathrm{D}1}}$ とノルムは小さくなっていると考えられる.

4.2　生成結果からの評価

M_{base}, $M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}1}}$ に対して, $M_{\mathrm{DPO}_{\mathrm{D}0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}1}})}$ の生成結果の類似度を確認する.
表 27 に結果を示す.

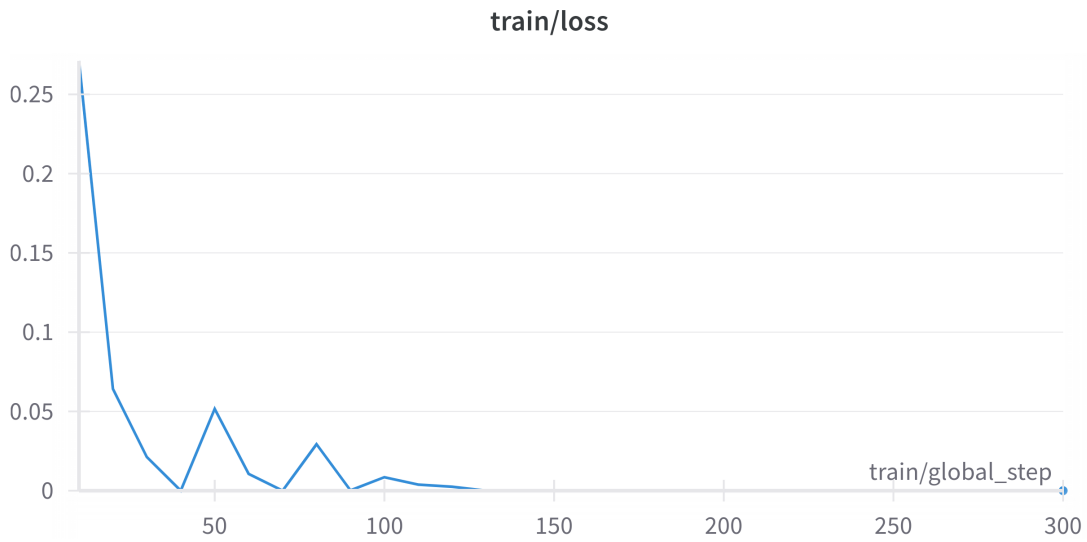

図 10: D_2 で学習後のモデル に対して D_1 を chosen とした DPO を適用した Loss の推移

M_{base}, $M_{\text{DPO}_{\text{D}_0}}$ を比較すると, M_{base} との類似度が両指標とも大きくなっている. このことから, $M_{\text{DPO}_{\text{D}_0}(\text{M}_{\text{SFT}_{\text{D}_1}})}$ の生成結果は, $M_{\text{DPO}_{\text{D}_0}}$ の生成結果に比べてベースモデル M_{base} の生成結果が語彙や文の共通部分が多くなっていることがわかる.

また, $M_{\text{SFT}_{\text{D}_0}}$, $M_{\text{SFT}_{\text{D}_1}}$ を比較すると, BLEU は $M_{\text{SFT}_{\text{D}_1}}$ が大きい一方で, BERTScore は $M_{\text{SFT}_{\text{D}_0}}$ の方が大きくなっている. このことから, $M_{\text{DPO}_{\text{D}_0}(\text{M}_{\text{SFT}_{\text{D}_1}})}$ の生成結果は, 語彙などの完全一致部分は $M_{\text{SFT}_{\text{D}_1}}$ の生成結果に見られるものが多く, 意味などを考慮すると $M_{\text{SFT}_{\text{D}_0}}$ の生成結果と類似した文が多いと考えられる.

また, 生成結果に対する GPT-4o-mini による評価は データセットのクエリに対する 202 件の応答のスコアの平均は, 5.73, オリジナルのクエリに対する 100 件のスコアの平均は 6.00 となった. 表 25 を参照すると前者と最も近い結果となったのは $M_{\text{SFT}_{\text{D}_0}}$, 後者と最も近い結果となったのは $M_{\text{DPO}_{\text{D}_0}}$ となった.

5　タスクベクトルの増加傾向

D_1 で SFT 後, D_1 を chosen として DPO を適用したモデル $M_{\text{DPO}_{\text{D}_1}(\text{M}_{\text{SFT}_{\text{D}_1}})}$ に関して, L2 ノルムがどう増加するか確かめる.

表 28 に $(M_{\text{SFT}_{\text{D}_1}}, M_{\text{base}})$, $(M_{\text{DPO}_{\text{D}_1}(\text{M}_{\text{SFT}_{\text{D}_1}})}, M_{\text{base}})$, $(M_{\text{DPO}_{\text{D}_1}(\text{M}_{\text{SFT}_{\text{D}_1}})}, M_{\text{SFT}_{\text{D}_1}})$ 間の L1 ノルム, L2 ノルムを示す. $M_{\text{DPO}_{\text{D}_1}(\text{M}_{\text{SFT}_{\text{D}_1}})}$, $M_{\text{SFT}_{\text{D}_1}}$ に関して, ベースモデルとの L1 ノルム, L2 ノルムは比較的大きな値を取っており, $M_{\text{DPO}_{\text{D}_1}(\text{M}_{\text{SFT}_{\text{D}_1}})}$, $M_{\text{SFT}_{\text{D}_1}}$ 間では小さな値となっている.

また, $M_{\text{SFT}_{\text{D}_1}}$ にさらに D_1 で DPO を適用することの応答生成の性能強化に関して考察すると, BLEU スコアに関しては 0.0947, BERTScore は 0.7787 となった. 評価者 LLM による生成結果の評価に関しては, データセットのクエリ 202 件に対する応答の評価の平均が 8.3069, オリジナルのクエリ 100 件に対する応答の評価の平均が 8.13 となった.

参考文献

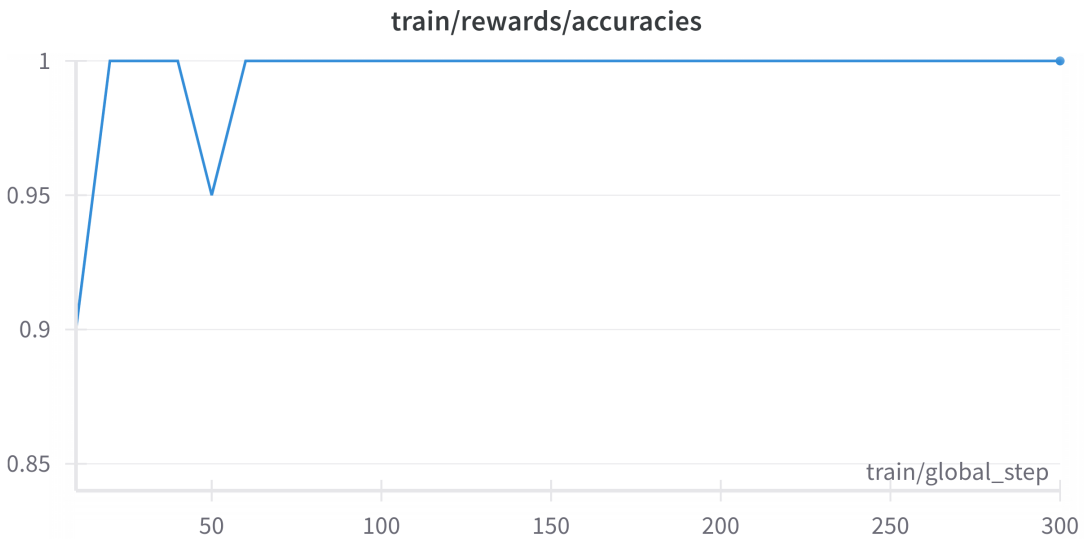

図 11: D_2 で学習後のモデル に対して D_1 を chosen とした DPO を適用した Rewards/accuracies の推移

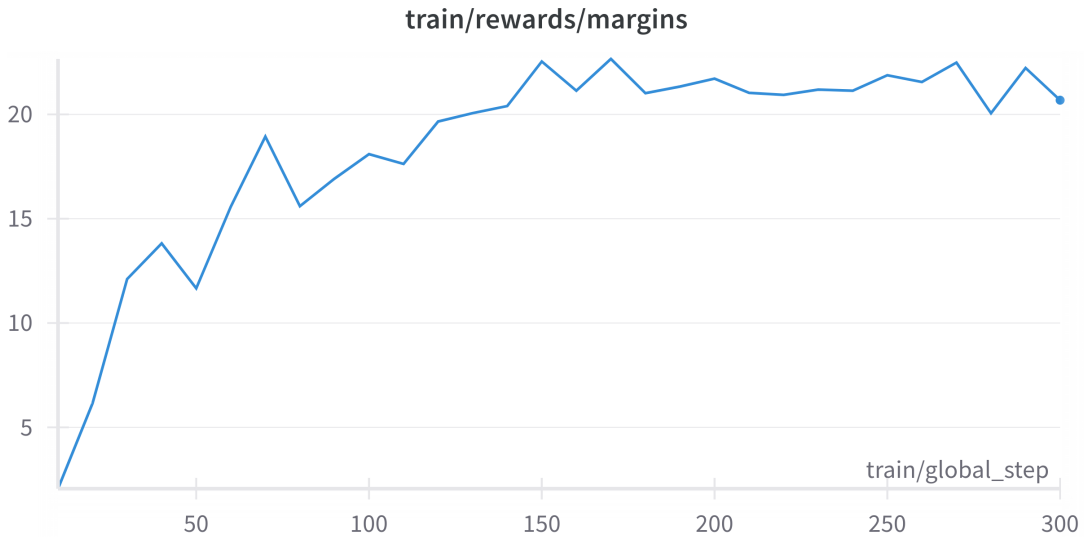

図 12: D_2 で学習後のモデル に対して D_1 を chosen とした DPO を適用した Rewards/margins の推移

図 13: D_2 で学習後のモデル に対して D_0 を chosen とした DPO を適用した Loss の推移

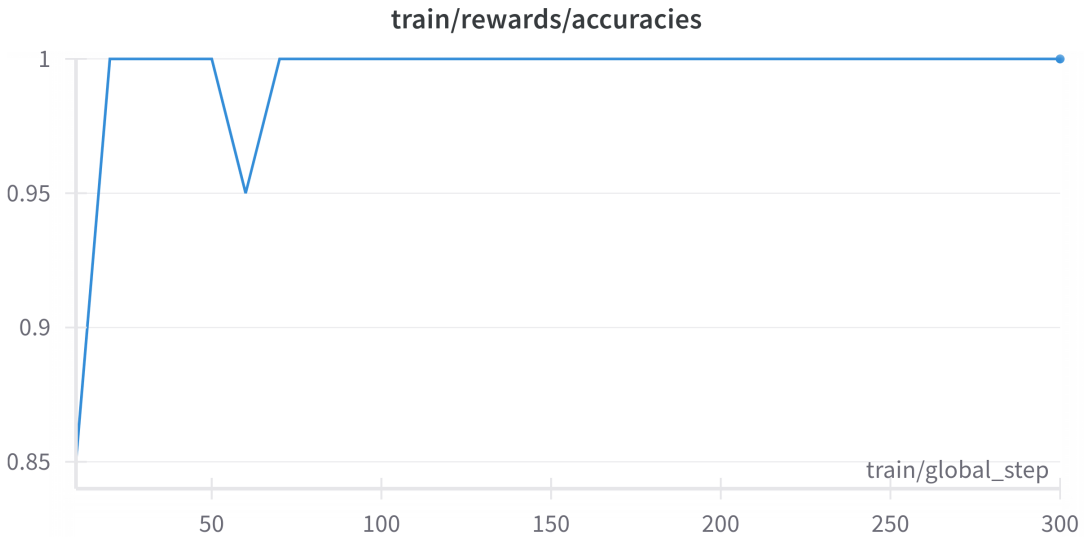

図 14: D_2 で学習後のモデル に対して D_0 を chosen とした DPO を適用した Rewards/accuracies の推移

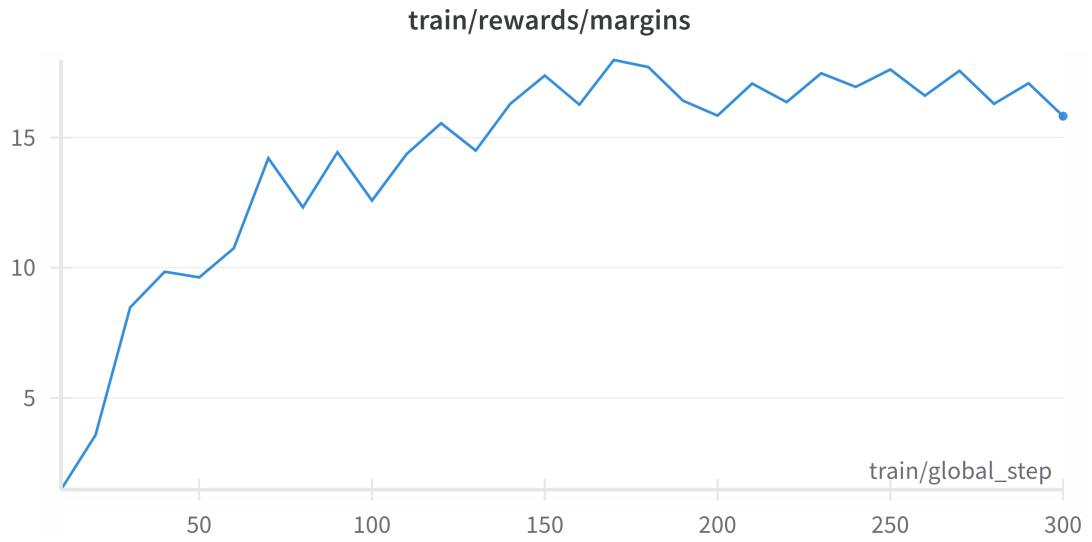

図 15: D_2 で学習後のモデル に対して D_0 を chosen とした DPO を適用した Rewards/margins の推移

```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(128256, 4096)
    (layers): ModuleList(
      (0-31): 32 x LlamaDecoderLayer(
        (self_attn): LlamaSdpaAttention(
          (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (k_proj): Linear(in_features=4096, out_features=1024, bias=False)
          (v_proj): Linear(in_features=4096, out_features=1024, bias=False)
          (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear(in_features=4096, out_features=14336, bias=False)
          (up_proj): Linear(in_features=4096, out_features=14336, bias=False)
          (down_proj): Linear(in_features=14336, out_features=4096, bias=False)
          (act_fn): SiLU()
        )
        (input_layernorm): LlamaRMSNorm((4096,), eps=1e-05)
        (post_attention_layernorm): LlamaRMSNorm((4096,), eps=1e-05)
      )
    )
    (norm): LlamaRMSNorm((4096,), eps=1e-05)
    (rotary_emb): LlamaRotaryEmbedding()
  )
  (lm_head): Linear(in_features=4096, out_features=128256, bias=False)
)
```

図 16: Llama3 のアーキテクチャ

表 5: $M_{\mathrm{SFT}_{\mathrm{D}0}}$, M_{base} 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	26506.5	2.2401313525243336
model.layers.1	26531.5	2.3194574392516882
model.layers.2	26726.0	2.2646280133970125
model.layers.3	26819.5	2.2781576858080412
model.layers.4	26810.5	2.274685546605838
model.layers.5	27126.5	2.303309969675434
model.layers.6	27656.0	2.3565017687108787
model.layers.7	27674.5	2.3608749371260647
model.layers.8	27684.5	2.3570715162702425
model.layers.9	27982.0	2.402238945884651
model.layers.10	28338.5	2.440753037614621
model.layers.11	28259.5	2.4434023090252657
model.layers.12	28394.0	2.448816877915231
model.layers.13	28113.0	2.4315833640346365
model.layers.14	28275.5	2.4511337459683835
model.layers.15	28366.5	2.4628093079854967
model.layers.16	28338.5	2.4673156434422814
model.layers.17	28421.0	2.455785778332976
model.layers.18	28338.0	2.441178114341004
model.layers.19	28259.5	2.4320100434748415
model.layers.20	28304.5	2.425023155495118
model.layers.21	28153.0	2.418142284766345
model.layers.22	28091.5	2.41634954366861
model.layers.23	28297.5	2.4265831253147914
model.layers.24	28192.0	2.419353526125316
model.layers.25	28257.5	2.424066548759027
model.layers.26	28332.5	2.4265579724684097
model.layers.27	28268.0	2.4345434794073815
model.layers.28	28171.5	2.4178898664434656
model.layers.29	28358.0	2.4527517630077242
model.layers.30	28068.5	2.445898991168687
model.layers.31	28270.5	2.4779791886484843
total	893386.5	13.584439160476501

表 6: $M_{\mathrm{SFT}_{\mathrm{D1}}}$, M_{base} 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	27233.0	2.313766545051596
model.layers.1	27102.0	2.3806470693385022
model.layers.2	27404.5	2.3292782378614625
model.layers.3	27238.5	2.3167981593003524
model.layers.4	27486.5	2.336734291364232
model.layers.5	27667.0	2.36025438909665
model.layers.6	28520.5	2.441178114341004
model.layers.7	28401.5	2.436573341485435
model.layers.8	28554.0	2.447121435546773
model.layers.9	28857.5	2.4802934239218555
model.layers.10	29240.5	2.5190436785985866
model.layers.11	29330.5	2.540876755669192
model.layers.12	29640.0	2.5649997334917014
model.layers.13	29266.0	2.537535598580343
model.layers.14	29236.0	2.5395552406499253
model.layers.15	29080.0	2.522022145645236
model.layers.16	28925.0	2.507240491618724
model.layers.17	29010.5	2.4986812928272784
model.layers.18	29039.5	2.509673666846449
model.layers.19	29041.0	2.50309866559181
model.layers.20	29328.5	2.516571056579269
model.layers.21	28930.0	2.489675948416781
model.layers.22	29211.0	2.515916132222018
model.layers.23	28913.0	2.4799488874521183
model.layers.24	29140.0	2.508068036046371
model.layers.25	28953.0	2.4921262727588664
model.layers.26	29006.5	2.503440016070387
model.layers.27	28941.5	2.4997070140818503
model.layers.28	29069.5	2.4870268865404332
model.layers.29	28951.0	2.518219741019933
model.layers.30	29158.5	2.5456764818020376
model.layers.31	28918.0	2.537126665886037
total	918794.5	14.002781183014145

表 7: $M_{\mathrm{DPO}_{\mathrm{D}0}}$, M_{base} 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	4682.0625	0.04123881332836149
model.layers.1	4805.25	0.07502072365867637
model.layers.2	4842.1875	0.06273796555824628
model.layers.3	4790.25	0.0639764906202372
model.layers.4	4884.125	0.07380319002695139
model.layers.5	4920.125	0.06575500689686554
model.layers.6	4867.1875	0.06507298040491212
model.layers.7	4982.375	0.06942141998918064
model.layers.8	5077.375	0.07380197859285326
model.layers.9	5178.375	0.09415855395273816
model.layers.10	5389.5	0.11215598171028791
model.layers.11	5564.6875	0.13653443037097748
model.layers.12	5501.75	0.14042076538774917
model.layers.13	5518.875	0.14939052421311871
model.layers.14	5182.4375	0.12412960284439457
model.layers.15	5074.4375	0.11627451976479285
model.layers.16	4840.5625	0.09287681842733461
model.layers.17	4853.6875	0.10967105579192545
model.layers.18	4612.375	0.0770900843645298
model.layers.19	4593.375	0.07240735152208896
model.layers.20	4733.4375	0.07025058987732981
model.layers.21	4727.1875	0.08890071697482363
model.layers.22	4955.0625	0.10945399020586168
model.layers.23	4726.0	0.07252949147720292
model.layers.24	4854.625	0.11366155518554752
model.layers.25	4812.6875	0.09594229571256829
model.layers.26	4617.0	0.07284028730504429
model.layers.27	4974.3125	0.10980467186666111
model.layers.28	4789.25	0.07321817640274447
model.layers.29	4699.375	0.08074229143982183
model.layers.30	4685.25	0.10154812050774749
model.layers.31	4679.9375	0.11393601184402531
total	157415.125	0.5359644668956891

表 8: $M_{\mathrm{DPO_{D1}}}$, M_{base} 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	4514.9375	0.035282332776249835
model.layers.1	4580.0625	0.07046153909198222
model.layers.2	4518.625	0.04240463327967811
model.layers.3	4516.125	0.053553123979957463
model.layers.4	4561.9375	0.0539920731837073
model.layers.5	4630.25	0.06020232578740129
model.layers.6	4611.375	0.059178297574178966
model.layers.7	4593.5	0.05603811836590746
model.layers.8	4687.375	0.05756077995588349
model.layers.9	4605.0	0.0641193416842948
model.layers.10	4660.375	0.06113955914499096
model.layers.11	4701.0	0.05929200244128473
model.layers.12	4699.875	0.07383145117961974
model.layers.13	4894.625	0.07662788246296551
model.layers.14	4796.8125	0.08158352694100413
model.layers.15	4863.375	0.09664114752268173
model.layers.16	4885.125	0.09552015603150191
model.layers.17	4868.875	0.09188092219151074
model.layers.18	4958.4375	0.12034448966267292
model.layers.19	4709.625	0.08434519771253576
model.layers.20	5065.625	0.10436290693753171
model.layers.21	5205.875	0.13546109082174643
model.layers.22	5310.125	0.1418526363812221
model.layers.23	5274.875	0.1547608943179614
model.layers.24	5277.25	0.16220935358248745
model.layers.25	5218.5	0.13393689084326157
model.layers.26	5289.5625	0.12507579412915423
model.layers.27	5361.4375	0.16717661884585688
model.layers.28	5367.125	0.13153992375373225
model.layers.29	4968.1875	0.0987216058784708
model.layers.30	5189.375	0.1693436881538837
model.layers.31	5145.6875	0.13721398169673268
total	156530.9375	0.5838905182501243

表 9: $M_{\mathrm{SFT}_{\mathrm{D2}}}$, M_{base} 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	25799.5	2.18205236854263
model.layers.1	25552.0	2.2295347486218957
model.layers.2	25650.0	2.164725191854615
model.layers.3	25782.0	2.1813809505821076
model.layers.4	26007.5	2.1969661254216004
model.layers.5	26399.0	2.236395502533038
model.layers.6	27247.0	2.318246660328663
model.layers.7	27128.0	2.3141358234500844
model.layers.8	27193.0	2.3092513075805545
model.layers.9	27475.5	2.3493163067954472
model.layers.10	27968.0	2.395088765643666
model.layers.11	28040.5	2.411444271917247
model.layers.12	28187.0	2.4307548901395837
model.layers.13	28132.5	2.436698586202241
model.layers.14	28180.5	2.4417031069521946
model.layers.15	28231.5	2.4415781189504258
model.layers.16	28195.5	2.4396024586497695
model.layers.17	28228.5	2.4280415445761014
model.layers.18	28161.0	2.4233110661732016
model.layers.19	28301.0	2.4376752741469687
model.layers.20	28388.0	2.4316586658467303
model.layers.21	28240.5	2.432612287040929
model.layers.22	28255.0	2.4333147161459614
model.layers.23	28160.0	2.4208919375795155
model.layers.24	28198.0	2.421648174862009
model.layers.25	28234.0	2.4236132874671856
model.layers.26	28364.0	2.4355460918703633
model.layers.27	28515.0	2.464964463169297
model.layers.28	28363.5	2.431282133463741
model.layers.29	28403.0	2.462140083151749
model.layers.30	28267.5	2.47571209768523
model.layers.31	28315.0	2.484645229014094
total	885563.0	13.461171011468505

表 10: $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, M_{base} 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	26517.0	2.254038259782318
model.layers.1	26371.0	2.314214946869132
model.layers.2	26454.5	2.2446496716497877
model.layers.3	26572.5	2.2606356700743486
model.layers.4	26834.5	2.278867548737201
model.layers.5	27231.5	2.318641549013668
model.layers.6	28063.0	2.3981702823454802
model.layers.7	27913.0	2.3912631909594353
model.layers.8	27933.5	2.3823898190682398
model.layers.9	28265.0	2.42605486077706
model.layers.10	28778.5	2.474207772512749
model.layers.11	28904.5	2.497288568690551
model.layers.12	29045.0	2.5141445566161265
model.layers.13	28897.5	2.5109866011067243
model.layers.14	28962.5	2.5151154375222027
model.layers.15	29040.5	2.5188982974333243
model.layers.16	29013.0	2.5185590420789223
model.layers.17	28984.0	2.499145361731186
model.layers.18	28928.5	2.4994872521054194
model.layers.19	29089.0	2.513391865960221
model.layers.20	29197.5	2.5101113879044292
model.layers.21	29096.0	2.5175652455547204
model.layers.22	29204.5	2.5243443610420906
model.layers.23	29048.5	2.5059986625844397
model.layers.24	29041.0	2.504512528978943
model.layers.25	29124.5	2.510378846533925
model.layers.26	29289.0	2.523715638696246
model.layers.27	29373.0	2.546922928996871
model.layers.28	29242.0	2.5158676125245543
model.layers.29	29274.0	2.542365639886403
model.layers.30	29307.5	2.562118874096204
model.layers.31	29038.0	2.5906655229038154
total	912034.0	13.9206517770698345

表 11: $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, M_{base} 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	26415.5	2.244268959831352
model.layers.1	26183.0	2.2943225613729643
model.layers.2	26335.5	2.234101818964391
model.layers.3	26466.0	2.2495252320281383
model.layers.4	26720.0	2.267348482396343
model.layers.5	27139.5	2.3108630185766095
model.layers.6	27895.5	2.381877377983594
model.layers.7	27805.0	2.3821848558623655
model.layers.8	27884.0	2.378171710283711
model.layers.9	28101.0	2.4098998293497593
model.layers.10	28569.5	2.451880655741629
model.layers.11	28601.0	2.46699403532163
model.layers.12	28750.0	2.483760733809217
model.layers.13	28766.5	2.492934351259074
model.layers.14	28874.0	2.5053896980435
model.layers.15	28858.0	2.5020743347395777
model.layers.16	28910.5	2.5039275788873168
model.layers.17	28911.0	2.492297705243597
model.layers.18	28891.5	2.495259176794066
model.layers.19	28961.0	2.501293610626809
model.layers.20	29197.5	2.5122259645332465
model.layers.21	29005.5	2.5064613766847677
model.layers.22	29054.5	2.5107435169312895
model.layers.23	28970.0	2.5002929515858736
model.layers.24	29107.5	2.5125903668450413
model.layers.25	28991.5	2.4951124097923927
model.layers.26	29173.5	2.5127604031721766
model.layers.27	29353.0	2.5497730827345206
model.layers.28	29154.0	2.5055115027923938
model.layers.29	29208.5	2.5336360282087678
model.layers.30	29149.0	2.5524526731043418
model.layers.31	29330.0	2.577367312903809
total	908732.5	13.85587346550868

表 12: $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}$, $M_{\mathrm{SFT}_{\mathrm{D}_2}}$ 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	5235.75	0.06629395807966705
model.layers.1	5553.125	0.17041365393079547
model.layers.2	5308.25	0.08503281633027253
model.layers.3	5252.625	0.0841106087106773
model.layers.4	5391.0	0.1010332254770631
model.layers.5	5421.0	0.10081855024521709
model.layers.6	5587.0	0.10875553730865473
model.layers.7	5662.25	0.1274086628793701
model.layers.8	5655.8125	0.12158962976695016
model.layers.9	5620.0625	0.13080812276122672
model.layers.10	5882.1875	0.14007652132868365
model.layers.11	5881.625	0.15624008147230237
model.layers.12	5913.5625	0.16550681738051237
model.layers.13	5760.125	0.1335998087618938
model.layers.14	5578.5	0.12735157573969702
model.layers.15	5674.5	0.14398212764350132
model.layers.16	5824.0	0.15211077751684718
model.layers.17	5710.875	0.12463803335467272
model.layers.18	5658.1875	0.15523734962903946
model.layers.19	5654.8125	0.1478801100066449
model.layers.20	5959.9375	0.15333314100889225
model.layers.21	5998.6875	0.16928296167988172
model.layers.22	6259.25	0.18989809616283554
model.layers.23	6180.3125	0.1827600909368164
model.layers.24	6035.25	0.1792476079400782
model.layers.25	6091.4375	0.18350493535153314
model.layers.26	6228.5	0.18319463974552896
model.layers.27	6421.3125	0.22718410064378877
model.layers.28	6198.4375	0.1583823613335519
model.layers.29	6259.5625	0.1861535043813329
model.layers.30	6578.125	0.24858839143148295
model.layers.31	6329.0	0.5099919426973439
total	186765.0625	0.9965761998485839

表 13: $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{SFT}_{\mathrm{D2}}}$ 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	5149.1875	0.07164934643955131
model.layers.1	5127.9375	0.10112138467746036
model.layers.2	5237.6875	0.07620473299721453
model.layers.3	5249.0	0.08533299853090331
model.layers.4	5457.8125	0.09596372662150737
model.layers.5	5557.1875	0.11516788956967715
model.layers.6	5378.5	0.09595285646879026
model.layers.7	5441.8125	0.10302474032181633
model.layers.8	5514.3125	0.10928640864923689
model.layers.9	5345.875	0.10305135004305874
model.layers.10	5492.9375	0.1058469983287495
model.layers.11	5513.625	0.12104328703766908
model.layers.12	5484.0625	0.11806338726818964
model.layers.13	5380.4375	0.11037370577752198
model.layers.14	5393.8125	0.11357945621186329
model.layers.15	5420.5	0.1379485270798664
model.layers.16	5405.375	0.12201268170842186
model.layers.17	5168.125	0.09930072912004208
model.layers.18	5244.5	0.12709975944214152
model.layers.19	5017.6875	0.10505028111140147
model.layers.20	5336.9375	0.12364270500632756
model.layers.21	5313.8125	0.13233502896341193
model.layers.22	5534.3125	0.1529283311571797
model.layers.23	5388.875	0.1396753245259631
model.layers.24	5553.8125	0.14525115656433296
model.layers.25	5420.875	0.13058648194049216
model.layers.26	5695.6875	0.15643033111886429
model.layers.27	5848.3125	0.19994222879496784
model.layers.28	5658.0	0.1437223324894262
model.layers.29	5629.0625	0.15128269913989342
model.layers.30	6011.4375	0.18038803871914416
model.layers.31	5532.75	0.16326127363596757
total	173904.25	0.7142688962114501

表 14: SFT を施したモデルとベースモデルとの L2 ノルムの平均値とレイヤーごとの平均値からの差分 (太字は上位 10 層)

Key	$M_{\mathrm{SFT}_{\mathrm{D}_0}}$	$M_{\mathrm{SFT}_{\mathrm{D}_1}}$	$M_{\mathrm{SFT}_{\mathrm{D}_2}}$	$M_{\mathrm{DPO}_{\mathrm{D}_0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}$	$M_{\mathrm{DPO}_{\mathrm{D}_0}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}_2}})}$
model.layers.0	-0.16040	-0.16057	-0.19563	-0.20486	-0.20315
model.layers.1	-0.08107	-0.09369	-0.14815	-0.14468	-0.15309
model.layers.2	-0.13590	-0.14506	-0.21296	-0.21425	-0.21331
model.layers.3	-0.12237	-0.15754	-0.19630	-0.19826	-0.19789
model.layers.4	-0.12585	-0.13761	-0.18072	-0.18003	-0.18007
model.layers.5	-0.09722	-0.11409	-0.14129	-0.14026	-0.13655
model.layers.6	-0.04403	-0.03316	-0.05944	-0.06073	-0.06554
model.layers.7	-0.03966	-0.03777	-0.06355	-0.06763	-0.06523
model.layers.8	-0.04346	-0.02722	-0.06843	-0.07651	-0.06924
model.layers.9	0.00171	0.00595	-0.02837	-0.03284	-0.03752
model.layers.10	0.04022	**0.04470**	0.01740	0.01531	0.00446
model.layers.11	**0.04287**	**0.06654**	0.03376	0.03839	0.01958
model.layers.12	**0.04829**	**0.09066**	0.05307	0.05525	0.03635
model.layers.13	0.03105	**0.06320**	**0.05901**	0.05209	0.04552
model.layers.14	**0.05060**	**0.06522**	**0.06402**	0.05622	0.05797
model.layers.15	**0.06228**	**0.04768**	**0.06389**	**0.06000**	0.05466
model.layers.16	**0.06678**	0.03290	**0.06192**	**0.05966**	0.05651
model.layers.17	**0.05526**	0.02434	0.05036	0.04025	0.04488
model.layers.18	**0.04065**	0.03533	0.04563	0.04059	0.04784
model.layers.19	0.03148	0.02876	**0.05999**	0.05449	0.05388
model.layers.20	0.02449	**0.04223**	0.05397	0.05121	**0.06481**
model.layers.21	0.01761	0.01534	0.05493	**0.05867**	**0.05905**
model.layers.22	0.01582	0.04158	0.05563	**0.06545**	**0.06333**
model.layers.23	0.02605	0.00561	0.04321	0.04710	0.05288
model.layers.24	0.01882	0.03373	0.04396	0.04562	**0.06517**
model.layers.25	0.02354	0.01779	0.04593	0.05148	0.04770
model.layers.26	0.02603	0.02910	**0.05786**	**0.06482**	**0.06534**
model.layers.27	0.03401	0.02537	**0.08728**	**0.08803**	**0.10236**
model.layers.28	0.01736	0.01269	0.05360	**0.05697**	**0.05810**
model.layers.29	**0.05222**	**0.04388**	**0.08445**	**0.08347**	**0.08622**
model.layers.30	**0.04537**	**0.07134**	**0.09803**	**0.10322**	**0.10504**
model.layers.31	**0.07745**	**0.06279**	**0.10696**	**0.13177**	**0.12995**
average	2.4005	2.4743	2.3777	2.4588	2.4474

表 15: DPO を適用モデルと DPO を適用する前のモデルとの L2 ノルムの平均値とレイヤーごとの平均値からの差分 (太字は上位 10 層)

Key	$M_{\mathrm{DPO}_{\mathrm{D}0}}$, M_{base}	$M_{\mathrm{DPO}_{\mathrm{D}1}}$, M_{base}	$M_{\mathrm{DPO}_{\mathrm{D}0(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}}$, $M_{\mathrm{SFT}_{\mathrm{D}2}}$	$M_{\mathrm{DPO}_{\mathrm{D}1(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}}$, $M_{\mathrm{SFT}_{\mathrm{D}2}}$
model.layers.0	-0.04997	-0.06021	-0.09353	-0.05137
model.layers.1	-0.01619	-0.02503	**0.01059**	-0.02189
model.layers.2	-0.02847	-0.05308	-0.07479	-0.04681
model.layers.3	-0.02723	-0.04194	-0.07571	-0.03768
model.layers.4	-0.01741	-0.04150	-0.05879	-0.02705
model.layers.5	-0.02546	-0.03529	-0.05901	-0.00785
model.layers.6	-0.02614	-0.03631	-0.05107	-0.02706
model.layers.7	-0.02179	-0.03945	-0.03242	-0.01999
model.layers.8	-0.01741	-0.03793	-0.03824	-0.01373
model.layers.9	0.00295	-0.03137	-0.02902	-0.01996
model.layers.10	**0.02094**	-0.03435	-0.01975	-0.01717
model.layers.11	**0.04532**	-0.03620	-0.00359	-0.00197
model.layers.12	**0.04921**	-0.02166	0.00568	-0.00495
model.layers.13	**0.05818**	-0.01886	-0.02623	-0.01264
model.layers.14	**0.03292**	-0.01391	-0.03247	-0.00944
model.layers.15	**0.02506**	0.00115	-0.01584	**0.01493**
model.layers.16	0.00167	0.00003	-0.00771	-0.00100
model.layers.17	**0.01846**	-0.00361	-0.03519	-0.02372
model.layers.18	-0.01412	0.02486	-0.00459	0.00408
a model.layers.19	-0.01880	-0.01114	-0.01195	-0.01797
model.layers.20	-0.02096	0.00887	-0.00649	0.00063
model.layers.21	-0.00231	**0.03997**	0.00946	0.00932
model.layers.22	0.01824	**0.04636**	**0.03007**	**0.02991**
model.layers.23	-0.01868	**0.05927**	**0.02293**	**0.01666**
model.layers.24	**0.02245**	**0.06672**	**0.01942**	**0.02223**
model.layers.25	0.00473	**0.03845**	**0.02368**	0.00757
model.layers.26	-0.01837	**0.02959**	**0.02337**	**0.03341**
model.layers.27	**0.01859**	**0.07169**	**0.06736**	**0.07693**
model.layers.28	-0.01799	**0.03605**	-0.00144	**0.02071**
model.layers.29	-0.01047	0.00323	**0.02633**	**0.02827**
model.layers.30	0.01034	**0.07385**	**0.08876**	**0.05737**
model.layers.31	**0.02272**	**0.04172**	**0.35017**	**0.04025**
average	0.09121	0.09549	0.15983	0.12302

表 16: 学習方法が同一でデータセットが異なる 2 モデル間の L1 ノルム, L2 ノルム

モデル	L1 ノルム	L2 ノルム
$M_{\mathrm{SFT}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}1}}$	1073358.0	16.658565739582745
$M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{DPO}_{\mathrm{D}1}}$	283154.5	2.7612353904133857
$M_{\mathrm{DPO}_{\mathrm{D}0(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}}$, $M_{\mathrm{DPO}_{\mathrm{D}1(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D}2}})}}$	333209.25	3.789889467585258

表 17: $M_{\mathrm{SFT}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}1}}$ 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	32760.0	2.847717870561004
model.layers.1	31470.0	2.830023534023171
model.layers.2	31862.5	2.7632404974730305
model.layers.3	31730.5	2.7561853379603303
model.layers.4	31711.0	2.7526842368518225
model.layers.5	31705.5	2.751752814968579
model.layers.6	31737.5	2.762356826040944
model.layers.7	31680.0	2.7564510626777867
model.layers.8	31984.0	2.7815572161785562
model.layers.9	32126.0	2.8031528705414908
model.layers.10	32176.0	2.811262748692747
model.layers.11	32082.5	2.8080694268927897
model.layers.12	31984.0	2.791938730524006
model.layers.13	32681.5	2.8773979469848276
model.layers.14	32977.0	2.890181553823652
model.layers.15	33543.0	2.949029898754843
model.layers.16	34086.5	3.003944347125126
model.layers.17	34604.0	3.0392131228740773
model.layers.18	35219.5	3.0954264712737887
model.layers.19	34911.0	3.069407866093638
model.layers.20	35735.5	3.1254687148490223
model.layers.21	35152.0	3.0841074478688317
model.layers.22	35395.5	3.1083541037621503
model.layers.23	35715.0	3.1285721770781474
model.layers.24	35373.0	3.116139783199319
model.layers.25	35406.0	3.1057807429755084
model.layers.26	35001.0	3.069527173737431
model.layers.27	34953.5	3.0697657751142025
model.layers.28	35543.5	3.1061541106929966
model.layers.29	34426.5	3.044630583605341
model.layers.30	34176.5	3.036601268791311
model.layers.31	33448.0	2.9895856017703357
total	1073358.0	16.658565739582745

表 18: $M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{DPO}_{\mathrm{D}1}}$ 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	8246.625	0.35940286279556055
model.layers.1	8482.875	0.42431837285577056
model.layers.2	8409.0	0.40373005962438757
model.layers.3	8357.0	0.41534742680656395
model.layers.4	8480.875	0.4292817020251428
model.layers.5	8502.375	0.4201398416863909
model.layers.6	8422.25	0.40898444495432784
model.layers.7	8476.625	0.4087506140841383
model.layers.8	8591.75	0.4160987162105647
model.layers.9	8721.125	0.4491030336694178
model.layers.10	8877.375	0.453323846792556
model.layers.11	9090.0	0.487039436513848
model.layers.12	9100.875	0.510888744209501
model.layers.13	9292.375	0.52259508353781
model.layers.14	8904.0	0.4881249749919968
model.layers.15	8905.375	0.4968083558883332
model.layers.16	8786.125	0.48132545573233243
model.layers.17	8814.5	0.49510466730964764
model.layers.18	8680.625	0.49147005622124057
model.layers.19	8432.25	0.4313289846430032
model.layers.20	8892.375	0.491299266722862
model.layers.21	8991.375	0.5200082983675297
model.layers.22	9316.125	0.5510409128608187
model.layers.23	9106.5	0.544820359047652
model.layers.24	9193.25	0.5702414272388103
model.layers.25	9140.0	0.5315802052827376
model.layers.26	9005.625	0.5033533182450288
model.layers.27	9452.125	0.5878224887840823
model.layers.28	9288.125	0.5269746972511762
model.layers.29	8847.0	0.4919394216072239
model.layers.30	9148.625	0.5982144809226427
model.layers.31	9199.375	0.5890153537167452
total	283154.5	2.7612353904133857

表 19: $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$ 間の L1 ノルム, L2 ノルム

Layer	**L1 ノルム**	**L2 ノルム**
model.layers.0	9391.625	0.48659964339450573
model.layers.1	9767.875	0.583098591393562
model.layers.2	9649.75	0.5389280289706189
model.layers.3	9627.625	0.5457411865417419
model.layers.4	9954.875	0.5813098507492583
model.layers.5	10051.875	0.5929527954045962
model.layers.6	10029.75	0.5905127363131526
model.layers.7	10119.75	0.602848412535103
model.layers.8	10165.5	0.6069318260357403
model.layers.9	9997.625	0.6002706553040392
model.layers.10	10334.875	0.6155288233883731
model.layers.11	10307.125	0.6246611630344211
model.layers.12	10364.75	0.6478165560428377
model.layers.13	10178.25	0.6061739844418123
model.layers.14	10000.125	0.6010422872433249
model.layers.15	10121.25	0.6312281236540375
model.layers.16	10366.75	0.6478489422335465
model.layers.17	10040.0	0.6047248454230031
model.layers.18	10050.625	0.6433568253782206
model.layers.19	9882.5	0.607776599598513
model.layers.20	10423.625	0.6668599166522906
model.layers.21	10458.25	0.682471176489001
model.layers.22	10943.625	0.7320024840932918
model.layers.23	10763.75	0.7139413275881158
model.layers.24	10765.25	0.7163044331955687
model.layers.25	10794.375	0.7229729035973185
model.layers.26	11168.0	0.7567244121458617
model.layers.27	11487.875	0.827590481977273
model.layers.28	11206.875	0.75011697492739
model.layers.29	11291.625	0.7760815364828177
model.layers.30	12041.75	0.8870800475392215
model.layers.31	11461.75	0.9909872169016611
total	333209.25	3.789889467585258

表 20: $M_{\mathrm{SFT}_{\mathrm{D}_0}}$, $M_{\mathrm{SFT}_{\mathrm{D}_1}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2

モデル	コンフリクト数	コンフリクト率	**Conflict Limited L2**
model.layers.0	10079150	0.0462108916519953	1.6077244266450796
model.layers.1	10024245	0.04595916318221831	1.622613181000404
model.layers.2	8811511	0.04039901976965962	1.5005797219585253
model.layers.3	8680592	0.039798782276995306	1.494956119365423
model.layers.4	8579514	0.039335359815140844	1.4843904193278086
model.layers.5	8542967	0.03916779911238263	1.480406908338992
model.layers.6	8588343	0.03937583901848592	1.4850327490075765
model.layers.7	8465185	0.038811184162265255	1.479788349585951
model.layers.8	8706212	0.03991624486502347	1.501728333714607
model.layers.9	8940091	0.04098853341402582	1.5268491446593129
model.layers.10	8916216	0.0408790713028169	1.5348878582178853
model.layers.11	8852400	0.04058648767605634	1.529834326457035
model.layers.12	8585148	0.039361190580985916	1.5079238138963205
model.layers.13	9663056	0.04430318368544601	1.6242862014012494
model.layers.14	9659245	0.04428571101085681	1.6255257562939445
model.layers.15	10302303	0.047234003631161973	1.6938267539112803
model.layers.16	11030795	0.050573994094776996	1.757829861025378
model.layers.17	11562130	0.053010059052230045	1.7946487567644678
model.layers.18	12085029	0.05540744663292253	1.8564676587457025
model.layers.19	11966245	0.0548628456939554	1.8297435409863454
model.layers.20	12635255	0.0579301230560446	1.877699879090772
model.layers.21	12214439	0.0560007656616784	1.8438493087661747
model.layers.22	12344707	0.05659801844923709	1.8746011945669605
model.layers.23	12708948	0.058267990757042254	1.8928512202990229
model.layers.24	12696652	0.058211616050469484	1.901216458859301
model.layers.25	12431305	0.05699505300029343	1.8682005619964228
model.layers.26	11942011	0.054751737639377934	1.8209410797759562
model.layers.27	11963729	0.05485131033597418	1.8277577088300025
model.layers.28	12464054	0.057145200630868545	1.8588004627060566
model.layers.29	11792149	0.05406465027142019	1.8122979413233906
model.layers.30	11559803	0.05299939022153756	1.8226329764373503
model.layers.31	11049796	0.05066110988849765	1.7666272045022684
total	337843225	0.04840449301849509	9.605842846771697

表 21: $M_{\mathrm{DPO}_{\mathrm{D}_0}}$, $M_{\mathrm{DPO}_{\mathrm{D}_1}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2

モデル	コンフリクト数	コンフリクト率	**Conflict Limited L2**
model.layers.0	1459	6.689223884976526e-06	0.01546395275596984
model.layers.1	4809	2.204830545774648e-05	0.028705996133556757
model.layers.2	3113	1.4272483861502348e-05	0.021976724883726196
model.layers.3	4079	1.8701401115023476e-05	0.025239937467422043
model.layers.4	5096	2.3364143192488264e-05	0.028650919144883987
model.layers.5	4909	2.2506785504694836e-05	0.028036869570225765
model.layers.6	4415	2.0241894072769953e-05	0.02728810779770109
model.layers.7	3442	1.578088321596244e-05	0.02273786904663096
model.layers.8	3573	1.638149207746479e-05	0.023276731321392633
model.layers.9	8579	3.933300322769953e-05	0.03727204657353858
model.layers.10	5900	2.7050322769953052e-05	0.03056636724987528
model.layers.11	8291	3.8012580692488264e-05	0.03557422968766989
model.layers.12	15347	7.036293280516432e-05	0.04898716774117485
model.layers.13	18217	8.352131015258215e-05	0.05428987087384065
model.layers.14	14183	6.502622505868544e-05	0.04741359280152299
model.layers.15	18840	8.637764084507043e-05	0.05462206604365083
model.layers.16	14205	6.512709066901409e-05	0.04712103805279391
model.layers.17	19564	8.969703638497653e-05	0.055999281878635294
model.layers.18	20580	9.435519366197184e-05	0.05754887038248279
model.layers.19	11849	5.43253007629108e-05	0.04471507362166005
model.layers.20	11511	5.277563820422535e-05	0.041995026248524456
model.layers.21	25966	0.00011904892899061033	0.06365285015811903
model.layers.22	33987	0.00015582361355633804	0.07283210392255525
model.layers.23	25475	0.0001167977919600939	0.06324703091976741
model.layers.24	46823	0.00021467411238262912	0.08765974449223178
model.layers.25	21221	9.72940507629108e-05	0.058050049069931754
model.layers.26	12677	5.8121515551643194e-05	0.045344479555640235
model.layers.27	33490	0.00015354496772300468	0.0739927364083714
model.layers.28	15284	7.007409037558685e-05	0.04950820338543516
model.layers.29	12341	5.658102259389672e-05	0.04436579971574641
model.layers.30	58637	0.000268838945129108	0.09763671679648428
model.layers.31	33901	0.00015542932071596245	0.07454969006070362
total	521763	7.475560147997359e-05	0.2892127806275728

表 22: $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$ 間の M_{base} を基準とした コンフリクト 数, コンフリクト率, Conflict Limited L2

モデル	コンフリクト数	コンフリクト率	**Conflict Limited L2**
model.layers.0	1440	6.602112676056338e-06	0.01611328125
model.layers.1	25245	0.00011574328785211268	0.09969879325587881
model.layers.2	3215	1.4740133509389672e-05	0.02368415740492368
model.layers.3	3373	1.5464531983568074e-05	0.02399419229491791
model.layers.4	8062	3.6962661384976526e-05	0.03831608024658376
model.layers.5	7324	3.3579078638497655e-05	0.03539702300792379
model.layers.6	7822	3.586230927230047e-05	0.03862440490264221
model.layers.7	8095	3.7113959800469485e-05	0.03801152712007562
model.layers.8	7999	3.667381895539906e-05	0.037699771281390945
model.layers.9	8537	3.914044160798122e-05	0.03887129718772133
model.layers.10	8445	3.8718639964788734e-05	0.03909529263191851
model.layers.11	10046	4.605890551643192e-05	0.04209213781632487
model.layers.12	13853	6.351324090375587e-05	0.049869271085593365
model.layers.13	8222	3.76962294600939e-05	0.03888969344866017
model.layers.14	8461	3.879199677230047e-05	0.03882296552860565
model.layers.15	12482	5.7227479460093894e-05	0.048060370310016726
model.layers.16	12913	5.920352846244132e-05	0.04840145615838384
model.layers.17	6333	2.9035541373239436e-05	0.03335983097381218
model.layers.18	16172	7.414539319248827e-05	0.05506548364459965
model.layers.19	11209	5.139102846244132e-05	0.04550325428150738
model.layers.20	12210	5.598041373239437e-05	0.04642204144427783
model.layers.21	16307	7.476434125586854e-05	0.053794102872389414
model.layers.22	25206	0.00011556448063380281	0.06777618138807913
model.layers.23	20717	9.498331132629108e-05	0.06078562272138425
model.layers.24	20821	9.546013057511737e-05	0.060969691454541675
model.layers.25	19114	8.763387617370892e-05	0.05890420682485283
model.layers.26	28796	0.00013202391431924883	0.07294291069404908
model.layers.27	48346	0.00022165676349765259	0.09456030028836097
model.layers.28	21246	9.740867077464789e-05	0.06154038937465008
model.layers.29	29347	0.00013455013937793426	0.07264650341002406
model.layers.30	64540	0.0002959030223004695	0.10916116792841717
model.layers.31	46044	0.0002111025528169014	0.4470326179022129
total	541942	7.7646748001027e-05	0.5433293524208188

表 23: $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$ 間の $M_{\mathrm{SFT}_{\mathrm{D2}}}$ を基準とした コンフリクト 数, コンフリクト率, Conflict Limited L2

モデル	コンフリクト数	コンフリクト率	Conflict Limited L2
model.layers.0	7809	3.580270686619718e-05	0.03533972441854576
model.layers.1	36533	0.0001674965155516432	0.11199111283328556
model.layers.2	14677	6.729111649061033e-05	0.04794489264152872
model.layers.3	15831	7.258197623239436e-05	0.04906133273734622
model.layers.4	28304	0.0001297681924882629	0.06749770831703665
model.layers.5	29930	0.0001372230780516432	0.06757404998664725
model.layers.6	26565	0.00012179522447183098	0.06659715923173312
model.layers.7	34050	0.0001561124559859155	0.07314935523144662
model.layers.8	31566	0.00014472381161971832	0.07037816689990004
model.layers.9	34367	0.0001575658377347418	0.07341248106382214
model.layers.10	37027	0.00016976140698356807	0.07718782988566589
model.layers.11	41869	0.00019196101085680752	0.08145447457646407
model.layers.12	50664	0.0002322843309859155	0.09060479300580626
model.layers.13	35286	0.00016177926936619718	0.07619534644387509
model.layers.14	36342	0.00016662081866197182	0.076073215855071
model.layers.15	54490	0.0002498257775821596 5	0.09435395568253561
model.layers.16	54706	0.0002508160944835680 6	0.09377542786669668
model.layers.17	26787	0.00012281305017605633	0.06521342963473184
model.layers.18	64880	0.0002974618544600939	0.10429120566286153
model.layers.19	45694	0.00020949787265258217	0.0873136443252231
model.layers.20	48544	0.00022256455399061032	0.08691337670902193
model.layers.21	72842	0.00033396603579812206	0.10761368706024432
model.layers.22	108230	0.0004962129548122066	0.13154672052347394
model.layers.23	90242	0.00041374156396713614	0.1195608763038629
model.layers.24	83733	0.0003838990977112676	0.11577260673683602
model.layers.25	83597	0.00038327556484741784	0.11716079407612678
model.layers.26	107380	0.0004923158744131455	0.13228164494719327
model.layers.27	198335	0.0009093264011150235	0.1804250424785017
model.layers.28	88908	0.00040762544014084505	0.1179808149420785
model.layers.29	115235	0.0005283294821009389	0.1353939722068931
model.layers.30	258835	0.0011867068295187794	0.20625707440814775
model.layers.31	160002	0.0007335772447183099	0.4638939137037548
total	2123260	0.00030421010765111504	0.733824178400172

表 24: 各モデルの生成結果と各データセットとの BLEU, BERTScore

モデル	BLEU(D_0)	BERTScore(D_0)	BLEU(D_1)	BERTScore(D_1)
M_{base}	0.00685	0.74843	0.00686	0.74532
$M_{\text{SFT}_{\text{D0}}}$	**0.10713**	**0.79208**	0.02214	0.76002
$M_{\text{SFT}_{\text{D1}}}$	0.01768	0.74124	**0.09295**	**0.78445**
$M_{\text{DPO}_{\text{D0}}}$	0.00340	0.75291	0.00045	0.74060
$M_{\text{DPO}_{\text{D1}}}$	0.00242	0.72455	0.01723	0.75312
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.04136	0.76400	0.01842	0.75454
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.03608	0.74590	0.07062	0.77633

表 25: 各モデルの生成結果とデータセットとの BLEU, BERTScore

モデル	データセットのクエリに対する生成結果 (202 件)	オリジナルのクエリに対する生成結果 (100 件)
M_{base}	6.62	7.08
$M_{\text{SFT}_{\text{D0}}}$	5.81	6.68
$M_{\text{SFT}_{\text{D1}}}$	**8.60**	8.71
$M_{\text{DPO}_{\text{D0}}}$	5.57	5.92
$M_{\text{DPO}_{\text{D1}}}$	8.27	**8.73**
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	5.37	6.18
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	8.23	8.38

表 26: ($M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, M_{base}), ($M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{SFT}_{\text{D1}}}$), ($M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{DPO}_{\text{D0}}}$) 間の L1 ノルム, L2 ノルム

モデル	L1 ノルム	L2 ノルム
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{SFT}_{\text{D1}}}$	188019.5	0.9321757445200005
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, M_{base}	937429.5	14.319758411266276
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{DPO}_{\text{D0}}}$	979227.5	15.046765478573377
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{SFT}_{\text{D0}}}$	1071540.0	16.608886252216386

表 27: 2 つのモデルの生成結果における BLEU, BERTScore

モデル	BLEU	BERTScore
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, M_{base}	0.02697	0.76962
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{DPO}_{\text{D0}}}$	0.01633	0.76247
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{SFT}_{\text{D0}}}$	0.082937	0.80572
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{SFT}_{\text{D1}}}$	0.08525	0.78747

表 28: $M_{\text{SFT}_{\text{D1}}}$, M_{base}), ($M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, M_{base}), ($M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{SFT}_{\text{D1}}}$) モデル間の L1 ノルム, L2 ノルム

モデル	L1 ノルム	L2 ノルム
$M_{\text{SFT}_{\text{D1}}}$, M_{base}	918794.5	14.002781183014145
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, M_{base}	948039.0	14.505088176492155
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$, $M_{\text{SFT}_{\text{D1}}}$	174754.5	0.7409783401721322

表 29: $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の M_{base} を基準とした コンフリクト 数, コンフリクト率, Conflict Limited L2

モデル	コンフリクト数	コンフリクト率	**Conflict Limited L2**
model.layers.0	34	1.5588321596244132e-07	0.0023544069240705456
model.layers.1	4486	2.0567414906103287e-05	0.03580303744851197
model.layers.2	0	0.0	0.0
model.layers.3	8	3.6678403755868543e-08	0.0009765625
model.layers.4	0	0.0	0.0
model.layers.5	2	9.169600938967136e-09	0.0005459150335692846
model.layers.6	0	0.0	0.0
model.layers.7	5	2.292400234741784e-08	0.0008457279333832408
model.layers.8	2	9.169600938967136e-09	0.0005459150335692846
model.layers.9	0	0.0	0.0
model.layers.10	2	9.169600938967136e-09	0.00048828125
model.layers.11	33	1.5129841549295775e-07	0.002237585788552656
model.layers.12	6	2.750880281690141e-08	0.0009134905729428568
model.layers.13	2	9.169600938967136e-09	0.00048828125
model.layers.14	7	3.2093603286384975e-08	0.0009765625
model.layers.15	10	4.584800469483568e-08	0.0011708573054962693
model.layers.16	3	1.3754401408450705e-08	0.000732421875
model.layers.17	0	0.0	0.0
model.layers.18	4	1.8339201877934272e-08	0.000732421875
model.layers.19	5	2.292400234741784e-08	0.0007720404443770458
model.layers.20	0	0.0	0.0
model.layers.21	1	4.584800469483568e-09	0.0004228639666916204
model.layers.22	9	4.126320422535211e-08	0.001118792894276328
model.layers.23	3	1.3754401408450705e-08	0.0005980199567341743
model.layers.24	5	2.292400234741784e-08	0.0007720404443770458
model.layers.25	5	2.292400234741784e-08	0.0008097228492078613
model.layers.26	5	2.292400234741784e-08	0.0007720404443770458
model.layers.27	108	4.951584507042254e-07	0.00390625
model.layers.28	0	0.0	0.0
model.layers.29	36	1.6505281690140845e-07	0.002142325289890655
model.layers.30	17	7.794160798122066e-08	0.0015049838874435977
model.layers.31	1567	7.184382335680751e-06	0.1323543950543304
total	6365	9.11945468383216e-07	0.13727130945427743