修 士 学 位 論 文

題　　目

大規模言語モデルにおけるユーザー嗜好学習方法の重みに基づくモデル変化解析

主査　　吉岡 理文 教授

副査　　森 直樹 教授

副査　　内海 ゆづ子 講師

令和　6　年（　2024　年）度修了

（No.　BGA23130　）　　西村昭賢

大阪公立大学大学院情報学研究科
基幹情報学専攻　知能情報学分野

目 次

図 目 次

表目次

1 はじめに

近年，Generative Pre-trained Transformers (GPT) [2] に代表される大規模言語モデル (Large Language Model, LLM) の急速な発展により，自然言語処理分野において高度なテキスト生成やテキスト理解が実現されている．LLM の応用先は多岐にわたっており，情報検索，文書要約，さらにはプログラミング支援など幅広い領域で活用が進展している．特に，人間と自然な対話を実現するチャットボットの可能性が大きく広がっており，従来の単なる情報伝達手段を超えて，対話を通じてユーザーに親しみや共感を抱かせる能力によりユーザー満足度の向上が期待される．また，ゲームや VR といったエンターテインメント分野においては，キャラクター性を持たせたチャットボットが魅力的かつ没入感のあるコミュニケーションを演出し，ユーザーの感情や体験に深い影響を及ぼす存在として期待されている．

キャラクター性を持たせたチャットボットの実現に向けて，LLM にユーザーの嗜好や属性を組み込みそれらを反映したテキスト生成を実現する研究が注目されている．具体的には，Supervised Fine Tuning (SFT) などの手法を用いてモデルを微調整し，応答スタイルを反映させるアプローチや，Direct Preference Optimization (DPO)[4] のようにユーザーの好みに基づく報酬設計する手法が代表例として挙げられる．これらの手法により，LLM は特定のスタイルの応答を実現するとともに，社会的・倫理的に好ましくない応答の生成を抑制し，安全かつ適切な応答の提供が可能となる．しかし，これらの手法を用いた場合に，モデル内部でどのような変化が生じるか，またその結果として生成出力や性能にどのような影響が及ぶかについては，定量的に十分解明されていない．

本研究では，日本語におけるロールプレイタスクにおいて，SFT および DPO を用いてモデルを調整した場合の学習過程やモデル内部の重み変化を定量的に比較・検討し，それらが生成出力や性能に及ぼす影響を明らかにすることを目的とする．具体的には, モデルマージで用いられる Ties-Merging[5] の概念を用いてモデルの比較をする. さらに，両手法の併用によって期待される性能向上や，ベースモデルへの回帰可能性についても考察する．本研究の成果は，LLM におけるユーザー嗜好学習の最適化およびその内部動作の理解を深めるための基盤的知見を提供することが期待される．

以下に本論文の構成を示す. まず, 2 章で本研究で用いる要素技術について概説する. 次に 3 章で関連研究を紹介し, 本研究の位置づけを明らかにする. 4 章で提案手法, 5 章で数値実験に関して説明し, 6 章で結果と考察を示す. 最後に 7 章でまとめと

今後の課題を述べる.

2 要素技術

2.1 Transformer

Transformer [1] は, Long Short-Term Memory (LSTM) [6],[7] や Gated Recurrent Unit (GRU) [8] に代表される Recurrent Neural Network (RNN) を用いずに, Attention 機構 [9],[10] を基本構造とする Encoder-Decoder モデルである. ここで, Encoder が入力系列 $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$ を連続系列 $\boldsymbol{z} = (\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_n})$ へと写像し, Decoder が $\boldsymbol{z}$ から出力系列 $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_n})$ を生成する場合を考える.

図 2.1 に Transformer の概略を示す. Encoder は左側に, Decoder は右側にそれぞれ位置している. N は層数を表しており, $N = 6$ に設定されている.

Transformer は RNN のように再帰構造を持たないため, 入力系列の位置情報を Positional Encoding や Positional Embedding で考慮する. 前者は各位置に対して要素が固定のベクトルを加算する.

Positional Encoding の行列を PE とすると,

$$\begin{aligned} \mathrm{PE}(\mathrm{pos}, 2i) &= \sin\left(\frac{\mathrm{pos}}{p_{\mathrm{freq}}^{2i/d_{\mathrm{model}}}}\right) \\ \mathrm{PE}(\mathrm{pos}, 2i+1) &= \cos\left(\frac{\mathrm{pos}}{p_{\mathrm{freq}}^{2i/d_{\mathrm{model}}}}\right) \end{aligned} \tag{2.1}$$

となる. ただし, d_{model} は入力 Embedding の次元数, pos, i は Positional Encoding の位置および成分である.

一方で, Positional Embedding は各位置に対して要素が学習により可変なベクトルを加算する. 初期値は 0 や乱数などが用いられる.

Encoder の各層は 2 層のサブレイヤおよび Layer Normalization [11] と残差接続 [12] を持つ. 1 層目は Multi-Head Attention であり, 2 層目は単純な Position-wise Feed-Forward Network (FFN) である. FFN は, ReLU 関数を間に有する 2 層の全結合層で構成され, (2.2) 式のように表される.

$$\mathrm{FFN}(\boldsymbol{x}) = \max(\boldsymbol{0}, \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_2)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{2.2}$$

ただし, $\boldsymbol{W}_i, \boldsymbol{b}_i (i \in \{1, 2\})$ はそれぞれ全結合層の重みとバイアスである.

Decoder では, Encoder の 2 層のサブレイヤに加えて, Encoder の出力に対して Multi-Head Attention を計算するための 3 層目のサブレイヤが挿入される.

Attention 機構は, Query, Key および Value への写像として表現される. ここで, Query, Key および Value はそれぞれベクトルである. 出力は Value の加重合計により

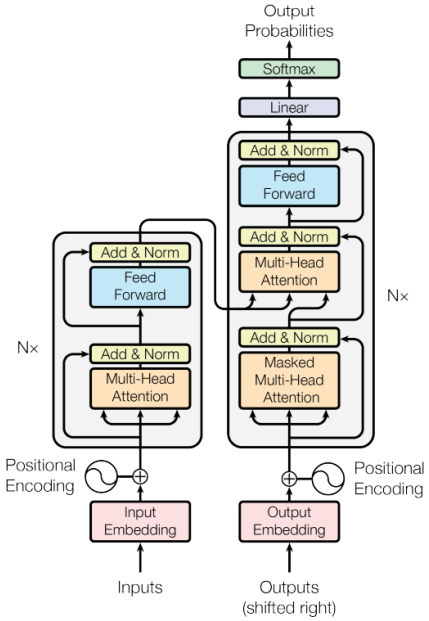

図 2.1: Transformer (文献 [1] Figure 1. 参照) のアーキテクチャ

求められる. なお, 各 Value に対する重みは Query と対応する Key から算出される値である.

2.2 Generative Pre-trained Transformers (GPT)

Transformer を用いたモデルとして, Bidirectional Encoder Representations from Transformers (BERT) [13], コンピュータビジョンの分野では Vision Transformer (ViT) [14] に代表される Encoder のみを用いたモデル, Text-to-Text Transfer Transformer (T5)[15] に代表される双方を使用した Encoder-Decoder モデルが存在する. そして Transformer の Decoder のみを使用した代表的なモデルには Generative Pre-trained Transformers (GPT) [2] が挙げられる.

図 2.2 に GPT のアーキテクチャを示す. GPT は Decoder only のモデルを用いているため, 図 2.1 の右側の Decoder に存在する Multi Head Attention 層 は無くなっている. このモデルにおいてトークン系列 $U = (u_{-k}, \ldots, u_{-1})$ からトークン u を予測する場合を考える. このとき, 入力テキストの埋め込み表現 W_e, 位置埋め込み表現 W_p を用いて Embedding 層の計算は (2.3) 式のように表される.

$$h_0 = UW_e + W_p \tag{2.3}$$

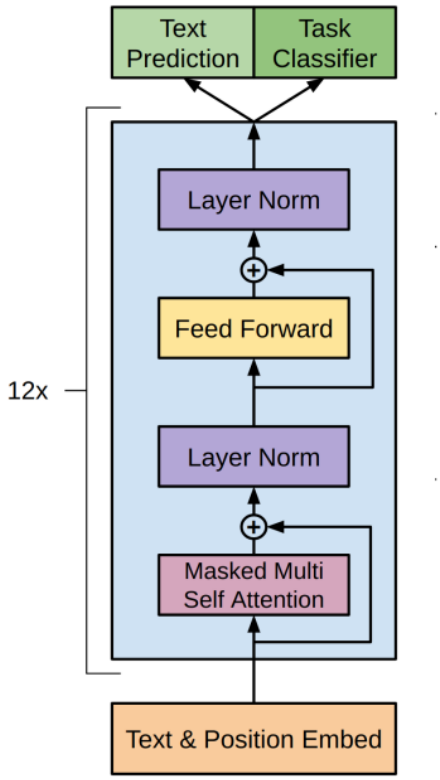

図 2.2: GPT (文献 [2] Figure 1. 参照) のアーキテクチャ

そして, Masked Multi Self Attention 層, Feed Forward 層, Layer Norm 層で構成される構造を transformer_block と置くと, transformer_block のレイヤー数を n として, トークン u の予測確率は (2.5) 式のように表される.

$$h_l = \text{transfomer\_block}(h_{l-1}), \forall l \in [1, n] \tag{2.4}$$

$$P(u) = \text{softmax}(h_n W_e^T) \tag{2.5}$$

GPT では $n = 12$ としており, このモデルをラベルなしデータでの事前学習と少量のラベル付きデータによるファインチューニングにより応用可能性が広いモデルを実現している.

GPT は近年急速に進化を続けており, 2020 年には OpenAI から 1750 億パラメータという大規模なパラメータをもつ GPT-3[16], 2023 年には GPT-3 の改良版である GPT-3.5, GPT-3.5 をさらに複雑な推論を可能にした GPT-4[17], 2024 年には GPT-4 を超える推論性能を残す OpenAI o1 が公開されている. GPT-4 以降のモデルはテキストだけでなく画像などマルチモーダルな入力が可能となっており注目を集めている

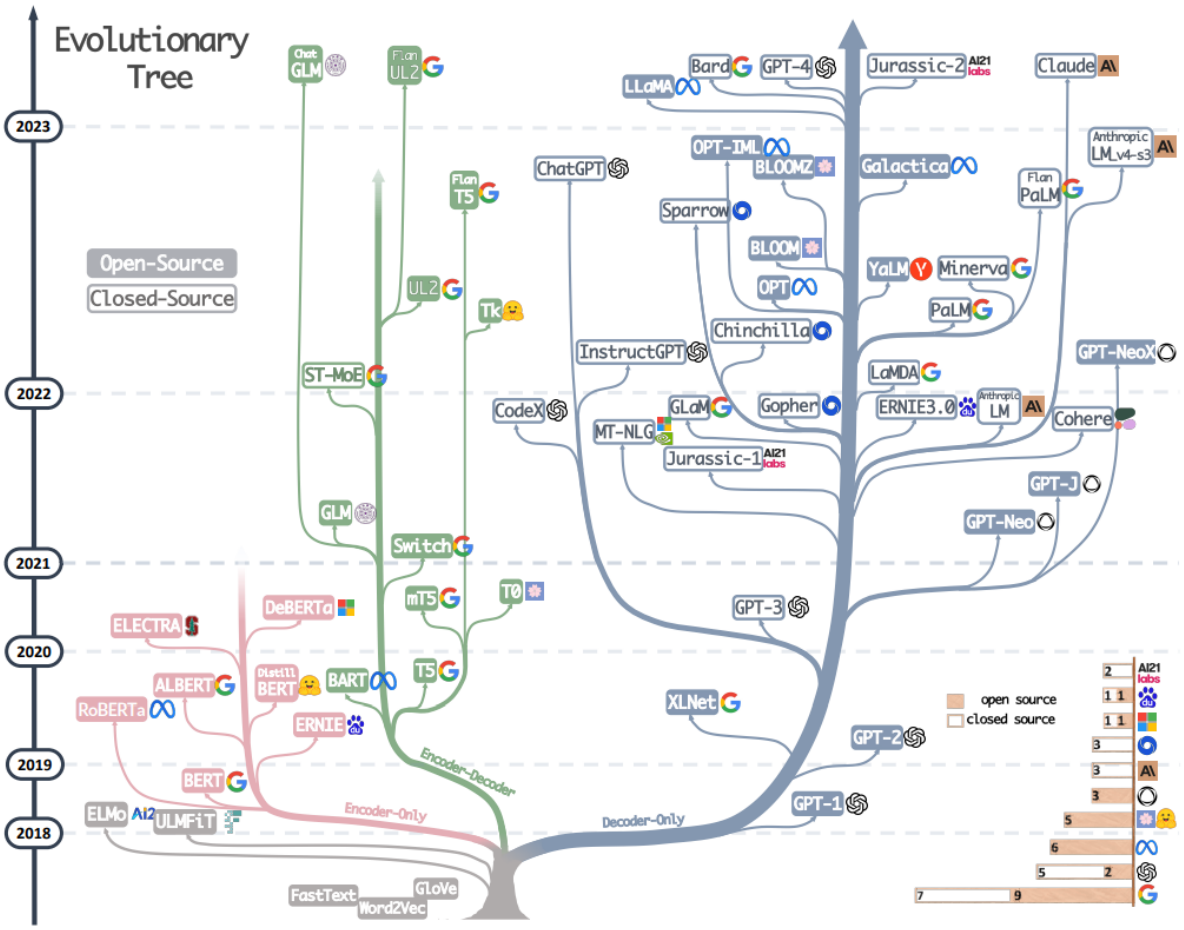

図 2.3: Transformer を用いた言語モデルの発展 (文献 [3] Figure 1. 参照)

が, 安全性や競争のリスクから具体的なパラメータ数や学習方法, 学習データセットなどは公開されていない.

2.3　Llama

図 2.3 に Transformer を用いた言語モデルの発展を示す. 近年の LLM はほとんどが Decoder only モデルとなっている.

Meta によって公開された Llama [18] も Decoder only のモデルの 1 つであり, 一般に利用可能なデータセットのみを用いて学習され, ほとんどのベンチマークで Llama2 を上回る精度を達成した. 2023 年に Llama を改良した Llama2[19] が登場し, 2024 年には更に複雑な推論を可能にした Llama3 [20] が登場した.Llama3 を改良し表現力や処理速度が向上した Llama3.1 は Llama3.1 は, 8B, 70B, 405B の 3 つのモデルサイ

ズが用意されており, 57 の多様なタスクで構成される言語理解ベンチマークである MMLU (Massive Multitask Language Understanding)[21], 与えられた指示に基づいて Python コードを生成する能力を評価するベンチマークである HumanEval[22], 小学レベルの数学の文章問題を解く能力を評価するベンチマークである GSM-8K[23] において, Llama3 405B は GPT-4 を超える性能を残している.

2.4 Qwen

Qwen[24] は Alibaba が公開している大規模言語モデルであり, 2023 年に公開された Qwen-72B は MMLU, HumanEval, GSM-8K などのベンチマークにおいて Llama2-70B と比較して高い数値を記録した.

2024 年に公開された Qwen-2.5 [25] は 0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B などの幅広いパラメータ数のモデルが公開されており, それに加えてコード生成に特化した Qwen2.5-Coder[26], 数学向けに特化した Qwen2.5-Math[27] といった特定のタスクに特化したバージョンも公開されている.

2.5 ユーザー嗜好の LLM 出力の制御

LLM にはプロンプトと呼ばれる実行指示を記述した文を入力として与えることで, 与えた指示に従ったユーザーの求める応答を生成する. より高性能かつユーザーの要望に沿った応答を実現するための最も簡単なアプローチとしてプロンプトエンジニアリングが挙げられる. プロンプトエンジニアリングとは LLM に与えるプロンプトを最適化する手法である. 特定のタスクに取り組む際にいくつかの (入力, 期待される出力) の例を与える Few-shot Prompting, 中間的な推論を例として与えることで複雑な推論を可能にする Chain of Thoughts (CoT) [28] など, プロンプトエンジニアリングは多くの手法が考案されている [29]. また, 遺伝的アルゴリズムを用いてプロンプトエンジニアリングを自動化する PromptBreeder[30] といった研究もなされている. このようなプロンプトエンジニアリングは, LLM 内部の重みを変えずにユーザーが求める応答を得ることができるという点で OpenAI API に代表される API 経由でのみアクセスすることができる LLM を活用する際に有効なアプローチとなる.

さらにモデル内部の重みを変更可能なローカル LLM を活用できる場合には, Supervised Fine-Tuning (SFT) や Direct Preference Optimization (DPO) のような手法で

LLM をファインチューニングし LLM 内部の重みまで変えることでユーザーが望む応答が実現しやすくなることが期待できる.

2.5.1 Supervised Fine-Tuning (SFT)

Supervised Fine-Tuning (SFT) はベースとなる LLM に対して, 入力と人間が「好ましい」と考える模範解答のペアを用意しこれを直接の教師データとしてモデルを教師ありファインチューニングする手法である. SFT の代表的な例として instruction tuning が挙げられる. 膨大な量のテキストデータで事前学習された LLM はそのままでは入力となる文の続きを生成する振る舞いとなり, ユーザーの指示を踏まえた対話的な形でのテキスト生成はできない. そこで Alpaca Dataset[1] に代表される大規模な instruction tuning 用のデータセットで instruction tuning することで, より広範な指示に対応できる LLM を生み出すことができる.

2.5.2 Reinforcement Learning from Human Feedback (RLHF)

プロンプトに対して人間が「好ましい」と考える応答, また「好ましくない」と考える応答のペアを学習データセットとして, 好ましい応答の生成確率を高め, 好ましくない応答の生成確率を抑制する形で LLM を学習する手法を preference training と呼ぶ. preference training の代表的な手法として Reinforcement Learning from Human Feedback (RLHF)[31], Direct Preference Optimization (DPO) がある.

RLHF は以下の 3 プロセスから構成される.

1. データ収集と言語モデルの事前学習
2. 報酬モデルの学習
3. PPO[32] を用いた強化学習による言語モデルのファインチューニング

データ収集と言語モデルの事前学習ではラベル付け担当者がデータセットからランダムに抽出されたサンプルに対して好ましい応答を提示しそのデータで LLM を SFT する.

報酬モデルの学習では, 同じプロンプトを条件とする 2 つの言語モデルから生成されたテキストをラベル付け担当者が比較し集まった比較データでユーザー嗜好をスカラー値で返す報酬モデルを学習する.

[1] https://huggingface.co/datasets/tatsu-lab/alpaca

最後に強化学習アルゴリズムである PPO で事前学習済みのモデルをファインチューニングする. 具体的にはプロンプトと回答が与えられると, 報酬モデルによって報酬を生成する. 次に報酬モデルの過剰最適化を緩和するために, 事前学習済みモデルと強化学習モデルのトークンごとの KL divergence[33] を求める. 報酬と KL divergence を基にモデルの重みを更新することで強化学習モデルが事前学習済みから大きく離れることを防ぎつつ高い報酬を獲得するように学習を進行させることができる.

RLHF は事前学習済みモデル, 報酬モデル, 強化学習モデルの 3 つを同時に扱うため, 大量の計算資源を必要とする, 他のモデルの出力を基に学習をする必要があるため並列化が難しく学習に時間がかかるといった問題がある.

2.5.3 Direct Preference Optimization (DPO)

DPO は RLHF における報酬モデルを用いずに直接最適化を可能とした手法であり, (2.6) 式と表される損失関数を最適化する手法である. ここで, D_{pair} はデータセット, x はプロンプト, y^w はプロンプト x に対する好ましい出力, y^l はプロンプト x に対する好ましくない出力であり, π_θ は学習対象の LLM, π_{ref} は事前学習済みのモデルである.

$$L_{\text{DPO}} = -\mathbb{E}_{(x,y^w,y^l)\sim D_{\text{pair}}}\left[\log \sigma(\beta \log \frac{\pi_\theta(y^w \mid x)}{\pi_{\text{ref}}(y^w \mid x)} - \beta \log \frac{\pi_\theta(y^l \mid x)}{\pi_{\text{ref}}(y^l \mid x)})\right] \tag{2.6}$$

DPO は数学的に RLHF と等価であることが知られており, 単純な勾配法でモデルを直接最適化する手法であるため上記で述べた RLHF の欠点を改善した手法といえる.

2.6 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) [34] とは, 学習するパラメータ数を削減しつつ fine-tuning する手法である. モデルの線形層のパラメータを $D_{\text{in}} \times D_{\text{out}}$ 次元の行列 $\boldsymbol{W}$ とし, 入力ベクトルを $\boldsymbol{x}$ とすると, 出力 $\boldsymbol{h}$ は (1) 式で表される.

$$\boldsymbol{h} = \boldsymbol{W}\boldsymbol{x} \tag{2.7}$$

LoRA では線形層のパラメータ $\boldsymbol{W}$ と同次元の差分行列 $\Delta\boldsymbol{W}$ を用意し, 出力 $\boldsymbol{h}$ は (2) 式で表される. 学習の際には $\boldsymbol{W}$ を固定し, $\Delta\boldsymbol{W}$ のみを学習する.

$$\boldsymbol{h} = \boldsymbol{W}\boldsymbol{x} + \Delta\boldsymbol{W}\boldsymbol{x} \tag{2.8}$$

この時, ランク r を設定し, $D_{\text{in}} \times r$ 次元の行列 $\boldsymbol{A}$, $r \times D_{\text{out}}$ 次元の行列 $\boldsymbol{B}$ で, $\Delta\boldsymbol{W}$ は (3) 式で表せる.

$$\Delta\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B} \tag{2.9}$$

$\boldsymbol{W}$ のパラメータ数は $D_{\text{in}} \times D_{\text{out}}$ となる一方で $\Delta\boldsymbol{W}$ のパラメータ数は $r\times(D_{\text{in}} + D_{\text{out}})$ となり, 一般的に r は D_{in}, D_{out} に比べて非常に小さい値であるため, 学習パラメータ数を大きく減らすことができる.

2.7 量子化

大規模なニューラルネットワークや LLM などの膨大なパラメータを持つモデルでは, 演算の際に膨大な数の乗加算を必要とするため演算に時間がかかる. また, 多くのパラメータを保持するために多くのメモリが必要となる. これらの問題を軽減するためのアプローチとして量子化がある. 一般的に LLM の学習や推論では 16 ビット浮動小数点 (FP16) が用いられることが多いが, これを量子化して 4 ビットに変換する際には FP16 で表現されているパラメータを 2^4通りの値いずれかにマッチングする. 量子化により精度は減少するものの消費するメモリ量を大幅に軽減することができる.

本研究で用いた Quantized Low-Rank Adaptation (QLoRA) [35] では LLM のパラメータの多くが正規分布に従うことを利用した NormalFloat4 という量子化手法により 4 ビット量子化した LLM において効率的に fine-tuning できることを示している.

2.8 Ties-Merging

複数のファインチューニング済みのモデルを組み合わせて単一のモデルにまとめることで, 様々なタスクに対応する汎用的なモデルを構築する方法がモデルマージである. Ties-Merging[5] はモデルマージにおけるマージ手法の 1 つであり, 事前学習モデルと各タスクに対してファインチューニングしたモデルとの差分, すなわちタスクベクトルを用いることで異なるタスクにおける知識を統合し複数タスクに汎用的に対応できるモデルマージを実現する.

ここで, あるベースモデル M_{base} を特定のタスク T に対してファインチューニングして得られたモデル M_{T} について, M_{base} のパラメータを θ_{base}, M_{T} のパラメータを θ_{T} とするとタスクベクトル Δ_{T} は (2.10) 式と表される.

$$\Delta_{\text{T}} = \theta_{\text{T}} - \theta_{\text{base}} \tag{2.10}$$

Ties-Merging を用いた場合, 以下の 3 ステップでモデルマージをする.

1. パラメータの変化量が小さい値を切り捨て

全てのパラメータを含んだタスクベクトルの単純な平均を取ると, あるモデルにおける重要な変化が他のモデルの冗長な変化により相殺される問題がある. 論文中では各タスクベクトルにおいて変化量上位 20 % のみ保持し, 残り 80 % は変化量を 0 とし冗長なパラメータによる干渉を防ぐ.

2. パラメータの符号の一致

同一パラメータに置いてタスクベクトル間で変化の方向が異なることがある. タスクベクトルの単純な平均を取ると正負の変化が相殺され両タスクについて性能が低下する問題がある (符号のコンフリクト). そのため, 各パラメータについてタスクベクトルの対応するパラメータの支配的な符号を決定する.

3. 符号が一致するパラメータのみを平均化して統合

タスクベクトルの各パラメータに対して, 支配的な符号と一致する成分のみ平均を計算しマージに用いるタスクベクトルの成分とする.

2.9 テキスト評価指標

テキスト生成タスクにおける評価指標として, モデルが生成した文と参照文の類似度を測定することが一般的となっている. 文同士の類似度を定量化した指標として BiLingual Evaluation Understudy (BLEU)[36], BERTScore[37] がある.

2.9.1 BLEU

BLEU スコアは機械翻訳の品質を評価するために広く使用される指標であり, 生成文と参照文とをコーパス単位で比較して計算される. BLEU スコアは次のように計算される.

$$\mathrm{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.11}$$

ここで, p_n は n グラムの精度を表し, w_n は各 n グラム精度に割り当てられた重みである. Brevity Penalty (BP) は生成された文が参照文の内容を完全にカバーしていない場合に調整をする.

$$BP = \min\left(1, \exp\left(1 - \frac{\text{参照文の総単語数}}{\text{生成文の総単語数}}\right)\right) \tag{2.12}$$

n グラム精度 p_n は, 生成された n グラムの総数に対する一致した n グラムの割合として計算される.

$$p_n = \frac{\sum_{i=1}^{M} m_n}{\sum_{i=1}^{M} \text{生成文における n グラム数}} \tag{2.13}$$

ここで, m_n は生成されたテキストの n グラムと参照文との一致数を示す.

$$m_n = \sum_{i=1}^{M} \text{生成文と参照文で一致した n グラム数} \tag{2.14}$$

BLEU スコアは 0 から 1 の範囲で評価される. ただし, このスコアは n グラムの精度に大きく依存しており, 生成文と参照文との間で多くの一致する n グラムがあれば BLEU スコアは高くなる. BLEU スコアは文同士の類似度を測る指標として有用であるが, より詳細な評価には人間の評価と組み合わせて使用することが重要である.

2.9.2 BERTScore

BERTScore は, 事前学習された BERT モデルを使用し文の意味的類似度を測定するために考案された指標である. BLEU に代表される n グラムベースの精度評価方法とは異なり, BERTScore は文脈を考慮した埋め込みベースでの類似度を計算するためテキスト中の語が完全に一致する必要がなく, より人間に近い翻訳評価を提供する.

BERTScore の計算には, まず各単語の埋め込みベクトルが BERT モデルを用いて抽出される. ここで, トークン長 N の参照文, トークン長 M の生成文をそれぞれ BERT に入力し, 生成文のトークン埋め込み列 $\mathbf{x}_1, ..., \mathbf{x}_N$ と, 参照文のトークン埋め込み列 $\mathbf{y}_1, ..., \mathbf{y}_M$ を得た後にコサイン類似度を用いてトークン埋め込み列間の類似度を算出し, 以下の式で適合率, 再現率, F 値を計算する.

$$\text{適合率} = \frac{1}{M} \sum_{j=1}^{M} \max_{1 \le i \le N} \frac{\mathbf{x}_i^\top \mathbf{y}_j}{\|\mathbf{x}_i\| \|\mathbf{y}_j\|} \tag{2.15}$$

$$\text{再現率} = \frac{1}{N} \sum_{i=1}^{N} \max_{1 \le j \le M} \frac{\mathbf{x}_i^\top \mathbf{y}_j}{\|\mathbf{x}_i\| \|\mathbf{y}_j\|} \tag{2.16}$$

$$\text{F 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}} \tag{2.17}$$

2.9.3 LLM による自動評価 (LLM-as-a-judge)

従来の NLP タスクの学習では, 評価指標に基づく高評価が得られても人間の評価から見た品質の高さを保証するものではない点が問題点として挙げられる. たとえば

BLEU, ROUGE[38] のようなテキスト要約用の評価指標は忠実性や事実性との相関が低いことが報告されており, ROUGE の値が高い場合でも Hallucination (幻覚) が生成されている [39]. また, 正解が複数ある場合に対応できないため本研究のような LLM のロールプレイの出力などの評価は BLEU や ROUGE 単体だけでは難しい.

そのような背景の中, LLM の発展により GPT-4 などの高性能の LLM を評価者として利用する LLM-as-a-judge という手法が登場した. LLM-as-a-judge は従来の評価指標よりも柔軟で精度の高い評価を可能にしており, 対話型 AI, 情報検索といった分野で有望な応用が期待できる手法である [40]. LLM-as-a-judge によって GPT-4 を評価者 LLM として他の LLM の性能を評価する Japanese Vicuna QA Benchmark といったベンチマークも存在する [41].

しかし, LLM-as-a-judge には以下の課題もあり, テキストの評価には複数の評価指標と人から見た定性的な評価を組み合わせて評価する必要がある.

- LLM の評価は完全に公平ではなく, 事前学習データなどから発生するバイアスを含む可能性がある
- 評価者 LLM による評価の一貫性が保証されていない
- 評価者の性能に大きく結果が依存される

2.10 OjousamaTalkScriptDataset

OjousamaTalkScriptDataset [2] は一般的な問いかけとお嬢様スタイルの応答がペアとなっている会話データを 202 件収録している MIT ライセンスで公開されたデータセットである. また, OjousamaTalkScriptDataset におけるお嬢様のキャラクターは, 以下の設定が与えられている.

- 高校生
- 女性
- 外見の設定はあえてしていません
- ミュージカルが好きでミュージカル女優に憧れていた
- 両親は不動産業
- 兄はアメリカに留学中
- バイオリンを子供の頃から習っている

[2] https://github.com/matsuvr/OjousamaTalkScriptDataset

- 一時期、祖父の住む長野で暮らしていた
- 現在は東京在住

3 関連研究

本章では LLM による特定のキャラクターのロールプレイに関する研究, また Transformer の解析に関する研究を紹介し, 本研究の位置づけを明確にする.

3.1 LLM によるキャラクターロールプレイ

ロールプレイの概念は, 単なる AI アシスタントの役割を超えて人々が求める心理的・娯楽的なニーズを満たすために発展してきた.

Character-LLM: A Trainable Agent for Role-Playing[42] では, ルートヴィヒ・ヴァン・ベートーヴェンなどの歴史上の人物の精神的な活動や物理的な行動を模倣し再構築した経験を用いて LLM に SFT を適用することでキャラクターを演じる LLM, Character-LLM を提案した. 研究では, ChatGPT によって生成された質問に対する応答を用いてその応答データに対して GPT-3.5 を用いて, Memorization, Values, Personality, Hallucination, Stability の 5 段階で評価した.

ChatHaruhi: Reviving Anime Character in Reality via Large Language Model[43] では, ユーザーのクエリに対してシステムプロンプト, キャラクターの記憶, 対話履歴などを組み合わせることで LLM のロールプレイの性能を向上させるフレームワークを提案している. また, アニメやドラマなどに登場する 32 人のキャラクターについて実際の脚本に登場した会話例に加えて SFT を適用したモデルによって対話データを拡張し, 合計約 54000 個の会話データからなるデータセットを構築した.

RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models[44] では, 100 個のロールについて GPT を用いてロール固有の QA ペアを生成し, 各ロールに対して 400 以上の質問データセットを作成した. また, システムプロンプトを組み込んだローカル LLM のファインチューニングにより従来のオープンソース LLM と比較してロールプレイ能力が大幅に向上し, かつ新しいロールへの適応能力も高い LLM を学習する方法を提案した. また, LLM のロールプレイ能力を測る RoleBench[3] といった新しいベンチマークを提案した.

3.2 Transformer 内部の定量的な解析

What do you learn from context? Probing for sentence structure in contextualized word representations [45] では, CoVe[46], ELMo[47], BERT, GPT の 4 つの異なる文埋め込みモ

[3] https://huggingface.co/datasets/ZenMoore/RoleBench

デルに 9 種類の NLP タスクを適用して 4 種類のモデルの性能を解析し, 言語表現をどのように内部に保持しているかを調査した. NLP タスクは以下の通り. なお, SPR においてはデータセットが 2 種類存在し論文中では分けて 2 タスクとして扱っている.

- Part-of-speech tagging (POS)
  各単語に対して品詞を割り当てるタスク
- Constituent labeling
  構造木の中で句や節といった構成素にラベルを割り当てるタスク
- Dependency labeling
  文中の係り受け関係などの依存関係とそのラベルを予測するタスク
- Named entity labeling
  人名や地名といった固有表現にラベルを割り当てるタスク
- Semantic role labeling (SRL)
  動詞や述語を中心に，文中の各項のどんな意味的役割を割り当てるタスク
- Coreference
  文中で同じ実体を指す表現を結びつけるタスク
- Semantic proto-role (SPR)
  SRL と比較して, より微細な概念的・意味的特徴をアノテートしているデータセット. 動詞と引数のペアについてその引数がどんな意味的性質を持つかを複数ラベルで判定するタスク
- Relation Classification (Rel.)
  2 つのエンティティに対して，それらの関係を割り当てるタスク

このような単語情報から意味関係まで幅広い言語表現のタスクにより文埋め込みモデルがどの層でどの言語情報を埋め込んでいるのかを分析することができ, この手法を edge probing と呼ぶ.

edge probing による解析の結果, BERT が最も優れたスコアを残し, 従来の単語埋め込みと比べ文埋め込みモデルは依存構造や句構造など統語的タスクに置いては顕著に改善が見られたが, より高次の意味的タスクでは限定的な改善しか見られないといった結果を得た.

BERT Rediscovers the Classical NLP Pipeline[48] では edge probing を用いて BERT の各層が文法や意味情報をどのように処理しているか解析した. 結果として, BERT において POS や構文解析などの文法情報は 1 から 7 層の浅い層で処理され, 意味役割付与 (SRL) や Coreference のような高レベルの位置情報は 9 から 20 層の深い層で

処理される傾向があった.

3.3 本研究の位置づけ

Character-LLM, ChatHaruhi, RoleLLM においてファインチューニングした LLM の性能評価では生成結果を ROUGE-L, LLM-as-a-judge といった評価指標による評価 人手評価, 定性的な評価がなされている. ロールプレイのタスクにおいて SFT や DPO といったファインチューニングでユーザー嗜好学習をした LLM の重みの変動に基づいた定量的な考察は十分になされていない.

また, Transformer の重みに関しても 先行研究では BERT という分類モデルでの考察にとどまっており, LLM のような decoder only の生成モデルに関しては十分に考察されていない.

そのため, 本研究では日本語でのデータセットを用いたロールプレイタスクにおいて SFT, DPO などの手法を用いて手法とデータセットがそれぞれ LLM 内部にどのような重みの変動をもたらすか調査し両手法, データセットによる違いを定量的に解析する.

4 提案手法

4.1 データセット

本研究では OjousamaTalkScriptDataset を基に ChatGPT o1 pro を用いて新たに一般的な応答を 202 件収録したデータセット, 男子大学生の応答を 202 件収録したデータセットを新たに構築した.

OjousamaTalkScriptDataset を参考に男子大学生のキャラクターには以下の設定を与えている.

- 大学生
- 男性
- 外見の設定はあえてしていません
- 一人称は俺で関西弁
- ぶっきらぼうな物言いだが、内面は熱い情熱と誠実さを秘めている
- 実家は滋賀の大津市にあり、両親は共働きのサラリーマン
- 小さい頃から病弱な二歳年下の妹の面倒を見てきた
- 大阪の公立大学に進学し、公務員となるため勉学に励んでいる

データセット構築のため, ChatGPT o1 pro に与えたプロンプトを以下に示す.

一般的な応答を収録したデータセットを構築するプロンプト

以下はお嬢様の応答のデータセットである.
お嬢様の comletion 以下の応答部分をすべて同じ意味でありながら可能な限り没個性的な応答に書き換えよ。
出力は同じ prompt 順の同じ json 形式とし、応答内の文以外は変えないこと。
–ここから–
{ OjousamaTalkScriptDataset の jsonl 形式のデータセット }

男子大学生の応答を収録したデータセットを構築するプロンプト

以下はお嬢様の応答のデータセットである。
お嬢様の comletion 以下の応答部分をすべて同じ意味でありながら以下に定義された男子大学生の応答に書き換えよ。
出力は同じ prompt 順の同じ json 形式とし、応答内の文以外は変えないこと。
変換する男子大校生の定義：
- 大学生 - 男性 - 外見の設定はあえてしていません - 一人称は俺で関西弁 - ぶっきらぼうな物言いだが、内面は熱い情熱と誠実さを秘めている - 実家は滋賀の大津市にあり、両親は共働きのサラリーマン - 小さい頃から病弱な二歳年下の妹の面倒を見てきた - 大阪の公立大学に進学し、公務員となるため勉学に励んでいる
生成時の注意点：- このペルソナは参考としているお嬢様とは完全に無関係である - このペルソナに合致しない生成はしないこと． 以下の内容に関連する生成は今回のペルソナでは不可: なルール - 兄がいる - 高級志向で高級という言葉を多用する - 海外旅行の経験がある - ミュージカルやバイオリンに興味があり話題にする - 女性的に内容 - 知らないこと答えられないことはわからないと答えて良い - データ数の追加欠損は付加 - 生成したデータが上記の不可なルールに抵触しないか厳密にチェック．男子大学生のペルソナに合致しない場合は合致するまで生成をやり直すこと
–ここからデータ–
{ OjousamaTalkScriptDataset の jsonl 形式のデータセット }

表 4.1 に「夕日に向かって走ろう」といったクエリに対する各データセットの応答を示す.

表 4.1: データセットの例

応答スタイル	「夕日に向かって走ろう」のクエリに対する応答
お嬢様	とてもロマンチックな気分になりそうですわね
一般的	ロマンチックな気分が味わえそうです
男子大学生	めっちゃロマンチックな気分になりそうやな

4.2 Conflict Limited L2 ノルム

本手法は Ties-Merging[5] から着想を得ている. Ties-Merging におけるタスクベクトルの符号のコンフリクトに関してはどのタスク同士が強く対立しやすいか, あるいはどのパラメータ領域で相反的に学習が進むかを示していると解釈できる.

本研究では, この考えのもとデータセット間や学習手法間の差異を定量的に確認するため符号のコンフリクトが生じている部分に限定して L2 ノルムを計算する手法 Conflict Limited L2 ノルムを提案する。具体的には ベースモデル M_{base} を特定のタスク $T1, T2$ に対してファインチューニングして得られたモデルのパラメータ $\theta_{\text{T1}}, \theta_{\text{T2}}$ においてそれぞれのベースモデルからのタスクベクトル $\Delta_{\text{T1}}, \Delta_{\text{T2}}$ を計算する. これらのタスクベクトルにおいて, 符号の衝突が生じているパラメータ領域のみを対象としその部分の L2 ノルムを計算する. なおコンフリクト数を C とすると (4.1) 式と表される.

$$\text{Conflict Limited } \|\Delta W\|_2 = \sqrt{\sum_{i \in C} \left(\Delta^i_{\text{T1}} - \Delta^i_{\text{T2}}\right)^2} = \sqrt{\sum_{i \in C} \left(\theta^i_{\text{T1}} - \theta^i_{\text{T2}}\right)^2} \tag{4.1}$$

5 数値実験

本研究で用いたデータセットを以下のように記載する.

- D_0 : 一般的な応答を 202 件収録したデータセット
- D_1 : OjousamaTalkScriptDataset (お嬢様スタイルの応答を 202 件収録したデータセット)
- D'_1 : 男子大学生スタイルの応答を 202 件収録したデータセット
- D_2 : D_0, D_1 を半数ずつ用いて計 202 件収録したデータセット
- D'_2 : D_0, D'_1 を半数ずつ用いて計 202 件収録したデータセット

また, 本研究で使用したベースモデルは以下の 2 つ. 図 5.1, 5.2 に 2 つのモデルのアーキテクチャを示す. 大きな違いとして transformer_block の層数があり, elyza/Llama-3-ELYZA-JP-8B が 32 層, Qwen/Qwen2.5-7B-Instruct は 28 層となっている.

- elyza/Llama-3-ELYZA-JP-8B[4]

  ELYZA 社が公開しているモデル. Llama3 に instruction tuning を適用した meta-llama/Meta-Llama-3-8B-Instruct に対してさらに大規模な日本語コーパスを用いて追加事前学習, instruction tuning をして日本語に対する指示遂行能力を強化した.

- Qwen/Qwen2.5-7B-Instruct[5]

  Qwne2.5 に大規模なデータセットで instruction tuning をして対話性能を強化したモデル. 中国語, 英語をはじめとしてフランス語やスペイン語, 日本語など 29 以上の言語に対応している.

5.1 実験 1

ベースモデル M_{base} として, elyza/Llama-3-ELYZA-JP-8B を用いる. ベースモデルに対して, 以下のモデルをファインチューニングする. ファインチューニングするにあたって各データセットは事前に train 162 件, test 40 件となるよう分割されている.

[4]https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
[5]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

```
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 5.1: elyza/Llama-3-ELYZA-JP-8B のアーキテクチャ

```
Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(152064, 3584)
    (layers): ModuleList(
      (0-27): 28 x Qwen2DecoderLayer(
        (self_attn): Qwen2SdpaAttention(
          (q_proj): Linear(in_features=3584, out_features=3584, bias=True)
          (k_proj): Linear(in_features=3584, out_features=512, bias=True)
          (v_proj): Linear(in_features=3584, out_features=512, bias=True)
          (o_proj): Linear(in_features=3584, out_features=3584, bias=False)
          (rotary_emb): Qwen2RotaryEmbedding()
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=3584, out_features=18944, bias=False)
          (up_proj): Linear(in_features=3584, out_features=18944, bias=False)
          (down_proj): Linear(in_features=18944, out_features=3584, bias=False)
          (act_fn): SiLU()
        )
        (input_layernorm): Qwen2RMSNorm((3584,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((3584,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((3584,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=3584, out_features=152064, bias=False)
)
```

図 5.2: Qwen/Qwen2.5-7B-Instruct のアーキテクチャ

- $M_{\mathrm{SFT_{D0}}}$: M_{base} に D_0 を用いて SFT を適用
- $M_{\mathrm{SFT_{D1}}}$: M_{base} に D_1 を用いて SFT を適用
- $M_{\mathrm{SFT_{D2}}}$: M_{base} に D_2 を用いて SFT を適用
- $M_{\mathrm{DPO_{D0}}}$: M_{base} に D_0 を chosen, D_1 を rejected として DPO を適用
- $M_{\mathrm{DPO_{D1}}}$: M_{base} に D_1 を chosen, D_0 を rejected として DPO を適用
- $M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$: $M_{\mathrm{SFT_{D2}}}$ に D_0 を chosen, D_1 を rejected として DPO を適用
- $M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$: $M_{\mathrm{SFT_{D2}}}$ に D_1 を chosen, D_0 を rejected として DPO を適用
- $M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$: $M_{\mathrm{SFT_{D2}}}$ に D_0 を chosen, D_1 を rejected として DPO を適用
- $M_{\mathrm{DPO_{D1}(M_{SFT_{D1}})}}$: $M_{\mathrm{SFT_{D2}}}$ に D_1 を chosen, D_0 を rejected として DPO を適用

学習した 9 個のモデルに対して, 以下の評価手法を用いて SFT や DPO の特性や影響範囲, 生成結果への影響を評価する.

- パラメータの重みからの評価

  L1 ノルム, L2 ノルム, Conflict Limited L2 ノルム

- 生成結果からの評価

  BLEU, BERTScore, LLM-as-a-judge, 定性的な評価

ここで, SFT には Transformers Reinforcement Learning (TRL) ライブラリ [49] 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 TRL 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 5.1, 5.2, 5.3 に各手法のパラメータを示す. SFTTrainer, DPOTrainer ともに学習率とエポック数を揃えている.

また, 表 5.4 にテキスト生成の際のパラメータを示す.

学習の際には ChatHaruhi [43] を参考として, 推論の際にも用いるシステムプロンプトを学習データに含めた. さらに, 学習後のモデルの出力を安定させるために LLM の末尾に生成の終わりを意味する tokenizer の EOS (End Of Sentense) トークンを付与した. 上記の処理を適用した学習データを示す.

表 5.1: 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 5.2: 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-5
学習率スケジューラ	cosine

実験 1,2 におけるシステムプロンプト, EOS トークンを加えた学習データ

I want you to act like a young lady. She always behaves gracefully. I want you to respond and answer like her, using the tone, manner and vocabulary she would use. I would like you to keep your response to about one sentence in length and brief, about 30 words. Please respond to the following questions with the above instructions and information. 必ず日本語で応答してください。
question
{ クエリ }
{ 応答 }{EOS トークン }

また, BLEU を計算する際のトークナイザーは rinna/japanese-roberta-base [6], BERTScore を計算する際の BERT には京都大学が公開している deberta-v2-tiny-japanese [**?**] を用いた.

LLM-as-a-judge において評価者 LLM は GPT-4o-mini を用いて以下のプロンプトでエージェントの口調を 10 段階でスコア化して評価する. プロンプトは Character-LLM のプロンプトを参考にした. ここでプロンプト内の {profile} はデータセットで定義されているペルソナを記述した文, {conversation_example} はデータセット内からランダムに抽出した会話例, {question}, {answer} は評価する会話のクエリと LLM の応答が入る.

[6] https://huggingface.co/rinna/japanese-roberta-base

表 5.3: 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
β	0.3

表 5.4: テキスト生成の際のパラメータ

パラメータ	値
max_token	128
do_sample	True
temperature	0.1
top_p	1.0

評価者 LLM に与えるプロンプト

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]

Tone (1-10): How well do the responses reflect the character's tone of voice?

[Evaluation Steps]

1. Review the profile and conversation examples to identify the manner of speaking or style (tone) of the original character.

2. Read through the interactions and examine the interactions and note the AI assistant's tone.

3. Compare the AI assistant's tone to the character's, based on the gathered information. Assess whether the response is natural and of appropriate length.

4. Use a 1–10 scale to rate how well the assistant's tone reflects the character's style. 1 means it does not reflect the character's tone at all, and 10 means it perfectly reflects the character's tone.

***

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.

5.2　実験 2

実験 1 で elyza/Llama-3-ELYZA-JP-8B としていたベースモデルを Qwen/Qwen2.5-7B-Instruct に変更し実験 1 と同条件で実験し, ベースモデルが異なる場合の影響を確かめる.

5.3　実験 3

実験 1 で用いていたデータセット D_1, D_2 を D_1', D_2' に変更し, データセットが異なる場合の結果を確かめる. ロールプレイするペルソナがお嬢様から男子大学生に変わるため, システムプロンプトは以下のようになる.

実験 3 における処理後の学習データ

I want you to act like a young male college student . he always behaves bluntly. I want you to respond and answer like him, using the tone, manner and vocabulary he would use. I would like you to keep your response to about one sentence in length and brief, about 30 words. Please respond to the following questions with the above instructions and information. 必ず日本語で応答してください。
question
{ クエリ }
{ 応答 }{EOS トークン }

6 結果と考察

以下, 各実験の結果とその考察を示す.

6.1 実験 1

表 6.1 にファインチューニング後のモデルとベースモデル間の重みの変化について, L1 ノルムおよび L2 ノルムを計算した結果を示す. 表 6.1 から, 同じエポック数かつ同じ学習率の場合でも SFT と DPO ではベースモデルからの重みの変化の大きさが SFT の方が大きくなっていることがわかる. これは、SFT がクロスエントロピー損失を最小化する手法であるのに対し, DPO は (2.6) 式に示す目的関数の最適化過程において係数 β によりベースモデルからの変化を抑制する働きを持つためと考えられる.

表 6.1: 実験 1 における (学習後のモデル, 学習前のモデル) の関係にある 2 モデル間の L1 ノルム, L2 ノルム

モデル組	L1 ノルム	L2 ノルム
$M_{\mathrm{SFT_{D0}}}$, M_{base}	2.32498×10^5	1.51147
$M_{\mathrm{SFT_{D1}}}$, M_{base}	2.37030×10^5	1.64791
$M_{\mathrm{SFT_{D2}}}$, M_{base}	2.34274×10^5	1.57759
$M_{\mathrm{DPO_{D0}}}$, M_{base}	1.54711×10^5	0.51305
$M_{\mathrm{DPO_{D1}}}$, M_{base}	1.38068×10^5	0.36422
$M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$, M_{base}	3.08548×10^5	2.67890
$M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$, M_{base}	2.97963×10^5	2.50216
$M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$, $M_{\mathrm{SFT_{D2}}}$	1.63760×10^5	0.58680
$M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$, $M_{\mathrm{SFT_{D2}}}$	1.47138×10^5	0.38904
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, M_{base}	3.12721×10^5	2.75796
$M_{\mathrm{DPO_{D1}(M_{SFT_{D1}})}}$, M_{base}	2.94477×10^5	2.46993
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D1}}}$	1.66718×10^5	0.58686
$M_{\mathrm{DPO_{D1}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D1}}}$	1.37648×10^5	0.30320

表 6.2, 6.3 に, それぞれ SFT を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値, DPO を適用したモデルとベースモデルとのタスクベ

クトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値を示す.

表 6.2 では全てのモデルに共通して変化量の大きい層が中間層と, 29, 30, 31 層の深い層となっている. LLM においてロールプレイを実現するために, SFT では中間層から深い層にかけて重みを変化していることがわかる. これは BERT での先行研究と同様に文脈情報などロールプレイを実現するための高度な言語情報は中間層から深い層で重みとして保持されていると考えられる. また最も L2 ノルム, すなわちベースモデルからの重みの変化量が大きいのは 31 層に共通しており, ロールプレイのタスクに最も重要な層は 31 層となっていることが考えられる.

表 6.3 ではデータセットによる違いが現れている. D_0 を chosen として DPO を適用したモデルは学習前と比較して 7 から 14 層で変化が大きくなっており, D_1 を chosen として DPO を適用したモデルは学習前と比較して 19 から 31 層まで変化が大きくなっている傾向が見られる. D_0 は一般的な応答を収録したデータセットであるため, ベースモデルから出力のスタイルは大きく変わらないためデータセット内に存在する語彙の情報を強く保持するために 7 から 14 層の中間層の変化量が大きくなっていると考えられる. D_1 はお嬢様スタイルを反映するため一人称などの固有名詞だけでなく文全体としてお嬢様スタイルの応答を統一する必要がある. そのためより高次の文脈情報を保持する深い層の重みの変化が激しくなっていると考えられる.

SFT ではこのような変化が見られず DPO だけでこのような傾向が見られたのは先述したように, DPO はベースモデルから大きく離れないような制約がある中でロールプレイを実現するために変化の影響が一定の層に集中したためと考えられる.

データセットの違いによる重みの変化の違いについてより詳しく考察するために, 学習方法が同一でデータセットが異なるモデル間の Conflict Limited L2 ノルムを計算する. 表 6.4 に $M_{\mathrm{SFT_{D0}}}$, $M_{\mathrm{SFT_{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.5 に各層の各パラメータのコンフリクト率の値を示す. なお, 表中ではコンフリクトを C と略している.

同様に, 表 6.6 に $M_{\mathrm{DPO_{D0}}}$, $M_{\mathrm{DPO_{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.7 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.8 に $M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$, $M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$ 間の コンフリクト 数, コンフリク

表 6.2: 実験 1 における SFT を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを, 全層について最大値を 1, 最小値を 0 として正規化した値 (太字は上位 15 層)

	$M_{\mathrm{SFT_{D0}}}$	$M_{\mathrm{SFT_{D1}}}$	$M_{\mathrm{SFT_{D2}}}$	$M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$	$M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$	$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$	$M_{\mathrm{DPO_{D1}(M_{SFT_{D1}})}}$
model.layers.0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
model.layers.1	0.13410	0.08287	0.07900	0.07605	0.07900	0.11550	0.06442
model.layers.2	0.02092	0.04209	0.01540	0.02966	0.01540	0.08906	0.04230
model.layers.3	0.09263	0.05512	0.01834	0.01551	0.01834	0.02367	0.03623
model.layers.4	0.15102	0.10250	0.10683	0.15934	0.10683	0.10951	0.13033
model.layers.5	0.14437	0.12467	0.13322	0.15990	0.13322	0.19026	0.17067
model.layers.6	0.18998	0.18343	0.19941	0.21341	0.19941	0.22662	0.23424
model.layers.7	0.25934	0.26840	0.27043	0.28756	0.27043	0.32924	0.31785
model.layers.8	0.24156	0.27981	0.28216	0.33285	0.28216	0.32549	0.34705
model.layers.9	0.36209	0.35401	0.40809	0.40450	0.40809	0.37958	0.40365
model.layers.10	0.42796	0.40584	0.40159	0.50470	0.40159	**0.49743**	0.47190
model.layers.11	**0.52657**	**0.48287**	**0.63576**	**0.72791**	**0.63576**	**0.62444**	**0.57985**
model.layers.12	**0.49928**	**0.50983**	**0.54722**	**0.66446**	**0.54722**	**0.58896**	**0.57152**
model.layers.13	**0.55414**	**0.55934**	**0.60010**	**0.74386**	**0.60010**	**0.60790**	**0.62366**
model.layers.14	**0.66680**	**0.59857**	**0.60296**	**0.71559**	**0.60296**	**0.64534**	**0.65056**
model.layers.15	**0.61526**	**0.58329**	**0.64371**	**0.68638**	**0.64371**	**0.50257**	**0.58900**
model.layers.16	**0.53631**	0.45411	**0.47500**	0.57693	**0.47500**	0.38497	0.49126
model.layers.17	**0.47559**	0.41180	0.45981	0.56899	0.45981	0.38644	0.47564
model.layers.18	**0.47135**	**0.57992**	**0.51529**	**0.59643**	**0.51529**	**0.47645**	**0.62522**
model.layers.19	**0.43682**	0.46129	0.42421	**0.57979**	0.42421	0.42700	0.53758
model.layers.20	0.37255	0.44102	0.40165	0.53554	0.40165	0.39753	0.51835
model.layers.21	0.41377	0.46779	**0.54563**	**0.61201**	**0.54563**	0.40137	0.54010
model.layers.22	**0.46423**	**0.53340**	0.43098	0.56106	0.43098	0.41907	**0.55927**
model.layers.23	**0.43830**	**0.54627**	0.32915	0.48215	0.32915	**0.48303**	**0.60587**
model.layers.24	0.38633	**0.50581**	**0.51277**	**0.65135**	**0.51277**	**0.42724**	**0.55627**
model.layers.25	0.42944	**0.50628**	**0.47166**	**0.63140**	**0.47166**	**0.48970**	**0.54623**
model.layers.26	0.38967	**0.49374**	**0.55834**	**0.66423**	**0.55834**	**0.47696**	**0.55888**
model.layers.27	**0.44322**	**0.52050**	**0.66203**	**0.72231**	**0.66203**	**0.46868**	**0.58607**
model.layers.28	0.37124	0.42178	0.43188	0.50818	0.43188	0.31367	0.46352
model.layers.29	**0.57333**	**0.67104**	**0.72108**	**0.81091**	**0.72108**	**0.67625**	**0.72207**
model.layers.30	**0.73263**	**0.82922**	**0.71693**	**0.80265**	**0.71693**	**0.82262**	**0.86572**
model.layers.31	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**

表 6.3: 実験 1 における DPO を適用後モデルと DPO を適用前モデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを, 全層について最大値を 1, 最小値を 0 として正規化した値 (太字は上位 15 層)

	$M_{\mathrm{DPO}_{\mathrm{D0}}}$	$M_{\mathrm{DPO}_{\mathrm{D1}}}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$	$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$
model.layers.0	0.00000	0.02570	0.00000	0.03629	0.35412	0.06459
model.layers.1	**0.51392**	0.33070	**0.43154**	0.25414	**0.80564**	0.24159
model.layers.2	0.17469	0.01881	0.14897	0.04775	**0.45116**	0.00000
model.layers.3	0.11499	0.00000	0.02272	0.00000	0.36164	0.02483
model.layers.4	0.38950	0.08587	**0.44078**	0.31362	0.43634	0.23350
model.layers.5	0.46033	0.25848	**0.45191**	0.40394	**0.67274**	0.31159
model.layers.6	0.19621	0.06702	0.19728	0.14230	0.37911	0.11735
model.layers.7	0.38494	0.17672	0.26374	0.29261	**0.56199**	0.27951
model.layers.8	0.46251	0.17805	**0.37827**	0.38058	**0.48566**	0.37360
model.layers.9	**0.56682**	0.23138	**0.49778**	0.42722	**0.52218**	0.37833
model.layers.10	**0.61342**	0.26539	**0.74145**	0.47837	**0.79276**	0.55713
model.layers.11	**0.61426**	0.26251	**0.91107**	**0.62872**	**1.00000**	**0.75267**
model.layers.12	**0.63835**	0.26341	**0.71676**	0.38772	**0.77362**	0.55225
model.layers.13	**0.98024**	**0.56449**	**1.00000**	**0.82313**	**0.84414**	**0.85941**
model.layers.14	**1.00000**	**0.68910**	**0.93259**	**0.84080**	**0.94909**	**0.88384**
model.layers.15	0.45838	0.33923	0.32015	0.49715	0.36739	0.59123
model.layers.16	0.46371	0.46922	0.19744	0.59934	0.29063	0.55446
model.layers.17	0.26485	0.38952	**0.51576**	**0.78237**	**0.54068**	**0.73666**
model.layers.18	0.21221	0.42675	0.31207	**0.83363**	0.24699	**0.78532**
model.layers.19	**0.55992**	**0.68598**	0.19643	**0.68801**	**0.43923**	**0.84834**
model.layers.20	0.40030	**0.47351**	0.19188	0.58205	0.35840	**0.73911**
model.layers.21	0.39316	**0.76513**	0.24349	**0.87919**	0.36627	**1.00000**
model.layers.22	0.45029	**0.51644**	0.13862	0.57595	0.06331	0.51171
model.layers.23	0.37373	**0.86224**	0.23026	**1.00000**	0.42741	**0.89145**
model.layers.24	**0.80484**	**0.68092**	0.16918	**0.72749**	0.16719	**0.64208**
model.layers.25	**0.78144**	**0.83861**	**0.54875**	**0.81620**	0.43002	**0.66005**
model.layers.26	**0.49797**	**0.78954**	0.28471	**0.83349**	**0.49011**	**0.80845**
model.layers.27	**0.71971**	**0.81422**	0.35441	**0.63551**	0.30562	**0.74414**
model.layers.28	0.30317	**0.62917**	0.06225	0.44327	0.00000	0.53683
model.layers.29	**0.58542**	**0.62667**	**0.62005**	**0.61001**	0.42933	**0.76541**
model.layers.30	**0.49807**	**1.00000**	**0.62713**	**0.74523**	**0.47652**	**0.94908**
model.layers.31	**0.62646**	**0.72610**	**0.93942**	**0.69110**	0.33696	0.49060

表 6.4: 実験 1 における $M_{\mathrm{SFT}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム の値 (太字は上位 15 層)

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	120	5.5018e-07	0.00438	0.00000	0.00000
model.layers.1	583	2.6729e-06	0.01011	0.03321	0.12635
model.layers.2	229	1.0499e-06	0.00614	0.00782	0.03874
model.layers.3	303	1.3892e-06	0.00677	0.01313	0.05269
model.layers.4	336	1.5405e-06	0.00733	0.01549	0.06508
model.layers.5	283	1.2975e-06	0.00687	0.01169	0.05490
model.layers.6	307	1.4075e-06	0.00719	0.01341	0.06192
model.layers.7	613	2.8105e-06	0.00987	0.03536	0.12102
model.layers.8	483	2.2145e-06	0.00887	0.02604	0.09892
model.layers.9	806	3.6953e-06	0.01118	0.04921	0.14993
model.layers.10	1144	5.245e-06	0.01350	0.07345	0.20116
model.layers.11	1848	8.4727e-06	0.01696	0.12395	0.27740
model.layers.12	1434	6.5746e-06	0.01537	0.09425	0.24240
model.layers.13	3568	1.6359e-05	0.02396	0.24733	0.43180
model.layers.14	3842	1.7615e-05	0.02445	0.26698	0.44257
model.layers.15	**6688**	**3.0663e-05**	**0.03332**	**0.47113**	**0.63813**
model.layers.16	**4836**	**2.2172e-05**	**0.02809**	**0.33828**	**0.52285**
model.layers.17	**5096**	**2.3364e-05**	**0.02892**	**0.35693**	**0.54107**
model.layers.18	**10577**	**4.8493e-05**	**0.04275**	**0.75009**	**0.84612**
model.layers.19	**6446**	**2.9554e-05**	**0.03305**	**0.45377**	**0.63217**
model.layers.20	3707	1.6996e-05	0.02450	0.25730	0.44357
model.layers.21	**4067**	**1.8646e-05**	**0.02529**	**0.28312**	**0.46112**
model.layers.22	**8805**	**4.0369e-05**	**0.03798**	**0.62298**	**0.74082**
model.layers.23	**9957**	**4.5651e-05**	**0.04076**	**0.70562**	**0.80230**
model.layers.24	**5262**	**2.4125e-05**	**0.02951**	**0.36884**	**0.55416**
model.layers.25	**5635**	**2.5835e-05**	**0.03026**	**0.39560**	**0.57061**
model.layers.26	**4611**	**2.1141e-05**	**0.02858**	**0.32214**	**0.53366**
model.layers.27	**4707**	**2.1581e-05**	**0.02832**	**0.32903**	**0.52798**
model.layers.28	3172	1.4543e-05	0.02295	0.21892	0.40938
model.layers.29	**3981**	**1.8252e-05**	**0.02548**	**0.27695**	**0.46519**
model.layers.30	**14061**	**6.4467e-05**	**0.04973**	**1.00000**	**1.00000**
model.layers.31	**8320**	**3.8146e-05**	**0.03846**	**0.58819**	**0.75150**

表 6.5: 実験 1 における $M_{\mathrm{SFT}_{\mathrm{D0}}}, M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の各パラメータのコンフリクト率の値 (太字は上位 15 層)

パラメータ	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	2.9802e-07	8.8215e-06	**1.0014e-05**	0.0000e+00	4.2575e-07	1.8733e-07	0.0000e+00
model.layer.1	1.4007e-05	1.0729e-05	**1.7643e-05**	2.1458e-06	3.7466e-07	1.3624e-07	2.7759e-06
model.layer.2	1.7881e-07	**4.6253e-05**	1.9073e-06	0.0000e+00	3.0654e-07	1.0218e-07	0.0000e+00
model.layer.3	1.2517e-06	**3.3617e-05**	7.1526e-07	1.3709e-06	1.4646e-06	4.9387e-07	0.0000e+00
model.layer.4	1.2517e-06	**3.2663e-05**	1.9073e-06	2.9802e-06	1.7200e-06	3.2357e-07	0.0000e+00
model.layer.5	8.3447e-07	**1.5259e-05**	1.0729e-05	5.3048e-06	1.0899e-06	1.1921e-07	0.0000e+00
model.layer.6	3.6359e-06	**1.9550e-05**	5.9605e-06	3.2187e-06	7.6635e-07	6.8120e-07	0.0000e+00
model.layer.7	4.8876e-06	**4.3869e-05**	1.1921e-06	6.0201e-06	1.3965e-06	2.7078e-06	0.0000e+00
model.layer.8	2.9802e-06	5.0068e-06	**6.4373e-06**	4.2319e-06	3.7806e-06	1.5668e-06	0.0000e+00
model.layer.9	3.0398e-06	7.1526e-06	6.6757e-06	5.5432e-06	2.6226e-06	**7.6635e-06**	0.0000e+00
model.layer.10	5.8413e-06	**5.2691e-05**	5.9605e-06	5.4240e-06	9.1621e-06	2.9121e-06	0.0000e+00
model.layer.11	1.0729e-06	1.7643e-05	3.0994e-06	**2.8014e-05**	4.5129e-06	1.7166e-05	0.0000e+00
model.layer.12	3.9935e-06	**3.5524e-05**	2.3842e-06	6.6757e-06	9.3664e-06	9.2983e-06	0.0000e+00
model.layer.13	1.8001e-05	3.1710e-05	7.1526e-07	2.8014e-06	1.7200e-05	**3.5303e-05**	0.0000e+00
model.layer.14	3.1590e-06	1.0252e-05	9.7752e-06	6.0201e-06	1.8443e-05	**4.2932e-05**	0.0000e+00
model.layer.15	1.0133e-05	4.4107e-05	3.1471e-05	7.6890e-06	4.7990e-05	**5.5398e-05**	1.7030e-08
model.layer.16	3.4451e-05	3.6716e-05	**7.4387e-05**	1.0788e-05	2.5868e-05	3.5610e-05	1.7030e-08
model.layer.17	7.2718e-06	**9.1791e-05**	9.7752e-06	3.0220e-05	6.0490e-05	8.2765e-06	5.1090e-08
model.layer.18	3.7789e-05	1.5736e-04	**6.5780e-04**	3.7968e-05	4.7565e-05	5.2691e-05	0.0000e+00
model.layer.19	3.0041e-05	**9.0837e-05**	5.5075e-05	2.2650e-05	7.5749e-05	8.5490e-06	0.0000e+00
model.layer.20	8.9407e-06	**1.4162e-04**	8.0347e-05	1.8299e-05	2.2326e-05	1.7166e-05	0.0000e+00
model.layer.21	5.0068e-06	**1.2064e-04**	3.2425e-05	5.7936e-05	1.4952e-05	2.5392e-05	0.0000e+00
model.layer.22	2.3067e-05	9.7513e-05	5.4121e-05	5.0008e-05	1.9261e-05	**9.8978e-05**	0.0000e+00
model.layer.23	4.2915e-06	**1.0395e-04**	1.6689e-05	1.2994e-05	1.0223e-04	5.3780e-05	0.0000e+00
model.layer.24	2.0325e-05	**1.3614e-04**	8.1062e-06	1.2589e-04	1.4424e-05	2.3110e-05	0.0000e+00
model.layer.25	4.2915e-05	**8.3685e-05**	6.2227e-05	7.8380e-05	2.0317e-05	3.0569e-05	0.0000e+00
model.layer.26	6.0797e-06	**4.5657e-04**	3.9816e-05	1.2934e-05	2.2224e-05	1.5412e-05	0.0000e+00
model.layer.27	3.1948e-05	**2.3389e-04**	4.1008e-05	5.1320e-05	3.2561e-05	4.1723e-06	0.0000e+00
model.layer.28	3.5644e-05	**5.3644e-05**	3.8147e-06	1.9372e-05	2.7657e-05	6.5395e-06	0.0000e+00
model.layer.29	2.9862e-05	3.1471e-05	**4.2439e-05**	2.8253e-05	2.3586e-05	2.2156e-05	1.7030e-07
model.layer.30	1.3560e-04	**2.3103e-04**	1.1921e-05	3.7849e-05	8.2919e-05	8.9117e-05	5.1090e-07
model.layer.31	4.1962e-05	8.7500e-05	3.6716e-05	**1.2630e-04**	5.6096e-05	2.7861e-05	7.8338e-07

表 6.6: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{DPO}_{\mathrm{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム の値 (太字は上位 15 層)

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	2656	1.2177e-05	0.02065	0.00000	0.00000
model.layers.1	**13560**	**6.217e-05**	**0.05236**	**0.43126**	**0.66380**
model.layers.2	3158	1.4479e-05	0.02214	0.01985	0.03114
model.layers.3	3096	1.4195e-05	0.02162	0.01740	0.02019
model.layers.4	4698	2.1539e-05	0.02773	0.08076	0.14806
model.layers.5	9219	4.2267e-05	0.03799	0.25957	0.36298
model.layers.6	3417	1.5666e-05	0.02306	0.03010	0.05040
model.layers.7	5167	2.369e-05	0.02768	0.09931	0.14718
model.layers.8	5215	2.391e-05	0.02858	0.10121	0.16587
model.layers.9	7206	3.3038e-05	0.03367	0.17996	0.27247
model.layers.10	8399	3.8508e-05	0.03652	0.22714	0.33219
model.layers.11	7365	3.3767e-05	0.03420	0.18624	0.28350
model.layers.12	7945	3.6426e-05	0.03501	0.20918	0.30048
model.layers.13	**20207**	**9.2645e-05**	**0.05606**	**0.69415**	**0.74119**
model.layers.14	**25755**	**0.00011808**	**0.06372**	**0.91358**	**0.90143**
model.layers.15	**10848**	**4.9736e-05**	**0.04373**	**0.32400**	**0.48300**
model.layers.16	**12054**	**5.5265e-05**	**0.04252**	**0.37170**	**0.45764**
model.layers.17	9178	4.2079e-05	0.03825	0.25795	0.36841
model.layers.18	7346	3.368e-05	0.03406	0.18549	0.28055
model.layers.19	**15293**	**7.0115e-05**	**0.04918**	**0.49980**	**0.59705**
model.layers.20	6708	3.0755e-05	0.03247	0.16026	0.24734
model.layers.21	**18174**	**8.3324e-05**	**0.05371**	**0.61375**	**0.69202**
model.layers.22	10262	4.7049e-05	0.04033	0.30082	0.41179
model.layers.23	**15042**	**6.8965e-05**	**0.04850**	**0.48988**	**0.58300**
model.layers.24	**20134**	**9.231e-05**	**0.05802**	**0.69127**	**0.78225**
model.layers.25	**27940**	**0.0001281**	**0.06843**	**1.00000**	**1.00000**
model.layers.26	**14857**	**6.8116e-05**	**0.04873**	**0.48256**	**0.58772**
model.layers.27	**17719**	**8.1238e-05**	**0.05399**	**0.59575**	**0.69777**
model.layers.28	9048	4.1483e-05	0.03819	0.25281	0.36709
model.layers.29	**19183**	**8.795e-05**	**0.05525**	**0.65365**	**0.72413**
model.layers.30	**13767**	**6.3119e-05**	**0.04867**	**0.43945**	**0.58654**
model.layers.31	**25498**	**0.0001169**	**0.06452**	**0.90342**	**0.91816**

表 6.7: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{DPO}_{\mathrm{D}1}}$ 間の各パラメータのコンフリクト率の値 (太字は上位 15 層)

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	3.9816e-05	**9.3222e-05**	7.0333e-05	2.7895e-05	9.1280e-06	5.0749e-06	0.0000e+00
model.layer.1	**4.8971e-04**	7.4387e-05	1.6403e-04	6.9737e-06	4.2915e-06	7.1185e-06	6.0575e-05
model.layer.2	4.0889e-05	**1.1969e-04**	2.0742e-05	2.3007e-05	1.3965e-05	1.1529e-05	0.0000e+00
model.layer.3	7.9274e-06	**3.4571e-04**	3.4809e-05	1.0133e-05	1.1121e-05	9.2643e-06	0.0000e+00
model.layer.4	7.5698e-05	4.2439e-05	**8.2016e-05**	9.1791e-06	2.3007e-05	2.3859e-05	0.0000e+00
model.layer.5	7.2181e-05	**1.9503e-04**	6.0081e-05	9.7811e-05	6.6757e-05	2.3450e-05	0.0000e+00
model.layer.6	1.5855e-05	**1.2732e-04**	3.1710e-05	1.2279e-05	2.4302e-05	1.4492e-05	0.0000e+00
model.layer.7	1.5140e-05	6.4850e-05	2.6941e-05	**9.5844e-05**	2.1338e-05	2.8389e-05	0.0000e+00
model.layer.8	1.8895e-05	**1.6141e-04**	1.3995e-04	2.5570e-05	2.7980e-05	2.6601e-05	0.0000e+00
model.layer.9	1.4663e-05	9.7036e-05	**1.4687e-04**	1.4561e-04	4.0020e-05	1.9482e-05	0.0000e+00
model.layer.10	9.1255e-05	**3.2425e-04**	8.2016e-05	6.4909e-05	2.8576e-05	4.0821e-05	0.0000e+00
model.layer.11	9.9123e-05	7.2479e-05	**1.4758e-04**	7.4267e-05	3.7330e-05	2.2837e-05	0.0000e+00
model.layer.12	5.5075e-05	**1.5187e-04**	7.9393e-05	7.0095e-05	4.7088e-05	3.5933e-05	0.0000e+00
model.layer.13	1.5754e-04	**6.3038e-04**	5.6028e-05	1.7011e-04	3.8505e-05	1.6298e-04	0.0000e+00
model.layer.14	8.6546e-05	1.9622e-04	1.7238e-04	**6.6411e-04**	7.7827e-05	1.1998e-04	0.0000e+00
model.layer.15	1.7160e-04	4.5872e-04	**5.1641e-04**	3.0041e-05	3.5627e-05	2.1815e-05	3.4060e-08
model.layer.16	2.1994e-05	1.1039e-04	2.7657e-05	1.5020e-05	3.7977e-05	**1.4687e-04**	0.0000e+00
model.layer.17	1.0818e-04	**4.9758e-04**	7.6771e-05	2.5749e-05	4.3562e-05	3.3447e-05	0.0000e+00
model.layer.18	7.3314e-06	**3.1853e-04**	2.9588e-04	1.2636e-05	3.7074e-05	3.8436e-05	0.0000e+00
model.layer.19	9.3997e-05	**1.0400e-03**	2.0623e-04	1.3417e-04	3.6819e-05	6.9227e-05	0.0000e+00
model.layer.20	9.4175e-05	**2.1648e-04**	4.1962e-05	4.4107e-05	2.2394e-05	3.3872e-05	0.0000e+00
model.layer.21	3.7611e-05	**6.0439e-04**	5.6982e-05	5.9545e-05	2.2473e-04	9.7752e-06	0.0000e+00
model.layer.22	8.9169e-05	**3.4499e-04**	3.3998e-04	7.4506e-06	1.7575e-05	8.0654e-05	0.0000e+00
model.layer.23	1.8078e-04	**4.3416e-04**	1.6594e-04	2.1011e-04	5.3576e-05	4.8041e-05	0.0000e+00
model.layer.24	1.9741e-04	2.5129e-04	**9.9754e-04**	5.1093e-04	1.9584e-05	3.1710e-05	0.0000e+00
model.layer.25	**7.5442e-04**	5.1713e-04	2.6965e-04	7.0155e-05	1.6497e-04	1.9039e-05	1.7030e-08
model.layer.26	1.7506e-04	**7.0167e-04**	2.2173e-04	1.8436e-04	2.0317e-05	6.4049e-05	0.0000e+00
model.layer.27	3.1877e-04	**7.4363e-04**	3.5143e-04	1.4335e-04	8.2408e-05	9.0940e-06	0.0000e+00
model.layer.28	1.7405e-04	1.8644e-04	**2.7800e-04**	4.4465e-05	2.9768e-05	2.8712e-05	0.0000e+00
model.layer.29	1.5271e-04	5.8913e-04	**6.9308e-04**	3.3891e-04	1.9125e-05	7.5511e-05	0.0000e+00
model.layer.30	3.6091e-04	**4.5538e-04**	2.1577e-04	7.1764e-05	4.5725e-05	1.7132e-05	3.4060e-08
model.layer.31	4.2278e-04	**1.4100e-03**	2.7275e-04	1.7244e-04	1.2858e-05	1.0913e-04	2.2003e-05

ト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.9 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.10 に $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.9 に各層の各パラメータのコンフリクト率の値を示す.

表 6.4, 6.6, 6.8, 6.10 では 10 から 18 層目の中間層と 25 から 31 層目の深い層において Conflict Limited L2 ノルムが大きい値となる傾向が見られる. このことから, LLM はおおよそ 10 層目以降の中間層から深い層でデータセット間の違いを捉えていると考えることができ, データセット間の違いである一人称や固有名詞, 特徴的な語尾を捉えているといえる. 中間層では浅い層で得られた文脈理解を基によりデータセット内の固有名詞や特徴的な一人称や語尾などのデータセット固有の概念を学習し, 深い層で出力文のスタイルをコントロールする要素を保持していると考えられる. 実際に上位 15 層に共通して入っている層は, 29 から 31 層となっており, この部分がデータセットの違いをより敏感に捉えておりロールプレイのタスクを果たすために大きな意味を持っていると考えられる.

また, 表 6.5 , 表 6.7, 6.9, 表 6.11 において, 共通してほとんどの層において v_proj のコンフリクト率が高くなっている. このことは全層にわたって, q_proj, k_proj で作り出された Attention 分布にかかる重みがデータセット間で異なることを意味しており, キャラクター固有の出力を生成する必要があるロールプレイにおいて v_proj の重みがスタイルや表現の微細な調整に大きく寄与している可能性があるといえる.

また, SFT と DPO の手法の差を詳しく調べるために, 同じデータセットで手法が異なる $M_{\mathrm{DPO}_{\mathrm{D0}}}$ と $M_{\mathrm{SFT}_{\mathrm{D0}}}$, $M_{\mathrm{DPO}_{\mathrm{D1}}}$ と $M_{\mathrm{SFT}_{\mathrm{D1}}}$ に関する Conflict Limited L2 ノルム を調べる. 表 6.12 に $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.13 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.14 に $M_{\mathrm{DPO}_{\mathrm{D1}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.15 に各層の各パラメータのコンフリクト率の値を示す.

表 6.8: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$ 間の M_{base} から計算したタスクベクトルのコンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム の値 (太字は上位 15 層)

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	5652	2.5913e-05	0.03020	0.00654	0.00000
model.layers.1	10876	4.9864e-05	**0.04877**	0.18938	**0.37837**
model.layers.2	7108	3.2589e-05	0.03386	0.05750	0.07453
model.layers.3	6104	2.7986e-05	0.03179	0.02236	0.03233
model.layers.4	12350	5.6622e-05	0.04613	0.24097	0.32454
model.layers.5	**13841**	**6.3458e-05**	**0.04752**	**0.29315**	**0.35275**
model.layers.6	6892	3.1598e-05	0.03377	0.04994	0.07273
model.layers.7	8221	3.7692e-05	0.03631	0.09646	0.12441
model.layers.8	8760	4.0163e-05	0.03708	0.11532	0.14008
model.layers.9	**12416**	**5.6925e-05**	0.04531	**0.24328**	0.30779
model.layers.10	**14859**	**6.8126e-05**	**0.04963**	**0.32878**	**0.39583**
model.layers.11	**16570**	**7.597e-05**	**0.05141**	**0.38867**	**0.43216**
model.layers.12	12130	5.5614e-05	0.04429	0.23327	0.28696
model.layers.13	**24772**	**0.00011357**	**0.06408**	**0.67573**	**0.69027**
model.layers.14	**34037**	**0.00015605**	**0.07663**	**1.00000**	**0.94608**
model.layers.15	11218	5.1432e-05	0.04329	0.20135	0.26674
model.layers.16	9626	4.4133e-05	0.04008	0.14563	0.20122
model.layers.17	**20242**	**9.2806e-05**	**0.05833**	**0.51718**	**0.57323**
model.layers.18	**20111**	**9.2205e-05**	**0.05958**	**0.51260**	**0.59861**
model.layers.19	8751	4.0122e-05	0.03765	0.11501	0.15182
model.layers.20	7773	3.5638e-05	0.03576	0.08078	0.11325
model.layers.21	**15607**	**7.1555e-05**	**0.05087**	**0.35496**	**0.42112**
model.layers.22	8527	3.9095e-05	0.03676	0.10717	0.13350
model.layers.23	**18053**	**8.2769e-05**	**0.05459**	**0.44057**	**0.49702**
model.layers.24	12034	5.5173e-05	0.04450	0.22991	0.29138
model.layers.25	**17382**	**7.9693e-05**	**0.05438**	**0.41709**	**0.49265**
model.layers.26	**14684**	**6.7323e-05**	**0.05035**	**0.32266**	**0.41061**
model.layers.27	12345	5.6599e-05	0.04561	0.24080	0.31395
model.layers.28	5465	2.5056e-05	0.03044	0.00000	0.00473
model.layers.29	**15762**	**7.2266e-05**	**0.05139**	**0.36039**	**0.43172**
model.layers.30	**22541**	**0.00010335**	**0.06156**	**0.59765**	**0.63893**
model.layers.31	**20835**	**9.5524e-05**	**0.07928**	**0.53794**	**1.00000**

表 6.9: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$ 間の M_{base} から計算したタスクベクトルの各パラメータのコンフリクト率の値 (太字は上位 15 層)

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	5.9307e-05	**1.2207e-04**	1.0467e-04	8.3625e-05	2.5017e-05	1.4203e-05	0.0000e+00
model.layer.1	1.2642e-04	9.7275e-05	**1.8883e-04**	4.4107e-06	1.4748e-05	1.3794e-05	9.8859e-05
model.layer.2	**1.5301e-04**	8.4639e-05	5.9128e-05	4.0770e-05	2.9036e-05	2.6379e-05	0.0000e+00
model.layer.3	4.7147e-05	**5.3596e-04**	9.3460e-05	1.5378e-05	1.9363e-05	2.1764e-05	0.0000e+00
model.layer.4	1.3119e-04	1.0753e-04	**6.2537e-04**	5.6207e-05	5.0988e-05	5.3440e-05	0.0000e+00
model.layer.5	6.7770e-05	**4.9210e-04**	1.5807e-04	7.4744e-05	1.1047e-04	3.8079e-05	0.0000e+00
model.layer.6	4.7088e-05	**2.6512e-04**	1.0228e-04	2.7001e-05	3.9203e-05	3.0756e-05	0.0000e+00
model.layer.7	4.4405e-05	**1.4830e-04**	6.3658e-05	1.3202e-04	3.7023e-05	3.7432e-05	0.0000e+00
model.layer.8	2.3603e-05	**1.4687e-04**	8.1539e-05	1.3661e-04	3.6052e-05	5.1039e-05	0.0000e+00
model.layer.9	6.3837e-05	**3.4976e-04**	1.4639e-04	2.3806e-04	5.5075e-05	3.4673e-05	0.0000e+00
model.layer.10	1.4436e-04	**2.6798e-04**	2.4009e-04	1.9956e-04	4.4959e-05	7.3365e-05	1.7030e-07
model.layer.11	4.7147e-05	**2.9469e-04**	1.8477e-04	1.9568e-04	6.6928e-05	1.1161e-04	1.7030e-08
model.layer.12	4.9472e-05	**3.4356e-04**	8.8453e-05	1.5032e-04	7.1985e-05	4.6645e-05	0.0000e+00
model.layer.13	3.4046e-04	**7.8177e-04**	9.1791e-05	4.1211e-04	3.6223e-05	1.0823e-04	0.0000e+00
model.layer.14	**7.4446e-04**	2.5582e-04	2.0075e-04	4.5067e-04	5.9111e-05	1.4646e-04	0.0000e+00
model.layer.15	5.3108e-05	**4.8900e-04**	1.2064e-04	1.0419e-04	4.1502e-05	6.1035e-05	1.7030e-08
model.layer.16	6.9439e-05	7.3433e-05	**1.5998e-04**	1.1563e-05	5.2520e-05	7.1577e-05	1.7030e-08
model.layer.17	2.0903e-04	**6.1774e-04**	2.1696e-04	9.4652e-05	1.2364e-04	7.4693e-05	0.0000e+00
model.layer.18	4.9770e-05	5.6458e-04	**1.4000e-03**	3.9160e-05	8.5524e-05	9.1399e-05	1.7030e-08
model.layer.19	8.9526e-05	**2.3341e-04**	1.7381e-04	8.2254e-05	1.6655e-05	5.4206e-05	0.0000e+00
model.layer.20	5.7340e-05	**2.9254e-04**	9.3460e-05	4.5776e-05	3.7227e-05	3.8096e-05	1.7030e-08
model.layer.21	2.7180e-05	**2.6631e-04**	1.4091e-04	2.1911e-04	1.4419e-04	2.2122e-05	1.7030e-08
model.layer.22	1.7524e-05	**1.9741e-04**	8.8930e-05	9.2149e-05	4.0208e-05	5.3218e-05	0.0000e+00
model.layer.23	7.5996e-05	2.5535e-04	**4.9400e-04**	8.8871e-05	5.7765e-05	1.4905e-04	0.0000e+00
model.layer.24	7.0691e-05	**3.1590e-04**	2.2817e-04	3.4988e-05	3.4179e-05	1.0170e-04	0.0000e+00
model.layer.25	1.8036e-04	**6.2656e-04**	4.3702e-04	7.5161e-05	1.2394e-04	2.3093e-05	0.0000e+00
model.layer.26	2.9033e-04	3.0422e-04	**3.9959e-04**	5.1618e-05	3.9220e-05	6.2874e-05	0.0000e+00
model.layer.27	5.2094e-05	**6.8092e-04**	1.2231e-04	1.4067e-04	7.3569e-05	2.4199e-05	1.7030e-08
model.layer.28	5.9724e-05	**2.1672e-04**	5.6267e-05	1.2994e-05	3.0977e-05	2.1815e-05	0.0000e+00
model.layer.29	2.4462e-04	**4.3011e-04**	2.0504e-04	2.5171e-04	2.1560e-05	5.9690e-05	0.0000e+00
model.layer.30	2.9153e-04	2.5988e-04	1.9479e-04	**3.8791e-04**	4.6441e-05	1.1083e-04	0.0000e+00
model.layer.31	2.6017e-04	**4.3702e-04**	3.5214e-04	6.5267e-05	6.6672e-05	1.3835e-04	4.4278e-07

表 6.10: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$ 間の M_{base} から計算したタスクベクトルのコンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム の値 (太字は上位 15 層)

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	3483	1.5969e-05	0.02418	0.04638	0.09820
model.layers.1	8176	3.7485e-05	**0.04091**	0.27574	**0.50516**
model.layers.2	2534	1.1618e-05	0.02015	0.00000	0.00000
model.layers.3	3099	1.4208e-05	0.02212	0.02761	0.04811
model.layers.4	4654	2.1338e-05	0.02866	0.10361	0.20720
model.layers.5	6025	2.7623e-05	0.03082	0.17062	0.25970
model.layers.6	3111	1.4263e-05	0.02253	0.02820	0.05805
model.layers.7	4607	2.1122e-05	0.02700	0.10131	0.16675
model.layers.8	4464	2.0467e-05	0.02660	0.09433	0.15712
model.layers.9	5364	2.4593e-05	0.02965	0.13831	0.23135
model.layers.10	**10575**	**4.8484e-05**	**0.04166**	**0.39299**	**0.52364**
model.layers.11	**13179**	**6.0423e-05**	**0.04612**	**0.52026**	**0.63216**
model.layers.12	**9638**	**4.4188e-05**	**0.03934**	**0.34720**	**0.46714**
model.layers.13	**14319**	**6.565e-05**	**0.04826**	**0.57597**	**0.68413**
model.layers.14	**22995**	**0.00010543**	**0.06124**	**1.00000**	**1.00000**
model.layers.15	8257	3.7857e-05	0.03812	0.27970	0.43740
model.layers.16	6053	2.7752e-05	0.03219	0.17199	0.29315
model.layers.17	**13855**	**6.3522e-05**	**0.04802**	**0.55330**	**0.67823**
model.layers.18	**10253**	**4.7008e-05**	**0.04245**	**0.37725**	**0.54278**
model.layers.19	**10125**	**4.6421e-05**	**0.04203**	**0.37100**	**0.53262**
model.layers.20	**9725**	**4.4587e-05**	**0.04113**	**0.35145**	**0.51054**
model.layers.21	**12353**	**5.6636e-05**	**0.04551**	**0.47989**	**0.61716**
model.layers.22	4307	1.9747e-05	0.02765	0.08665	0.18260
model.layers.23	**10040**	**4.6031e-05**	**0.04057**	**0.36684**	**0.49690**
model.layers.24	5612	2.573e-05	0.03090	0.15043	0.26167
model.layers.25	7276	3.3359e-05	0.03550	0.23176	0.37373
model.layers.26	**8650**	**3.9659e-05**	0.03909	**0.29891**	0.46092
model.layers.27	8108	3.7174e-05	0.03697	0.27242	0.40937
model.layers.28	3125	1.4328e-05	0.02372	0.02888	0.08687
model.layers.29	**10239**	**4.6944e-05**	**0.04080**	**0.37657**	**0.50254**
model.layers.30	**14864**	**6.8148e-05**	**0.05089**	**0.60261**	**0.74824**
model.layers.31	**8713**	**3.9947e-05**	**0.03960**	**0.30199**	**0.47353**

表 6.11: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$ 間の M_{base} から計算したタスクベクトルの各パラメータのコンフリクト率の値 (太字は上位 15 層)

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	6.9618e-05	5.7459e-05	**1.1849e-04**	4.3035e-05	9.3324e-06	5.2282e-06	0.0000e+00
model.layer.1	6.4135e-05	4.6730e-05	6.4135e-05	2.8610e-06	6.1478e-06	5.4666e-06	**1.0056e-04**
model.layer.2	3.1352e-05	**6.2943e-05**	2.1219e-05	1.0073e-05	1.3045e-05	1.2262e-05	0.0000e+00
model.layer.3	2.6405e-05	**3.4690e-04**	2.4557e-05	7.0333e-06	6.5565e-06	1.0133e-05	0.0000e+00
model.layer.4	4.9949e-05	2.2888e-05	**3.3879e-04**	2.6941e-05	1.9653e-05	1.1802e-05	0.0000e+00
model.layer.5	2.2769e-05	**1.9240e-04**	3.6955e-05	2.7955e-05	5.3235e-05	1.8494e-05	0.0000e+00
model.layer.6	1.9550e-05	**1.1301e-04**	3.0756e-05	1.7226e-05	1.6349e-05	1.5855e-05	0.0000e+00
model.layer.7	2.4736e-05	5.3883e-05	3.6955e-05	**8.4043e-05**	2.0129e-05	2.0759e-05	0.0000e+00
model.layer.8	1.2577e-05	7.2479e-05	**9.4414e-05**	4.2260e-05	1.7405e-05	3.1028e-05	0.0000e+00
model.layer.9	2.3484e-05	**1.3971e-04**	6.7949e-05	8.4698e-05	3.1352e-05	1.4254e-05	0.0000e+00
model.layer.10	1.1110e-04	**3.8266e-04**	1.7262e-04	1.0711e-04	1.9346e-05	5.8736e-05	0.0000e+00
model.layer.11	5.8413e-05	**2.5511e-04**	1.7977e-04	9.9540e-05	4.8263e-05	9.9983e-05	0.0000e+00
model.layer.12	2.3782e-05	**3.1209e-04**	3.5286e-05	6.3360e-05	6.7353e-05	4.7071e-05	0.0000e+00
model.layer.13	1.3131e-04	**6.8617e-04**	9.7275e-05	1.8603e-04	2.8389e-05	6.8835e-05	0.0000e+00
model.layer.14	2.6095e-04	2.1577e-04	1.5116e-04	**4.6909e-04**	4.5623e-05	1.1119e-04	0.0000e+00
model.layer.15	9.6738e-05	**4.0746e-04**	2.1958e-04	6.1154e-05	2.7997e-05	2.2684e-05	3.4060e-08
model.layer.16	6.6400e-05	8.4162e-05	**1.4067e-04**	3.7551e-06	2.9002e-05	3.7977e-05	0.0000e+00
model.layer.17	1.0878e-04	**5.2309e-04**	1.3995e-04	1.5140e-04	8.1011e-05	3.3242e-05	0.0000e+00
model.layer.18	1.4067e-05	3.6263e-04	**7.1383e-04**	1.6749e-05	4.8161e-05	4.0753e-05	0.0000e+00
model.layer.19	2.0224e-04	**3.6740e-04**	2.4819e-04	3.2425e-05	1.1802e-05	4.9608e-05	0.0000e+00
model.layer.20	**2.8926e-04**	2.6631e-04	1.0514e-04	3.7253e-05	1.7319e-05	2.8457e-05	1.7030e-08
model.layer.21	2.5570e-05	**4.6587e-04**	2.0671e-04	1.6797e-04	9.6134e-05	1.0899e-05	0.0000e+00
model.layer.22	3.7014e-05	1.5926e-04	**1.9169e-04**	1.3173e-05	2.3382e-05	1.0559e-05	0.0000e+00
model.layer.23	7.9453e-05	**2.0671e-04**	1.5306e-04	8.3864e-05	2.2752e-05	7.5868e-05	0.0000e+00
model.layer.24	7.2360e-05	**2.2221e-04**	1.3542e-04	1.1861e-05	2.0402e-05	2.5562e-05	0.0000e+00
model.layer.25	5.7340e-05	**5.5480e-04**	1.3733e-04	5.7817e-05	3.3021e-05	8.5490e-06	0.0000e+00
model.layer.26	1.9974e-04	1.6785e-04	**3.0923e-04**	4.4942e-05	2.6158e-05	1.7166e-05	0.0000e+00
model.layer.27	3.3438e-05	**2.9206e-04**	1.8406e-04	1.0258e-04	5.7442e-05	7.7656e-06	0.0000e+00
model.layer.28	5.1141e-05	**1.2302e-04**	6.2704e-05	1.8299e-05	1.8171e-05	1.9244e-06	1.7030e-08
model.layer.29	5.7995e-05	**3.4642e-04**	5.9843e-05	2.3514e-04	1.3913e-05	4.7684e-05	0.0000e+00
model.layer.30	3.5679e-04	**4.2725e-04**	1.9693e-04	1.5831e-04	2.1390e-05	3.9986e-05	0.0000e+00
model.layer.31	1.6040e-04	**4.2367e-04**	2.3746e-04	1.7762e-05	1.0286e-05	3.9952e-05	1.7030e-08

表 6.12: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間 コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム の値 (太字は上位 15 層)

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	241	1.1049e-06	0.00658	0.00000	0.0003839
model.layers.1	**4028**	**1.8468e-05**	**0.03061**	**0.42488**	**0.67841**
model.layers.2	255	1.1691e-06	0.00656	0.00157	0.00000
model.layers.3	268	1.2287e-06	0.00668	0.00303	0.00330
model.layers.4	569	2.6088e-06	0.01000	0.03680	0.09696
model.layers.5	654	2.9985e-06	0.01045	0.04634	0.10962
model.layers.6	311	1.4259e-06	0.00730	0.00785	0.02086
model.layers.7	1033	4.7361e-06	0.01287	0.08886	0.17799
model.layers.8	1007	4.6169e-06	0.01276	0.08594	0.17491
model.layers.9	1617	7.4136e-06	0.01631	0.15438	0.27504
model.layers.10	2203	1.01e-05	0.01907	0.22013	0.35290
model.layers.11	2465	1.1302e-05	0.02085	0.24952	0.40295
model.layers.12	2531	1.1604e-05	0.02064	0.25693	0.39728
model.layers.13	**3843**	**1.7619e-05**	**0.02531**	**0.40413**	**0.52886**
model.layers.14	**7372**	**3.3799e-05**	**0.03532**	**0.80007**	**0.81138**
model.layers.15	**3872**	**1.7752e-05**	**0.02637**	**0.40738**	**0.55886**
model.layers.16	2680	1.2287e-05	0.02140	0.27365	0.41864
model.layers.17	1807	8.2847e-06	0.01771	0.17570	0.31463
model.layers.18	1199	5.4972e-06	0.01427	0.10748	0.21740
model.layers.19	3035	1.3915e-05	0.02374	0.31347	0.48459
model.layers.20	1213	5.5614e-06	0.01466	0.10905	0.22856
model.layers.21	1520	6.9689e-06	0.01640	0.14350	0.27766
model.layers.22	1589	7.2852e-06	0.01671	0.15124	0.28634
model.layers.23	1417	6.4967e-06	0.01576	0.13194	0.25935
model.layers.24	**4121**	**1.8894e-05**	**0.02631**	**0.43532**	**0.55724**
model.layers.25	**3468**	**1.59e-05**	**0.02480**	**0.36206**	**0.51466**
model.layers.26	2565	1.176e-05	0.02159	0.26074	0.42404
model.layers.27	**3740**	**1.7147e-05**	**0.02603**	**0.39257**	**0.54933**
model.layers.28	973	4.461e-06	0.01293	0.08213	0.17968
model.layers.29	**3227**	**1.4795e-05**	**0.02396**	**0.33502**	**0.49093**
model.layers.30	**4298**	**1.9705e-05**	**0.02846**	**0.45518**	**0.61794**
model.layers.31	**9154**	**4.1969e-05**	**0.04200**	**1.00000**	**1.00000**

表 6.13: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間の各パラメータのコンフリクト率の値 (太字は上位 15 層)

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	2.0266e-06	8.5831e-06	**2.5034e-05**	1.7285e-06	4.5981e-07	1.7030e-07	0.0000e+00
model.layer.1	8.4639e-06	1.8597e-05	9.2983e-06	1.7285e-06	3.2357e-07	2.5545e-07	**6.3113e-05**
model.layer.2	5.3644e-06	**8.8215e-06**	2.1458e-06	3.2187e-06	5.6199e-07	5.4496e-07	0.0000e+00
model.layer.3	2.7418e-06	**1.2875e-05**	5.9605e-06	2.7418e-06	9.5367e-07	6.9823e-07	0.0000e+00
model.layer.4	7.7486e-06	1.6928e-05	**2.6941e-05**	3.9339e-06	1.6689e-06	1.5497e-06	0.0000e+00
model.layer.5	4.5896e-06	7.1526e-06	8.5831e-06	**1.3411e-05**	3.0313e-06	1.8392e-06	0.0000e+00
model.layer.6	2.9802e-06	5.0068e-06	**8.8215e-06**	2.0266e-06	1.4986e-06	1.3794e-06	0.0000e+00
model.layer.7	1.9073e-06	**4.8876e-05**	6.6757e-06	1.0431e-05	1.9755e-06	8.1233e-06	0.0000e+00
model.layer.8	6.0797e-06	**3.2425e-05**	1.3351e-05	7.9274e-06	6.5225e-06	3.3549e-06	0.0000e+00
model.layer.9	2.9802e-06	3.1471e-05	1.3828e-05	**3.2723e-05**	8.5149e-06	5.5858e-06	0.0000e+00
model.layer.10	1.2517e-05	**3.9101e-05**	1.2636e-05	1.5855e-05	1.0559e-05	1.5157e-05	0.0000e+00
model.layer.11	1.1504e-05	**5.8413e-05**	1.7643e-05	4.3929e-05	7.3399e-06	1.3368e-05	0.0000e+00
model.layer.12	1.1265e-05	**5.8651e-05**	3.3855e-05	2.4438e-05	1.0524e-05	1.5770e-05	0.0000e+00
model.layer.13	1.8477e-05	**1.2875e-04**	2.7418e-05	2.3842e-05	1.4016e-05	2.8184e-05	0.0000e+00
model.layer.14	9.2983e-06	7.8201e-05	1.3590e-05	**1.1188e-04**	1.6434e-05	6.7932e-05	0.0000e+00
model.layer.15	3.0756e-05	**1.7023e-04**	3.6240e-05	6.7949e-06	1.7370e-05	2.3093e-05	0.0000e+00
model.layer.16	1.1444e-05	**5.5313e-05**	6.4373e-06	6.7353e-06	1.1138e-05	2.4898e-05	0.0000e+00
model.layer.17	8.2850e-06	**1.9455e-04**	5.9605e-06	7.4506e-06	8.2765e-06	3.6614e-06	1.7030e-08
model.layer.18	1.3053e-05	**5.3883e-05**	1.3113e-05	3.3975e-06	4.1383e-06	6.7949e-06	0.0000e+00
model.layer.19	3.3379e-05	**3.2187e-04**	6.3419e-05	1.3530e-05	6.0626e-06	4.7003e-06	0.0000e+00
model.layer.20	1.0133e-05	**5.1260e-05**	**5.1260e-05**	1.4007e-05	3.7977e-06	2.6396e-06	0.0000e+00
model.layer.21	1.1504e-05	**1.3804e-04**	1.2875e-05	1.2577e-05	5.1941e-06	3.0313e-06	0.0000e+00
model.layer.22	1.1325e-05	**9.2506e-05**	1.0729e-05	9.7752e-06	1.8052e-06	1.1853e-05	0.0000e+00
model.layer.23	4.2915e-06	**9.4414e-05**	1.1683e-05	1.0133e-05	4.7513e-06	7.6805e-06	0.0000e+00
model.layer.24	1.1563e-05	6.7949e-05	4.9353e-05	**1.8668e-04**	1.4986e-06	3.6614e-06	0.0000e+00
model.layer.25	5.4002e-05	**1.9073e-04**	6.9380e-05	4.0472e-05	8.5149e-06	4.9727e-06	0.0000e+00
model.layer.26	4.0829e-05	**1.6975e-04**	5.7220e-05	3.1054e-05	2.4693e-06	4.4618e-06	0.0000e+00
model.layer.27	4.6909e-05	**1.6212e-04**	7.7724e-05	9.8228e-05	3.6274e-06	1.4646e-06	0.0000e+00
model.layer.28	1.1742e-05	**4.1485e-05**	6.1989e-06	1.1861e-05	2.5375e-06	3.8828e-06	0.0000e+00
model.layer.29	3.6955e-05	1.4114e-04	**1.7738e-04**	2.1756e-05	3.4400e-06	1.1972e-05	1.7030e-08
model.layer.30	6.2943e-05	**1.7667e-04**	8.3447e-05	4.2319e-05	1.7268e-05	7.1696e-06	1.0218e-07
model.layer.31	5.6088e-05	**4.3440e-04**	1.4496e-04	1.4615e-04	1.6093e-05	2.9274e-05	1.1359e-05

表 6.14: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D1}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間 コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム の値 (太字は上位 15 層)

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	63	2.8884e-07	0.00347	0.00131	0.01196
model.layers.1	565	2.5904e-06	0.01059	0.07459	0.22439
model.layers.2	82	3.7595e-07	0.00384	0.00409	0.02291
model.layers.3	54	2.4758e-07	0.00307	0.00000	0.00000
model.layers.4	134	6.1436e-07	0.00480	0.01168	0.05172
model.layers.5	258	1.1829e-06	0.00677	0.02978	0.11028
model.layers.6	155	7.1064e-07	0.00530	0.01474	0.06668
model.layers.7	397	1.8202e-06	0.00834	0.05007	0.15713
model.layers.8	375	1.7193e-06	0.00798	0.04685	0.14656
model.layers.9	627	2.8747e-06	0.01057	0.08364	0.22388
model.layers.10	807	3.6999e-06	0.01194	0.10991	0.26471
model.layers.11	1075	4.9287e-06	0.01399	0.14903	0.32566
model.layers.12	1157	5.3046e-06	0.01453	0.16100	0.34200
model.layers.13	1648	7.5558e-06	0.01687	0.23267	0.41158
model.layers.14	**3057**	**1.4016e-05**	**0.02311**	**0.43833**	**0.59783**
model.layers.15	1469	6.7351e-06	0.01697	0.20654	0.41473
model.layers.16	1200	5.5018e-06	0.01430	0.16727	0.33497
model.layers.17	925	4.2409e-06	0.01258	0.12713	0.28384
model.layers.18	1688	7.7391e-06	0.01745	0.23851	0.42884
model.layers.19	2273	1.0421e-05	**0.02052**	0.32389	**0.52053**
model.layers.20	1344	6.162e-06	0.01546	0.18829	0.36968
model.layers.21	2006	9.1971e-06	0.01843	0.28492	0.45832
model.layers.22	**2286**	**1.0481e-05**	0.02014	**0.32579**	0.50916
model.layers.23	**3474**	**1.5928e-05**	**0.02483**	**0.49920**	**0.64909**
model.layers.24	**3169**	**1.4529e-05**	**0.02394**	**0.45468**	**0.62270**
model.layers.25	**2389**	**1.0953e-05**	**0.02061**	**0.34083**	**0.52313**
model.layers.26	**2813**	**1.2897e-05**	**0.02241**	**0.40271**	**0.57690**
model.layers.27	**4146**	**1.9009e-05**	**0.02776**	**0.59729**	**0.73652**
model.layers.28	2094	9.6006e-06	0.01972	0.29777	0.49658
model.layers.29	**3482**	**1.5964e-05**	**0.02541**	**0.50036**	**0.66636**
model.layers.30	**6905**	**3.1658e-05**	**0.03634**	**1.00000**	**0.99239**
model.layers.31	**6803**	**3.119e-05**	**0.03659**	**0.98511**	**1.00000**

表 6.15: 実験 1 における $M_{\mathrm{DPO}_{\mathrm{D1}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の各パラメータのコンフリクト率の値 (太字は上位 15 層)

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	1.3113e-06	3.3379e-06	**3.8147e-06**	1.1921e-07	6.8120e-08	8.5149e-08	0.0000e+00
model.layer.1	6.4969e-06	1.4305e-06	2.1458e-06	4.1723e-07	6.8120e-08	6.8120e-08	**7.2547e-06**
model.layer.2	1.1325e-06	**5.2452e-06**	1.4305e-06	2.3842e-07	2.3842e-07	2.8951e-07	0.0000e+00
model.layer.3	3.5763e-07	**2.8610e-06**	2.6226e-06	3.5763e-07	2.0436e-07	1.1921e-07	0.0000e+00
model.layer.4	1.2517e-06	**9.0599e-06**	9.5367e-07	6.5565e-07	4.9387e-07	5.2793e-07	0.0000e+00
model.layer.5	2.3246e-06	2.3842e-06	5.4836e-06	**6.4373e-06**	9.8773e-07	3.4060e-07	0.0000e+00
model.layer.6	1.3113e-06	2.6226e-06	9.5367e-07	**2.6822e-06**	7.8338e-07	4.5981e-07	0.0000e+00
model.layer.7	1.0729e-06	**1.5974e-05**	3.8147e-06	6.3181e-06	1.1921e-06	2.0436e-06	0.0000e+00
model.layer.8	5.3644e-07	**1.1921e-05**	1.1921e-06	3.5763e-06	2.7248e-06	1.5497e-06	0.0000e+00
model.layer.9	5.9605e-07	**1.0729e-05**	8.5831e-06	8.0466e-06	3.2187e-06	3.6103e-06	0.0000e+00
model.layer.10	5.2452e-06	**2.0742e-05**	2.8610e-06	3.3379e-06	4.5129e-06	5.0919e-06	0.0000e+00
model.layer.11	8.8811e-06	**1.5974e-05**	8.5831e-06	1.4603e-05	3.5252e-06	6.3181e-06	0.0000e+00
model.layer.12	6.3777e-06	**4.5776e-05**	1.2636e-05	1.0490e-05	4.2064e-06	6.5054e-06	0.0000e+00
model.layer.13	4.9472e-06	**2.1935e-05**	3.3379e-06	5.3048e-06	8.2595e-06	1.5071e-05	0.0000e+00
model.layer.14	3.5763e-06	1.3590e-05	5.7220e-06	**6.0797e-05**	9.7411e-06	2.2548e-05	0.0000e+00
model.layer.15	1.5974e-05	**3.6955e-05**	7.3910e-06	2.6226e-06	7.6975e-06	8.8385e-06	0.0000e+00
model.layer.16	4.1127e-06	1.2159e-05	5.4836e-06	3.9935e-06	3.6785e-06	**1.3181e-05**	0.0000e+00
model.layer.17	8.8215e-06	**4.9829e-05**	5.7220e-06	8.6427e-06	3.3038e-06	3.4911e-06	0.0000e+00
model.layer.18	2.1458e-06	2.9564e-05	**3.8147e-05**	1.1742e-05	9.7752e-06	1.0167e-05	0.0000e+00
model.layer.19	1.8477e-05	**1.6308e-04**	2.3603e-05	1.1086e-05	1.4135e-05	2.7929e-06	0.0000e+00
model.layer.20	8.4639e-06	**1.0848e-04**	7.6294e-06	3.9339e-06	3.0654e-06	7.9870e-06	0.0000e+00
model.layer.21	5.5432e-06	**7.6771e-05**	1.6451e-05	1.7285e-05	1.6791e-05	4.1894e-06	0.0000e+00
model.layer.22	6.3181e-06	**7.4863e-05**	3.4332e-05	7.7486e-06	3.2016e-06	2.3910e-05	0.0000e+00
model.layer.23	1.2934e-05	**8.4639e-05**	3.9339e-05	2.8014e-05	1.5497e-05	2.3110e-05	0.0000e+00
model.layer.24	2.0742e-05	**1.0204e-04**	9.1076e-05	6.2525e-05	5.7220e-06	1.0661e-05	0.0000e+00
model.layer.25	2.0981e-05	**8.3447e-05**	4.7922e-05	1.8179e-05	1.5361e-05	4.7513e-06	0.0000e+00
model.layer.26	2.4319e-05	**1.8406e-04**	7.5102e-05	1.7107e-05	3.4400e-06	1.4118e-05	0.0000e+00
model.layer.27	3.7253e-05	**2.6512e-04**	3.0994e-05	4.4942e-05	2.3620e-05	2.3501e-06	0.0000e+00
model.layer.28	2.0504e-05	**1.5807e-04**	1.5974e-05	1.9729e-05	9.5197e-06	2.2139e-06	0.0000e+00
model.layer.29	5.8532e-05	**9.6560e-05**	9.2030e-05	5.0306e-05	3.2016e-06	1.1512e-05	1.7030e-08
model.layer.30	3.1233e-05	**1.6785e-04**	5.0068e-05	1.2958e-04	3.0756e-05	2.5238e-05	8.5149e-08
model.layer.31	9.7156e-05	**2.5225e-04**	1.1063e-04	7.7605e-05	1.2193e-05	2.5851e-05	1.9584e-06

表 6.12 では 29 から 31 層の深い層に加えて 13 から 14 層の中間層において Conflict Limited L2 ノルムの値が大きくなっている一方で, 表 6.14 では 22 から 31 層まで局所的に Conflict Limited L2 ノルムの値が大きくなっている. これはデータセットの特徴の差と, DPO と SFT の重みの変化量の差によるものと考えられる. D_0 は一般的な応答を収録しており出力のスタイルはベースモデルと大きく変わらない. このため DPO では表 6.3 で見られたように一般的な語彙などの情報をより学習するため中間層の重みの変化が集中的に大きくなっているが, DPO よりも SFT より重みの変化量が大きい手法であるため中間的な層に加えて最終的な出力に関わる深い層もバランスよく重みを変化させている. そのため中間層と最も深い層でコンフリクトが大きくなっていると考えられる. 一方で D_1 はお嬢様の応答を収録しているため出力のスタイルはベースモデルと大きく変わる. このため文脈を踏まえた学習が必要であるため, DPO でも表 6.3 で見られたように 22 から 31 層までの層の重みの変化が大きくなっており, コンフリクトが大きくなったと考えられる.

表 6.16: 実験 1 における ($M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, M_{base}), ($M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D1}}}$), ($M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{DPO_{D0}}}$) 間の L1 ノルム, L2 ノルム

モデル	L1 ノルム	L2 ノルム
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D1}}}$	1.66718×10^5	0.58686
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, M_{base}	3.12721×10^5	2.75796
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{DPO_{D0}}}$	3.45665×10^5	3.32855
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D0}}}$	3.27395×10^5	3.05515

表 6.16 には, 実験 1 における ($M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, M_{base}), ($M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D1}}}$), ($M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{DPO_{D0}}}$) 間の L1 ノルム, L2 ノルム を示している. 一度 D_1 で SFT を適用し, そこから更に D_0 を chosen とした DPO を適用したモデル $M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$ に関して, ベースモデルへの回帰現象は見られず, D_0 でファインチューニングしたモデル $M_{\mathrm{DPO_{D0}}}$, $M_{\mathrm{SFT_{D0}}}$ にパラメータが近づくような現象は起こらなかった. このことから似たスタイルの応答をする LLM であっても重みは似た分布をするといったわけではないことがわかる. これは LLM が膨大なパラメータを保持し冗長性があるため重みの分布は学習過程によって重みの分布が異なっても似たようなタスクを実現できる性能があるといえる.

学習後のモデルの出力結果から LLM の性能を評価する. 表 6.17, 6.18 にそれぞれ実験 1 における学習後のモデルの D_0, D_1 との BLEU, BERTScore の値を示す. 全体

的にかなり BLEU の値が小さくなっている. 本研究のようなロールプレイタスクにおけるファインチューニングでは学習データと完全に一致するような文は出力されず語尾や出力のスタイルのみが反映された結果このような結果となっていると考えられる. また, 文の意味的類似度を測る BERTScore においてはベースモデルから SFT, DPO を適用したモデルがそれぞれのデータセットにおいて類似度が高くなっている.

表 6.17: 実験 1 における各モデルの生成結果とデータセット D_0 との BLEU, BERTScore

モデル	BLEU (D_0 train)	BLEU (D_0 test)	BERTScore (D_0 train)	BERTScore (D_0 test)
M_{base}	0.01320	0.01777	0.74986	0.74261
$M_{\text{SFT}_{\text{D0}}}$	**0.05969**	0.04509	**0.76087**	0.74243
$M_{\text{SFT}_{\text{D1}}}$	0.02933	0.02559	0.74803	0.73257
$M_{\text{DPO}_{\text{D0}}}$	0.02059	0.01595	0.75637	**0.74397**
$M_{\text{DPO}_{\text{D1}}}$	0.00388	0.00697	0.73052	0.71447
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.05811	**0.04643**	0.75684	0.73387
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.0	0.0	0.70501	0.68910
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	0.02708	0.03039	0.75352	0.74052
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	0.00386	0.00999	0.71788	0.70418

表 6.18: 実験 1 における各モデルの生成結果とデータセット D_1 との BLEU, BERTScore

モデル	BLEU (D_1 train)	BLEU (D_1 test)	BERTScore (D_1 train)	BERTScore (D_1 test)
M_{base}	0.06843	0.03588	0.74715	0.73788
$M_{\text{SFT}_{\text{D0}}}$	0.03146	0.04271	0.74864	0.73704
$M_{\text{SFT}_{\text{D1}}}$	**0.06843**	0.03588	**0.77558**	**0.75122**
$M_{\text{DPO}_{\text{D0}}}$	0.01197	0.01353	0.74928	0.73538
$M_{\text{DPO}_{\text{D1}}}$	0.02776	0.03289	0.76129	0.73996
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.03060	0.03334	0.74403	0.72435
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.00968	0.01290	0.74823	0.72593
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	0.00386	0.00999	0.71788	0.70418
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	0.02790	**0.04605**	0.75967	0.73866

表 6.19 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は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はベースモデルより上回る数値となった. 評価者 LLM はプロンプトで出力結果の口調を評価するため, 学習により D_1 の応答スタイルを踏まえたテキスト生成ができていると考えられる.

また, LLM が最もロールプレイで最も高いスコアをつけたのは $M_{\mathrm{DPO}_{\mathrm{D1}}}$ であった. DPO では SFT と異なり, 好ましくない応答を rejected として学習するためロールプレイの精度が高くなったと考えられる. ここで, SFT で学習後さらに同じデータセットで D_1 を chosen としたモデル $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$ に関しては LLM のスコアは $M_{\mathrm{DPO}_{\mathrm{D1}}}$ より大きくならず, SFT と DPO を組み合わせることによるロールプレイ性能のさらなる向上は本データセットでは定量的には示すことはできなかった. また, 表 6.20 に

表 6.19: 実験 1 における評価者 LLM による各モデルの生成結果の評価

モデル	train データのクエリ (162 件)	test データのクエリ (40 件)
dataset	7.79	8.1
M_{base}	6.72	6.23
$M_{\mathrm{SFT}_{\mathrm{D0}}}$	5.88	5.825
$M_{\mathrm{SFT}_{\mathrm{D1}}}$	8.20	8.45
$M_{\mathrm{DPO}_{\mathrm{D0}}}$	5.87	6.025
$M_{\mathrm{DPO}_{\mathrm{D1}}}$	**8.40**	**8.85**
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	4.81	4.75
$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	7.02	7.925
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	5.39	5.475
$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	7.67	7.78

train データに含まれている「このカフェ素敵ですね」に対する LLM の応答を示す.

D_0, D_1 で共通して登場する「雰囲気」という単語が頻出していることがわかる. M_{base} にも「雰囲気」という単語が出てくるためファインチューニングとベースモデル両方の影響を受けてこの現象が起こっていると考えられる.

また, D_1 を用いてファインチューニングしたモデルでは一人称が「わたくし」となっていたり, 語尾が「ですわね。」など D_1 内に見られるお嬢様スタイルの応答が実現されていることがわかる.

6.2 実験 2

(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 6.21 に結果を示す.

表 6.20: 「このカフェ素敵ですね」に対する LLM の応答

モデル	「このカフェ素敵ですね」に対する LLM の応答
D_0	私もこのお店が気に入っています。静かな雰囲気がいいですね
D_1	はい私もこちらのお店が大好きですわ。落ち着いた雰囲気がいいですわね
M_{base}	ありがとうございます！このカフェは、内装もおしゃれで、雰囲気もとても良く友達と来るにはぴったりなお店です。
$M_{\text{SFT}_{\text{D0}}}$	ありがとうございます。内装が気に入っています。雰囲気がいいですね。
$M_{\text{SFT}_{\text{D1}}}$	わたくしも大好きですの。雰囲気がとてもいいですわね。
$M_{\text{DPO}_{\text{D0}}}$	ありがとうございます、店内の雰囲気と美味しいコーヒーがとても気に入っています。
$M_{\text{DPO}_{\text{D1}}}$	まあ！こちらのカフェはとってもお洒落で可愛いわ！
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	店内のインテリアが気に入っています。雰囲気もいいです。
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	わたくしもそう思いますわ、お母様とよく来るのよ。
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	入り口のガラス張りがとても綺麗で、自然光が差し込むので明るい雰囲気ですよね。
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	わたくしもそう思いますわ、こちらのお店のインテリアはとてもお気に入りですわ♪

表 6.21 から, 同じ学習率においても SFT と DPO ではベースモデルからの重みの変化の大きさは SFT の方が大きくなっており, 実験 1 と同様の結果が得られた.

また, 表 6.22, 6.23 に, それぞれ SFT を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを, 全層について最大値を 1, 最小値を 0 として正規化した値, DPO を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値を示す.

表 6.2, 表 6.3 ともに 0, 1, 2 層目の浅い層が他の層と比較して大きく変動していることがわかった. これはベースモデルである Qwen/Qwen2.5-7B-Instruct が多言語向けに instruction tuning されたモデルであるためと考えられる. 実験 1 で用いた elyza/Llama-3-ELYZA-JP-8B が日本語タスクに特化している一方で Qwen/Qwen2.5-7B-Instruct は中国語, 英語を中心に日本語も対応しているが, 日本語に特化しているわけではない. そのため, 今回の実験で用いた日本語のデータセットでファインチューニングする際に日本語の品詞のタグ付けの情報などが浅い層の重みで保持されており結果として L2 ノルムが大きくなったと考えられる.

0, 1, 2 層目と比較すると値は小さいが, 19 層以降の層も他の層と比較して L2 ノルムが大きくなっており, ロールプレイを実現するための高度な言語情報は深い層で保持されることがわかる.

表 6.24 に $M_{\text{SFT}_{\text{D0}}}$, $M_{\text{SFT}_{\text{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.25 に各層の各パラメータのコンフリクト率の値を示す.

表 6.21: 実験 2 における (学習後のモデル, 学習前のモデル) の関係にある 2 モデル間の L1 ノルム, L2 ノルム

モデル組	L1 ノルム	L2 ノルム
$M_{\mathrm{SFT_{D0}}}$, M_{base}	2.65476 $\times 10^5$	4.66679
$M_{\mathrm{SFT_{D1}}}$, M_{base}	2.52975 $\times 10^5$	4.42202
$M_{\mathrm{SFT_{D2}}}$, M_{base}	2.53258 $\times 10^5$	4.02821
$M_{\mathrm{DPO_{D0}}}$, M_{base}	1.37400 $\times 10^5$	3.98816
$M_{\mathrm{DPO_{D1}}}$, M_{base}	1.41034 $\times 10^5$	3.49920
$M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$, M_{base}	3.29224 $\times 10^5$	6.63510
$M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$, M_{base}	3.23261 $\times 10^5$	5.80873
$M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$, $M_{\mathrm{SFT_{D2}}}$	1.61470 $\times 10^5$	4.64364
$M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$, $M_{\mathrm{SFT_{D2}}}$	1.53017 $\times 10^5$	3.36994
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, M_{base}	3.29579 $\times 10^5$	6.86016
$M_{\mathrm{DPO_{D1}(M_{SFT_{D1}})}}$, M_{base}	3.26070 $\times 10^5$	6.32263
$M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D1}}}$	1.63094 $\times 10^5$	4.61360
$M_{\mathrm{DPO_{D1}(M_{SFT_{D1}})}}$, $M_{\mathrm{SFT_{D1}}}$	1.57729 $\times 10^5$	3.75747

同様に, 表 6.26 に $M_{\mathrm{DPO_{D0}}}$, $M_{\mathrm{DPO_{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.27 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.28 に $M_{\mathrm{DPO_{D0}(M_{SFT_{D2}})}}$, $M_{\mathrm{DPO_{D1}(M_{SFT_{D2}})}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.29 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.30 に $M_{\mathrm{DPO_{D0}(M_{SFT_{D1}})}}$, $M_{\mathrm{DPO_{D1}(M_{SFT_{D1}})}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.29 に各層の各パラメータのコンフリクト率の値を示す.

表 6.24 , 表 6.26, 6.28, 表 6.30 に共通して 0, 1, 2 層目, 特に 1, 2 層目のコンフリクト数が非常に多いことが挙げられる. データセットの違いによるコンフリクトが浅い層で発生しているということは, ベースモデルが Qwen/Qwen2.5-7B-Instruct の場合に

表 6.22: 実験 2 における SFT を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値 (太字は上位 15 層)

	$M_{\mathrm{SFT}_{\mathrm{D0}}}$	$M_{\mathrm{SFT}_{\mathrm{D1}}}$	$M_{\mathrm{SFT}_{\mathrm{D2}}}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$	$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$
model.layers.0	**0.12979**	**0.13813**	**0.17924**	**0.08147**	**0.11990**	**0.08676**	**0.10519**
model.layers.1	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**
model.layers.2	**0.49916**	**0.44634**	**0.72400**	**0.51644**	**0.71301**	**0.57045**	**0.61946**
model.layers.3	0.02176	0.01351	0.02324	0.00248	0.00198	0.00124	0.00000
model.layers.4	0.01128	0.01871	0.02136	0.00306	0.00409	0.00494	0.00410
model.layers.5	0.00000	0.00945	0.00873	0.00000	0.00000	0.00303	0.00144
model.layers.6	0.00443	0.00195	0.00468	0.00501	0.00611	0.00000	0.00078
model.layers.7	0.00134	0.00216	0.00000	0.00590	0.00826	0.00523	0.00671
model.layers.8	0.00462	0.00000	0.00150	0.01007	0.01339	0.00326	0.00500
model.layers.9	0.00914	0.00014	0.00521	0.01438	0.01813	0.00416	0.00604
model.layers.10	0.01020	0.00040	0.00836	0.01550	0.01941	0.00354	0.00570
model.layers.11	0.02071	0.00895	0.01936	0.02314	0.03071	0.01305	0.01509
model.layers.12	0.02152	0.01504	0.02275	0.02245	0.03195	0.01273	0.02108
model.layers.13	0.03209	0.01958	0.03518	0.02463	0.03331	0.01648	0.02317
model.layers.14	**0.04987**	0.03021	0.05678	0.03482	0.04424	0.01692	0.02916
model.layers.15	**0.04128**	0.02775	0.05269	0.03409	0.04067	0.01476	0.02822
model.layers.16	**0.06866**	**0.05172**	**0.08244**	**0.04486**	**0.05408**	**0.02828**	**0.04048**
model.layers.17	**0.06239**	**0.04971**	**0.07889**	**0.04540**	**0.05656**	**0.03040**	**0.04491**
model.layers.18	**0.07762**	**0.06591**	**0.10389**	**0.05490**	**0.06844**	**0.04356**	**0.05595**
model.layers.19	**0.06293**	**0.06422**	**0.09583**	**0.04623**	**0.05958**	**0.04108**	**0.05000**
model.layers.20	**0.04874**	**0.06560**	**0.08970**	**0.04516**	**0.05841**	**0.04658**	**0.05548**
model.layers.21	0.02717	**0.06421**	**0.08040**	**0.04247**	**0.05273**	**0.04545**	**0.05134**
model.layers.22	**0.05458**	**0.07995**	**0.11635**	**0.05519**	**0.07050**	**0.05260**	**0.05827**
model.layers.23	**0.04291**	**0.06718**	**0.09857**	**0.04751**	**0.06452**	**0.05114**	**0.05585**
model.layers.24	**0.05192**	**0.08156**	**0.12228**	**0.04953**	**0.08110**	**0.05270**	**0.06500**
model.layers.25	**0.04422**	**0.08644**	**0.11191**	**0.05046**	**0.07554**	**0.05733**	**0.07279**
model.layers.26	0.03453	**0.08321**	**0.10697**	**0.05378**	**0.06612**	**0.06144**	**0.06439**
model.layers.27	**0.03704**	**0.09865**	**0.11060**	**0.04736**	**0.08462**	**0.09345**	**0.07258**

表 6.23: 実験 2 における DPO を適用後モデルと DPO を適用前モデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値 (太字は上位 15 層)

	$M_{\mathrm{DPO}_{\mathrm{D0}}}$	$M_{\mathrm{DPO}_{\mathrm{D1}}}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$	$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$
model.layers.0	**0.06874**	**0.08858**	**0.04723**	**0.09084**	**0.05956**	**0.09586**
model.layers.1	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**
model.layers.2	**0.68511**	**0.94277**	**0.46698**	**0.73145**	**0.83402**	**0.82160**
model.layers.3	**0.00946**	0.00899	0.00371	0.00619	0.00000	0.00774
model.layers.4	**0.01152**	**0.04689**	0.00539	0.00715	0.01188	0.01048
model.layers.5	0.00685	0.00755	0.00240	0.00440	0.00348	0.00495
model.layers.6	0.00436	0.00320	0.00390	0.00743	0.00622	0.00414
model.layers.7	0.00194	0.00000	0.00000	0.00000	0.01009	0.00287
model.layers.8	0.00000	0.00050	0.00380	0.00568	0.00763	0.00000
model.layers.9	0.00672	0.01059	0.01225	**0.01990**	0.01514	0.00958
model.layers.10	0.00230	0.00744	0.01019	0.01639	0.01390	0.00513
model.layers.11	0.00495	0.01542	0.01279	**0.02351**	**0.02301**	0.01035
model.layers.12	0.00375	0.00890	0.01141	**0.02462**	0.01402	0.01598
model.layers.13	0.00231	0.01306	0.01155	**0.02307**	0.01515	**0.02109**
model.layers.14	0.00270	0.01045	0.01304	0.01455	0.01295	0.01616
model.layers.15	0.00615	**0.01714**	**0.01623**	**0.01826**	**0.01868**	**0.02168**
model.layers.16	0.00134	0.01165	**0.01691**	0.01402	0.00916	0.01739
model.layers.17	0.00384	**0.01747**	**0.01578**	0.01612	0.01697	0.02030
model.layers.18	**0.00911**	0.01461	**0.01760**	0.01263	**0.02036**	**0.02722**
model.layers.19	**0.01104**	**0.02564**	**0.01349**	0.01799	**0.02614**	**0.02423**
model.layers.20	**0.00973**	**0.02091**	**0.01587**	**0.02370**	**0.02374**	**0.03551**
model.layers.21	**0.01223**	**0.02571**	**0.01515**	0.01576	**0.03255**	**0.02212**
model.layers.22	**0.01332**	**0.02767**	**0.01571**	**0.02086**	**0.03731**	**0.02420**
model.layers.23	**0.02465**	**0.04830**	**0.01317**	**0.02391**	**0.05536**	**0.03296**
model.layers.24	**0.01391**	**0.03172**	0.00131	**0.03478**	**0.04266**	**0.03706**
model.layers.25	**0.01758**	**0.04480**	**0.01511**	**0.03813**	**0.04313**	**0.05233**
model.layers.26	**0.02047**	**0.03401**	**0.03254**	**0.02455**	**0.04248**	**0.03479**
model.layers.27	**0.04647**	**0.02751**	**0.02571**	**0.10406**	**0.04443**	**0.06456**

表 6.24: 実験 2 における $M_{\mathrm{SFT}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	**914789**	**0.00393**	**0.51633**	**0.08453**	**0.14013**
model.layers.1	**10715964**	**0.04598**	**3.42104**	**1.00000**	**1.00000**
model.layers.2	**4005148**	**0.01719**	**1.70914**	**0.37318**	**0.49323**
model.layers.3	**50394**	**0.00021624**	**0.10364**	**0.00379**	**0.01797**
model.layers.4	**50375**	**0.00021616**	**0.12146**	**0.00379**	**0.02324**
model.layers.5	34303	0.00014719	0.08097	0.00229	0.01126
model.layers.6	17830	7.6508e-05	0.05758	0.00075321	0.00433
model.layers.7	14345	6.1554e-05	0.05209	0.0004277	0.00271
model.layers.8	10275	4.409e-05	0.04361	4.7543e-05	0.00019814
model.layers.9	13697	5.8774e-05	0.05065	0.00036717	0.00228
model.layers.10	11167	4.7918e-05	0.04569	0.00013086	0.00081321
model.layers.11	9766	4.1906e-05	0.04294	0.00000	0.00000
model.layers.12	10128	4.3459e-05	0.04297	3.3812e-05	8.4202e-06
model.layers.13	15683	6.7296e-05	0.05355	0.00055267	0.00314
model.layers.14	26421	0.00011337	0.07031	0.00156	0.00810
model.layers.15	22067	9.4689e-05	0.06379	0.00115	0.00617
model.layers.16	46790	0.00020078	0.09416	0.00346	0.01516
model.layers.17	41335	0.00017737	0.08897	0.00295	0.01362
model.layers.18	**57543**	**0.00024692**	**0.10510**	**0.00446**	**0.01840**
model.layers.19	**92875**	**0.00039853**	**0.13497**	**0.00776**	**0.02724**
model.layers.20	**101330**	**0.00043481**	**0.14012**	**0.00855**	**0.02877**
model.layers.21	**99965**	**0.00042895**	**0.13908**	**0.00842**	**0.02846**
model.layers.22	**206817**	**0.00088745**	**0.20598**	**0.01841**	**0.04826**
model.layers.23	**145366**	**0.00062377**	**0.16941**	**0.01267**	**0.03744**
model.layers.24	**193949**	**0.00083223**	**0.19713**	**0.01720**	**0.04564**
model.layers.25	**205414**	**0.00088143**	**0.20464**	**0.01827**	**0.04787**
model.layers.26	**157304**	**0.00067499**	**0.17830**	**0.01378**	**0.04007**
model.layers.27	**232881**	**0.00099929**	**0.22366**	**0.02084**	**0.05350**

表 6.25: 実験 2 における $M_{\mathrm{SFT}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の各パラメータのコンフリクト率の値

パラメータ	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	2.6391e-05	5.5041e-05	1.9618e-05	6.3371e-05	**1.3320e-02**	1.3223e-04	0.0000e+00
model.layer.1	9.0774e-05	6.3215e-05	5.7220e-05	7.3569e-05	**9.4270e-02**	6.3530e-02	0.0000e+00
model.layer.2	9.5757e-06	1.7984e-05	5.4496e-06	9.8481e-05	**3.8700e-02**	2.0270e-02	0.0000e+00
model.layer.3	1.7750e-05	2.0599e-04	1.6349e-05	4.1183e-05	2.3723e-04	**4.8784e-04**	0.0000e+00
model.layer.4	1.3857e-05	1.6894e-04	4.6866e-05	5.7532e-05	2.7974e-04	**4.4287e-04**	0.0000e+00
model.layer.5	3.8458e-05	1.1063e-04	1.1444e-05	1.9245e-04	**3.3977e-04**	1.1848e-04	0.0000e+00
model.layer.6	4.5387e-05	8.8283e-05	1.0354e-05	4.2818e-05	1.0628e-04	**1.3698e-04**	0.0000e+00
model.layer.7	2.7481e-05	**1.1717e-04**	1.3624e-05	8.2288e-05	7.7030e-05	1.0995e-04	0.0000e+00
model.layer.8	2.3277e-05	4.5776e-05	2.7248e-05	**1.1351e-04**	5.7500e-05	6.5984e-05	0.0000e+00
model.layer.9	3.6279e-05	**9.4822e-05**	7.6294e-06	8.7894e-05	8.1199e-05	9.4219e-05	5.8914e-08
model.layer.10	3.2697e-05	5.9945e-05	1.0899e-05	4.9980e-05	**8.5175e-05**	6.1742e-05	0.0000e+00
model.layer.11	1.8451e-05	**1.7205e-04**	3.7602e-05	1.7984e-05	4.7293e-05	5.9003e-05	0.0000e+00
model.layer.12	1.3235e-05	**1.8575e-04**	1.2534e-05	8.2833e-05	4.8884e-05	5.9371e-05	6.9224e-07
model.layer.13	2.8571e-05	**3.2814e-04**	2.1253e-05	8.8828e-05	8.3437e-05	7.6942e-05	1.4729e-07
model.layer.14	1.2923e-04	**4.4196e-04**	5.1771e-05	2.1580e-04	1.3728e-04	1.3625e-04	3.0930e-07
model.layer.15	2.3200e-05	**2.1362e-04**	7.2479e-05	4.4141e-05	1.6151e-04	1.1525e-04	2.9457e-07
model.layer.16	3.9237e-05	**4.6967e-04**	1.1989e-05	1.0409e-04	1.9409e-04	3.9163e-04	4.0062e-06
model.layer.17	8.8517e-05	**6.3526e-04**	3.3787e-05	1.8529e-04	2.1826e-04	2.4717e-04	5.1550e-07
model.layer.18	1.2363e-04	3.4932e-04	1.0354e-05	2.5154e-04	2.9666e-04	**4.6610e-04**	4.0651e-06
model.layer.19	2.1868e-04	5.8528e-04	6.6485e-05	**1.6300e-03**	4.3913e-04	5.5344e-04	7.2906e-06
model.layer.20	6.5239e-05	2.0981e-04	6.5940e-05	4.7248e-04	**7.8729e-04**	5.9399e-04	1.9884e-06
model.layer.21	2.4975e-04	**7.5586e-04**	2.7793e-05	5.0813e-04	6.7958e-04	6.2778e-04	4.2713e-07
model.layer.22	2.8883e-04	9.6512e-04	2.5123e-04	**2.1600e-03**	1.5800e-03	9.5098e-04	2.0399e-05
model.layer.23	3.9758e-04	9.8419e-04	3.5967e-04	**2.4900e-03**	4.8808e-04	1.0600e-03	1.1120e-05
model.layer.24	1.6100e-03	1.0800e-03	1.2300e-03	9.7376e-04	6.7067e-04	**1.6300e-03**	1.8853e-06
model.layer.25	9.8240e-04	**1.9000e-03**	5.0681e-04	1.7400e-03	8.0338e-04	1.6200e-03	2.6511e-07
model.layer.26	4.9015e-04	**2.6000e-03**	3.5095e-04	2.1700e-03	7.5370e-04	9.0871e-04	9.5736e-07
model.layer.27	1.0000e-03	**2.1700e-03**	5.2316e-04	1.1300e-03	1.1400e-03	1.8100e-03	9.6619e-06

表 6.26: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{DPO}_{\mathrm{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	**523224**	**0.0022452**	**0.33015**	**0.03848**	**0.07641**
model.layers.1	**13234595**	**0.05679**	**3.69124**	**1.00000**	**1.00000**
model.layers.2	**5573947**	**0.02392**	**3.28866**	**0.42053**	**0.88938**
model.layers.3	28588	0.00012267	0.07322	0.001069	0.005815
model.layers.4	**33304**	**0.00014291**	**0.15721**	**0.0014257**	**0.02889**
model.layers.5	27509	0.00011804	0.07132	0.00098736	0.0052931
model.layers.6	20254	8.691e-05	0.06077	0.00043857	0.0023931
model.layers.7	16607	7.1261e-05	0.05580	0.00016271	0.0010286
model.layers.8	14456	6.2031e-05	0.05206	0.00000	0.00000
model.layers.9	29171	0.00012517	0.07238	0.0011131	0.0055838
model.layers.10	21836	9.3698e-05	0.06329	0.00055824	0.0030859
model.layers.11	**33387**	**0.00014326**	**0.07760**	**0.001432**	**0.0070173**
model.layers.12	22334	9.5835e-05	0.06372	0.00059591	0.0032026
model.layers.13	25895	0.00011112	0.06889	0.00086527	0.0046239
model.layers.14	21354	9.163e-05	0.06216	0.00052178	0.0027749
model.layers.15	**31556**	**0.00013541**	**0.07581**	**0.0012935**	**0.0065249**
model.layers.16	19965	8.567e-05	0.06054	0.00041671	0.0023291
model.layers.17	25987	0.00011151	0.06890	0.00087223	0.0046281
model.layers.18	30751	0.00013195	0.07487	0.0012326	0.0062683
model.layers.19	**43708**	**0.00018755**	**0.08881**	**0.0022127**	**0.01010**
model.layers.20	**34003**	**0.00014591**	**0.07892**	**0.0014786**	**0.0073791**
model.layers.21	**46579**	**0.00019987**	**0.09185**	**0.0024299**	**0.01093**
model.layers.22	**45205**	**0.00019397**	**0.09138**	**0.0023259**	**0.01080**
model.layers.23	**99608**	**0.00042742**	**0.13522**	**0.0064411**	**0.02285**
model.layers.24	**52723**	**0.00022623**	**0.09866**	**0.0028946**	**0.01281**
model.layers.25	**93421**	**0.00040087**	**0.13195**	**0.0059731**	**0.02195**
model.layers.26	**80685**	**0.00034622**	**0.12198**	**0.0050097**	**0.01921**
model.layers.27	**80585**	**0.00034579**	**0.17247**	**0.0050021**	**0.03309**

表 6.27: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{DPO}_{\mathrm{D1}}}$ 間の各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	3.1218e-05	1.6349e-05	3.4877e-05	1.4792e-05	**7.6800e-03**	1.8190e-05	0.0000e+00
model.layer.1	3.2075e-05	7.0844e-05	9.7547e-05	4.0015e-05	5.2300e-02	**1.4261e-01**	0.0000e+00
model.layer.2	5.8933e-05	3.1607e-05	5.3406e-05	7.9797e-05	2.8670e-02	**5.3400e-02**	0.0000e+00
model.layer.3	7.2168e-05	2.9973e-05	3.2697e-05	3.1374e-05	1.9546e-04	**2.0431e-04**	0.0000e+00
model.layer.4	7.1311e-05	1.6240e-04	1.6403e-04	6.9132e-05	1.8701e-04	**2.6812e-04**	0.0000e+00
model.layer.5	1.6411e-04	**2.8883e-04**	9.3188e-05	7.4581e-05	2.2956e-04	1.2013e-04	0.0000e+00
model.layer.6	1.1997e-04	2.3978e-05	9.3733e-05	5.4262e-05	1.1475e-04	**1.4742e-04**	0.0000e+00
model.layer.7	1.1615e-04	**2.2507e-04**	4.8501e-05	4.1183e-05	1.2069e-04	8.6751e-05	0.0000e+00
model.layer.8	4.0093e-05	**1.3624e-04**	4.1962e-05	1.3297e-04	9.3865e-05	8.1493e-05	0.0000e+00
model.layer.9	2.7092e-05	1.4823e-04	2.0708e-05	1.1288e-04	1.5844e-04	**2.4016e-04**	0.0000e+00
model.layer.10	4.9591e-05	1.0899e-04	1.0790e-04	1.1600e-04	8.0138e-05	**2.0429e-04**	0.0000e+00
model.layer.11	7.2946e-05	1.0899e-04	1.1499e-04	1.7688e-04	1.4643e-04	**2.9199e-04**	0.0000e+00
model.layer.12	6.3838e-05	2.9700e-04	1.2098e-04	**3.6777e-04**	1.0194e-04	1.3404e-04	1.4729e-08
model.layer.13	1.3499e-04	**2.9700e-04**	1.0027e-04	2.0187e-04	1.1750e-04	1.8942e-04	0.0000e+00
model.layer.14	7.9953e-05	3.8474e-04	1.4332e-04	**5.0806e-04**	1.0612e-04	8.2863e-05	1.4729e-08
model.layer.15	2.2795e-04	**3.9291e-04**	1.0736e-04	1.6458e-04	1.1671e-04	2.6028e-04	0.0000e+00
model.layer.16	2.2437e-04	**2.5885e-04**	1.8747e-04	2.3137e-04	1.4765e-04	4.8118e-05	0.0000e+00
model.layer.17	1.8840e-04	**8.3869e-04**	1.7820e-04	5.9322e-04	1.2832e-04	7.9078e-05	0.0000e+00
model.layer.18	9.7236e-05	**6.6430e-04**	4.1417e-05	4.6773e-04	2.0365e-04	1.2331e-04	0.0000e+00
model.layer.19	8.6103e-05	**8.7302e-04**	7.9564e-05	9.7937e-05	1.8739e-04	3.9567e-04	1.3256e-07
model.layer.20	2.3604e-04	**8.9100e-04**	1.7112e-04	3.5726e-04	1.4521e-04	2.1465e-04	0.0000e+00
model.layer.21	4.2740e-04	**7.0790e-04**	4.8774e-04	1.4737e-04	1.2368e-04	4.2128e-04	2.9457e-08
model.layer.22	2.0771e-04	**1.9500e-03**	1.6512e-04	3.7353e-04	1.7779e-04	3.2054e-04	2.3566e-07
model.layer.23	1.5313e-04	**9.9782e-04**	2.2834e-04	6.7800e-04	5.9707e-04	6.7953e-04	1.0310e-07
model.layer.24	1.9159e-04	**6.3869e-04**	9.9182e-05	4.9684e-04	2.6103e-04	3.6531e-04	0.0000e+00
model.layer.25	1.1351e-04	9.6675e-04	2.0163e-05	**1.6600e-03**	1.6528e-04	8.4875e-04	1.6201e-07
model.layer.26	1.7065e-04	**1.7400e-03**	2.9319e-04	1.3300e-03	4.3660e-04	4.1349e-04	8.8371e-08
model.layer.27	2.6321e-04	**3.7400e-03**	2.4523e-05	1.3100e-03	7.8194e-05	7.0685e-04	1.4729e-06

表 6.28: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$ 間の M_{base} から計算したタスクベクトルのコンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	**355036**	**0.0015235**	**0.27748**	**0.03339**	**0.05011**
model.layers.1	**9895489**	**0.04246**	**4.19639**	**1.00000**	**1.00000**
model.layers.2	**3476695**	**0.01492**	**2.40902**	**0.34967**	**0.56677**
model.layers.3	39396	0.00016905	0.08820	0.0014136	0.004235
model.layers.4	45167	0.00019381	0.09398	0.0019983	0.0056354
model.layers.5	36244	0.00015552	0.08440	0.0010942	0.0033137
model.layers.6	43662	0.00018735	0.09151	0.0018458	0.0050372
model.layers.7	25444	0.00010918	0.07073	0.00000	0.00000
model.layers.8	36703	0.00015749	0.08475	0.0011407	0.0033987
model.layers.9	**71661**	**0.0003075**	**0.11710**	**0.0046826**	**0.01124**
model.layers.10	**65295**	**0.00028018**	**0.11206**	**0.0040376**	**0.01002**
model.layers.11	**84588**	**0.00036297**	**0.12739**	**0.0059923**	**0.01373**
model.layers.12	**72136**	**0.00030954**	**0.11756**	**0.0047307**	**0.01135**
model.layers.13	**74717**	**0.00032061**	**0.12008**	**0.0049922**	**0.01196**
model.layers.14	49247	0.00021132	0.09741	0.0024116	0.0064679
model.layers.15	58623	0.00025155	0.10615	0.0033616	0.0085853
model.layers.16	58147	0.00024951	0.10587	0.0033134	0.0085177
model.layers.17	51721	0.00022193	0.10009	0.0026623	0.0071156
model.layers.18	49586	0.00021277	0.09808	0.002446	0.0066289
model.layers.19	50815	0.00021805	0.09907	0.0025705	0.0068692
model.layers.20	**68732**	**0.00029493**	**0.11550**	**0.0043858**	**0.01085**
model.layers.21	**69073**	**0.00029639**	**0.11604**	**0.0044203**	**0.01098**
model.layers.22	**66031**	**0.00028334**	**0.11346**	**0.0041121**	**0.01036**
model.layers.23	**68066**	**0.00029207**	**0.11590**	**0.0043183**	**0.01095**
model.layers.24	43649	0.0001873	0.09224	0.0018445	0.0052137
model.layers.25	**97667**	**0.00041909**	**0.13845**	**0.0073174**	**0.01641**
model.layers.26	**96468**	**0.00041394**	**0.13756**	**0.0071959**	**0.01620**
model.layers.27	**62068**	**0.00026633**	**0.31409**	**0.0037106**	**0.05899**

表 6.29: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$ 間の M_{base} から計算したタスクベクトルの各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	1.1413e-04	2.4523e-05	2.5776e-04	3.2697e-05	**5.1600e-03**	3.0842e-05	0.0000e+00
model.layer.1	6.5473e-05	8.6648e-05	1.7275e-04	3.2697e-05	3.5670e-02	**1.1005e-01**	0.0000e+00
model.layer.2	1.3157e-04	6.3760e-05	1.2098e-04	8.8906e-05	1.8310e-02	**3.2850e-02**	0.0000e+00
model.layer.3	1.0152e-04	5.5586e-05	1.7057e-04	5.1070e-05	**2.8127e-04**	2.6399e-04	0.0000e+00
model.layer.4	4.1775e-04	2.7139e-04	**4.3052e-04**	7.3336e-05	3.0663e-04	2.4673e-04	0.0000e+00
model.layer.5	1.0175e-04	**3.5967e-04**	1.6131e-04	1.0128e-04	3.1510e-04	1.6623e-04	0.0000e+00
model.layer.6	2.7629e-04	9.9727e-05	1.4169e-04	1.9953e-04	2.2950e-04	**3.1703e-04**	0.0000e+00
model.layer.7	1.9759e-04	**3.6458e-04**	8.0654e-05	8.8750e-05	1.6740e-04	1.4114e-04	0.0000e+00
model.layer.8	1.0323e-04	**4.6321e-04**	5.5041e-05	2.5465e-04	2.5457e-04	2.0430e-04	0.0000e+00
model.layer.9	1.0440e-04	4.9918e-04	8.9373e-05	2.9537e-04	4.0903e-04	**5.5490e-04**	0.0000e+00
model.layer.10	3.2876e-04	2.6485e-04	4.7139e-04	**5.1039e-04**	2.8596e-04	4.9706e-04	2.9457e-08
model.layer.11	2.4048e-04	4.9428e-04	3.0736e-04	4.7396e-04	4.0752e-04	**6.8151e-04**	0.0000e+00
model.layer.12	2.0389e-04	6.3215e-04	3.0518e-04	**1.2200e-03**	3.5139e-04	4.1576e-04	3.0930e-07
model.layer.13	8.3012e-04	**1.2700e-03**	1.0700e-03	9.2526e-04	2.7177e-04	4.3339e-04	1.6201e-07
model.layer.14	1.4838e-04	5.7983e-04	4.5013e-04	**7.8902e-04**	3.2046e-04	1.9897e-04	7.2170e-07
model.layer.15	4.6205e-04	6.0272e-04	**6.2125e-04**	5.6294e-04	2.5389e-04	3.8196e-04	5.8914e-07
model.layer.16	4.8314e-04	**1.1500e-03**	6.0599e-04	7.0883e-04	3.9873e-04	1.8461e-04	1.3256e-07
model.layer.17	4.2273e-04	**1.5000e-03**	4.5504e-04	3.7423e-04	3.3334e-04	2.2483e-04	0.0000e+00
model.layer.18	6.0179e-05	**1.5500e-03**	1.7439e-04	7.5477e-04	3.3593e-04	1.9367e-04	0.0000e+00
model.layer.19	1.1631e-04	**1.6700e-03**	5.3406e-04	7.8225e-04	2.2109e-04	2.9752e-04	2.7984e-07
model.layer.20	4.2974e-04	**1.3200e-03**	6.4850e-05	4.9054e-04	3.4319e-04	4.5716e-04	4.5659e-07
model.layer.21	5.4223e-04	**7.3896e-04**	1.2970e-04	1.5095e-04	1.4882e-04	7.1389e-04	1.4729e-08
model.layer.22	7.8240e-04	**1.8100e-03**	2.8011e-04	1.4400e-03	2.2278e-04	2.7282e-04	7.2170e-07
model.layer.23	3.2129e-04	**9.3787e-04**	1.7548e-04	5.8256e-04	4.6961e-04	3.3179e-04	2.9457e-08
model.layer.24	1.1888e-04	**1.5800e-03**	1.1008e-04	1.3305e-04	2.3784e-04	3.1166e-04	4.4186e-08
model.layer.25	1.0500e-03	**2.7300e-03**	4.1689e-04	1.1200e-03	3.8604e-04	5.5827e-04	7.3643e-08
model.layer.26	1.0000e-03	1.5700e-03	5.5041e-04	**2.4900e-03**	3.4549e-04	3.5736e-04	0.0000e+00
model.layer.27	8.5893e-04	**2.0800e-03**	4.0926e-04	8.3332e-04	9.4233e-05	4.3234e-04	2.6511e-07

表 6.30: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$ 間の M_{base} から計算したタスクベクトルのコンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	**489336**	**0.0020997**	**0.34001**	**0.04510**	**0.06052**
model.layers.1	**10285618**	**0.04414**	**4.49179**	**1.00000**	**1.00000**
model.layers.2	**4452921**	**0.01911**	**3.10350**	**0.43145**	**0.68585**
model.layers.3	51120	0.00021936	0.10045	0.0023861	0.0063149
model.layers.4	47711	0.00020473	0.10047	0.0020538	0.0063201
model.layers.5	42169	0.00018095	0.09091	0.0015136	0.004157
model.layers.6	38388	0.00016472	0.08609	0.001145	0.0030664
model.layers.7	33694	0.00014458	0.08126	0.0006875	0.0019731
model.layers.8	26641	0.00011432	0.07254	0.00000	0.00000
model.layers.9	53701	0.00023043	0.10155	0.0026377	0.0065647
model.layers.10	40280	0.00017284	0.08866	0.0013295	0.0036474
model.layers.11	**62213**	**0.00026696**	**0.10954**	**0.0034674**	**0.0083711**
model.layers.12	59576	0.00025564	0.10718	0.0032104	0.0078383
model.layers.13	**84048**	**0.00036065**	**0.12746**	**0.0055958**	**0.01243**
model.layers.14	45379	0.00019472	0.09352	0.0018265	0.0047465
model.layers.15	47962	0.0002058	0.09585	0.0020783	0.0052729
model.layers.16	47919	0.00020562	0.09634	0.0020741	0.0053853
model.layers.17	46394	0.00019908	0.09485	0.0019254	0.0050475
model.layers.18	**65619**	**0.00028157**	**0.11318**	**0.0037994**	**0.0091949**
model.layers.19	**64756**	**0.00027787**	**0.11218**	**0.0037153**	**0.0089687**
model.layers.20	**119507**	**0.0005128**	**0.15305**	**0.0090522**	**0.01822**
model.layers.21	**90845**	**0.00038982**	**0.13325**	**0.0062583**	**0.01374**
model.layers.22	**84568**	**0.00036288**	**0.12877**	**0.0056465**	**0.01272**
model.layers.23	**127703**	**0.00054797**	**0.15905**	**0.0098511**	**0.01957**
model.layers.24	**103295**	**0.00044324**	**0.14293**	**0.0074719**	**0.01593**
model.layers.25	**156390**	**0.00067107**	**0.17648**	**0.01265**	**0.02352**
model.layers.26	**150769**	**0.00064695**	**0.17390**	**0.01210**	**0.02293**
model.layers.27	**77729**	**0.00033353**	**0.47941**	**0.0049798**	**0.09207**

表 6.31: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$ 間の M_{base} から計算したタスクベクトルの各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	1.5700e-03	1.0354e-05	6.8556e-04	3.4877e-05	**6.8500e-03**	3.4229e-05	0.0000e+00
model.layer.1	6.7886e-05	8.1199e-05	1.7384e-04	6.3838e-05	3.4090e-02	**1.1737e-01**	0.0000e+00
model.layer.2	1.4854e-04	7.5749e-05	1.3733e-04	1.8279e-04	2.2330e-02	**4.3190e-02**	0.0000e+00
model.layer.3	1.0051e-04	5.7765e-05	7.1934e-05	6.1035e-05	3.3550e-04	**3.8330e-04**	5.8914e-08
model.layer.4	2.0179e-04	2.7629e-04	1.5586e-04	1.7960e-04	3.0747e-04	**3.1141e-04**	0.0000e+00
model.layer.5	2.5177e-04	**3.4823e-04**	1.4114e-04	1.4083e-04	3.2357e-04	2.1001e-04	0.0000e+00
model.layer.6	9.7625e-05	6.7575e-05	6.8665e-05	1.0860e-04	2.5513e-04	**2.6757e-04**	0.0000e+00
model.layer.7	9.7392e-05	2.0163e-04	7.5749e-05	1.2697e-04	**2.2745e-04**	2.1887e-04	0.0000e+00
model.layer.8	9.1397e-05	**2.9210e-04**	4.0872e-05	1.3717e-04	1.6970e-04	1.7044e-04	0.0000e+00
model.layer.9	7.1545e-05	2.2452e-04	8.8828e-05	1.7298e-04	3.0531e-04	**4.3090e-04**	0.0000e+00
model.layer.10	1.9548e-04	1.8038e-04	1.3242e-04	2.6166e-04	2.3486e-04	**2.6346e-04**	0.0000e+00
model.layer.11	8.9529e-05	2.7411e-04	1.1553e-04	2.8462e-04	3.2977e-04	**5.0522e-04**	0.0000e+00
model.layer.12	1.2324e-04	5.1226e-04	1.7766e-04	**7.6979e-04**	3.3207e-04	3.5740e-04	3.9767e-07
model.layer.13	3.7400e-04	**8.5177e-04**	5.4768e-04	8.3269e-04	3.7593e-04	5.9580e-04	5.8914e-08
model.layer.14	2.4889e-04	**8.9427e-04**	5.8855e-04	5.3312e-04	3.1266e-04	1.6766e-04	2.9457e-08
model.layer.15	4.7154e-04	**6.5068e-04**	3.4169e-04	3.0494e-04	1.6110e-04	3.7147e-04	1.1783e-07
model.layer.16	1.3966e-04	**6.6103e-04**	1.6894e-04	5.1880e-04	3.3360e-04	2.2517e-04	0.0000e+00
model.layer.17	2.9949e-04	**1.0600e-03**	2.7902e-04	2.5574e-04	2.9653e-04	2.4545e-04	0.0000e+00
model.layer.18	8.6259e-05	**1.9200e-03**	1.5967e-04	5.1966e-04	3.7863e-04	4.1692e-04	2.9457e-08
model.layer.19	1.3811e-04	**3.2300e-03**	2.5177e-04	7.0050e-04	3.4888e-04	3.5141e-04	6.6279e-07
model.layer.20	3.8840e-04	1.1100e-03	9.9182e-05	**1.4600e-03**	3.9509e-04	9.8290e-04	3.0930e-07
model.layer.21	4.6423e-04	9.0136e-04	1.2806e-04	2.5403e-04	2.6363e-04	**9.1067e-04**	1.4729e-08
model.layer.22	5.4792e-04	**1.6900e-03**	7.2370e-04	1.2300e-03	4.1741e-04	4.2312e-04	3.7411e-06
model.layer.23	3.1654e-04	**1.8200e-03**	2.4033e-04	1.1300e-03	8.8142e-04	6.7077e-04	4.4186e-08
model.layer.24	2.8750e-04	**2.7700e-03**	2.6485e-04	4.9124e-04	5.5407e-04	7.3768e-04	2.9457e-07
model.layer.25	5.1164e-04	**6.5200e-03**	7.3188e-04	1.6300e-03	7.1537e-04	9.8569e-04	6.6279e-07
model.layer.26	1.0200e-03	2.9100e-03	1.6700e-03	**4.1600e-03**	6.8355e-04	4.3122e-04	9.5736e-07
model.layer.27	9.0813e-04	**3.8900e-03**	4.6594e-04	1.0300e-03	1.8543e-04	4.7410e-04	5.8914e-08

は浅い層でデータセットの違いを敏感に捉えていると考えられる. これは深い層にコンフリクトが集中していた実験 1 の結果と異なる.

この違いの原因として上で述べたベースモデルの事前学習の違いが挙げられる. Qwen/Qwen2.5-7B-Instruct が多言語に対応した汎用的なデータで事前学習した LLM であり, 多言語に共通する単語表現や語順・文法構造を浅い層で早期に獲得する必要があるため言語の違いを浅い層で捉えていると考えられる. D_0, D_1 は応答スタイルの特徴は異なるが日本語データセットという点で共通しているため, 1, 2 層目局所的にコンフリクトが発生している理由として, 言語識別とその言語体系に合わせたタスクに必要な言語情報の取得も浅い層で同時にしている可能性がある. 一方で, 実験 1 で用いた elyza/Llama-3-ELYZA-JP-8B は日本語に特化しているため, 浅い層はより汎用的な日本語の言語情報の処理に集中し中間層以降でデータセット間の違いのコンフリクトが生まれていると考えられる.

また表 6.25 , 表 6.27, 6.29, 表 6.31 においては, 共通してほとんどの層において v_proj のコンフリクト率が高くなっているという実験 1 と同様の結果が得られた. ロールプレイの演じ分けをするにあたって v_proj の重みが大きな意味をなしていると考えられる.

さらに, コンフリクト率が大きい 0, 1, 2 層については up_proj, down_proj のコンフリクト率が大きくなっている. 浅い層において up_proj, down_proj は入力の埋め込みといった低次の言語情報の段階で単語や局所的特徴に重みをつけるため, この部分でコンフリクトが多く発生するということは, 上で述べたように浅い層の段階で言語識別に加えて D_0, D_1 のタスクの違いを情報として保持している可能性が考えられる.

同じデータセットで手法が異なる $M_{\mathrm{DPO}_{\mathrm{D0}}}$ と $M_{\mathrm{SFT}_{\mathrm{D0}}}$, $M_{\mathrm{DPO}_{\mathrm{D1}}}$ と $M_{\mathrm{SFT}_{\mathrm{D1}}}$ に関する Conflict Limited L2 ノルム を調べる. 表 6.32 に $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間の コンフリクト数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.33 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.34 に $M_{\mathrm{DPO}_{\mathrm{D1}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.35 に各層の各パラメータのコンフリクト率の値を示す.

表 6.32, 表 6.34 に共通して 0 から 1, 16 から 27 層目の値が大きくなっている. DPO と SFT の重みの変化量の差によるものと考えられる. DPO と SFT では, DPO はベー

表 6.32: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間 コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	**647994**	**0.0027805**	**0.41853**	**0.06738**	**0.12781**
model.layers.1	**9577258**	**0.04110**	**3.11201**	**1.00000**	**1.00000**
model.layers.2	**4400722**	**0.01888**	**2.11211**	**0.45934**	**0.67622**
model.layers.3	6559	2.8145e-05	0.03678	0.00038854	0.0041961
model.layers.4	9535	4.0915e-05	**0.16654**	0.00069936	**0.04621**
model.layers.5	5104	2.1901e-05	0.03194	0.00023657	0.0026304
model.layers.6	3482	1.4941e-05	0.02621	6.7158e-05	0.00077544
model.layers.7	3114	1.3362e-05	0.02477	2.8722e-05	0.00030904
model.layers.8	2839	1.2182e-05	0.02382	0.00000	0.00000
model.layers.9	4086	1.7533e-05	0.02841	0.00013024	0.0014885
model.layers.10	3713	1.5932e-05	0.02717	9.1285e-05	0.0010867
model.layers.11	5229	2.2438e-05	0.03164	0.00024962	0.0025345
model.layers.12	4991	2.1416e-05	0.03116	0.00022477	0.0023778
model.layers.13	6813	2.9235e-05	0.03641	0.00041506	0.004079
model.layers.14	7485	3.2118e-05	0.03844	0.00048525	0.0047337
model.layers.15	7969	3.4195e-05	0.03933	0.0005358	0.005024
model.layers.16	**10316**	**4.4266e-05**	0.04596	**0.00078094**	0.0071712
model.layers.17	**12098**	**5.1912e-05**	**0.04886**	**0.00096706**	**0.0081078**
model.layers.18	**13873**	**5.9529e-05**	**0.05340**	**0.0011524**	**0.0095786**
model.layers.19	**19602**	**8.4112e-05**	**0.06315**	**0.0017508**	**0.01274**
model.layers.20	**16451**	**7.0591e-05**	**0.05757**	**0.0014217**	**0.01093**
model.layers.21	**22550**	**9.6762e-05**	**0.06781**	**0.0020587**	**0.01425**
model.layers.22	**31080**	**0.00013336**	**0.08089**	**0.0029496**	**0.01848**
model.layers.23	**35665**	**0.00015304**	**0.08339**	**0.0034285**	**0.01929**
model.layers.24	**32483**	**0.00013938**	**0.08170**	**0.0030962**	**0.01874**
model.layers.25	**50093**	**0.00021495**	**0.10143**	**0.0049354**	**0.02513**
model.layers.26	**33726**	**0.00014472**	**0.09068**	**0.003226**	**0.02165**
model.layers.27	**47955**	**0.00020577**	**0.10531**	**0.0047121**	**0.02639**

表 6.33: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間の各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	6.8509e-06	8.1744e-06	1.6349e-06	6.6952e-06	**9.5300e-03**	6.7310e-06	0.0000e+00
model.layer.1	6.8509e-06	2.1798e-05	3.2697e-05	6.9287e-06	3.6700e-02	**1.0436e-01**	0.0000e+00
model.layer.2	6.4616e-06	3.8147e-06	1.4714e-05	2.3745e-05	3.2250e-02	**3.2560e-02**	0.0000e+00
model.layer.3	3.5033e-06	6.5395e-06	4.9046e-06	6.1502e-06	3.6542e-05	**5.7927e-05**	0.0000e+00
model.layer.4	1.6816e-05	9.2643e-06	5.9945e-06	1.3235e-05	3.7646e-05	**9.6693e-05**	0.0000e+00
model.layer.5	1.5804e-05	3.2152e-05	8.1744e-06	2.7092e-05	**4.1873e-05**	2.4096e-05	0.0000e+00
model.layer.6	**2.6002e-05**	2.2343e-05	7.6294e-06	1.7127e-05	1.7556e-05	2.4759e-05	0.0000e+00
model.layer.7	8.6415e-06	**6.0490e-05**	4.3597e-06	1.1133e-05	1.6835e-05	2.3536e-05	0.0000e+00
model.layer.8	5.9945e-06	**4.1417e-05**	5.4496e-06	2.9817e-05	1.5200e-05	1.8573e-05	0.0000e+00
model.layer.9	1.1055e-05	**7.7384e-05**	9.2643e-06	2.4757e-05	2.0944e-05	3.0120e-05	0.0000e+00
model.layer.10	9.1086e-06	**4.0327e-05**	1.4714e-05	2.1175e-05	2.2579e-05	2.4891e-05	0.0000e+00
model.layer.11	1.4714e-05	3.9782e-05	3.5967e-05	**5.5897e-05**	2.3993e-05	3.7617e-05	0.0000e+00
model.layer.12	1.1911e-05	4.2507e-05	1.6894e-05	**6.8509e-05**	1.9942e-05	3.6748e-05	0.0000e+00
model.layer.13	1.9541e-05	**9.6457e-05**	3.2152e-05	8.1510e-05	2.5937e-05	5.1815e-05	0.0000e+00
model.layer.14	2.0241e-05	9.8092e-05	2.6703e-05	**1.7260e-04**	3.1504e-05	3.8839e-05	4.4186e-08
model.layer.15	2.9194e-05	8.5558e-05	6.9754e-05	**1.0066e-04**	3.3287e-05	5.5306e-05	1.4729e-08
model.layer.16	2.5535e-05	1.3624e-04	1.5259e-05	**1.3748e-04**	5.1521e-05	6.5483e-05	0.0000e+00
model.layer.17	2.8493e-05	1.5967e-04	2.6158e-05	**2.7427e-04**	5.1417e-05	6.4452e-05	1.4729e-08
model.layer.18	1.9774e-05	**3.3460e-04**	8.1744e-06	1.3912e-04	5.5703e-05	1.0930e-04	0.0000e+00
model.layer.19	4.2351e-05	2.2834e-04	2.1253e-05	**2.9015e-04**	7.1154e-05	1.4790e-04	0.0000e+00
model.layer.20	3.7524e-05	**1.8311e-04**	1.0899e-05	1.4247e-04	8.6030e-05	1.1697e-04	0.0000e+00
model.layer.21	6.5006e-05	**2.2071e-04**	7.6294e-06	1.0993e-04	1.2843e-04	1.6443e-04	0.0000e+00
model.layer.22	3.7368e-05	**3.7711e-04**	2.1253e-05	3.3367e-04	1.7182e-04	2.0455e-04	4.2713e-07
model.layer.23	1.0603e-04	3.0736e-04	9.4278e-05	**3.1452e-04**	1.5599e-04	2.7887e-04	1.4729e-08
model.layer.24	1.0276e-04	**3.8147e-04**	3.0518e-05	1.8217e-04	1.4061e-04	2.7249e-04	2.7984e-07
model.layer.25	7.9719e-05	**6.8120e-04**	3.9782e-05	4.2553e-04	2.3467e-04	3.8802e-04	2.9457e-08
model.layer.26	4.9591e-05	**6.4904e-04**	5.9945e-05	5.7267e-04	1.3580e-04	2.2398e-04	7.3643e-08
model.layer.27	1.4402e-04	**1.1100e-03**	8.1744e-06	4.8766e-04	1.1104e-04	4.4507e-04	3.9767e-07

表 6.34: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D1}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間 コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	**671970**	**0.0028834**	**0.42896**	**0.05227**	**0.10604**
model.layers.1	**12799376**	**0.05492**	**3.83702**	**1.00000**	**1.00000**
model.layers.2	**4269135**	**0.01832**	**2.16689**	**0.33338**	**0.56191**
model.layers.3	8746	3.7529e-05	**0.05938**	0.00044458	**0.0090946**
model.layers.4	8138	3.492e-05	**0.05018**	0.00039707	**0.0066813**
model.layers.5	4997	2.1442e-05	0.03183	0.00015161	0.0018682
model.layers.6	4611	1.9786e-05	0.03029	0.00012144	0.0014644
model.layers.7	3641	1.5624e-05	0.02703	4.5638e-05	0.00061101
model.layers.8	3057	1.3118e-05	0.02471	0.00000	0.00000
model.layers.9	4300	1.8451e-05	0.02902	9.7137e-05	0.001132
model.layers.10	3958	1.6984e-05	0.02806	7.0411e-05	0.00088003
model.layers.11	5354	2.2974e-05	0.03264	0.0001795	0.0020825
model.layers.12	4595	1.9717e-05	0.03026	0.00012019	0.001457
model.layers.13	5699	2.4454e-05	0.03390	0.00020647	0.0024113
model.layers.14	**9123**	**3.9147e-05**	0.04353	**0.00047404**	0.0049376
model.layers.15	7412	3.1805e-05	0.03872	0.00034033	0.0036749
model.layers.16	**10764**	**4.6188e-05**	**0.04798**	**0.00060228**	**0.006106**
model.layers.17	**11371**	**4.8793e-05**	**0.04875**	**0.00064972**	**0.0063068**
model.layers.18	**16692**	**7.1625e-05**	**0.05910**	**0.0010655**	**0.0090232**
model.layers.19	**15999**	**6.8652e-05**	**0.05811**	**0.0010114**	**0.0087626**
model.layers.20	**10954**	**4.7004e-05**	0.04740	**0.00061713**	0.0059531
model.layers.21	8147	3.4959e-05	0.04003	0.00039777	0.00402
model.layers.22	**16835**	**7.2239e-05**	**0.05958**	**0.0010767**	**0.009148**
model.layers.23	**16523**	**7.09e-05**	**0.05746**	**0.0010523**	**0.0085924**
model.layers.24	**15256**	**6.5463e-05**	**0.05586**	**0.00095332**	**0.008173**
model.layers.25	**16655**	**7.1467e-05**	**0.05869**	**0.0010626**	**0.0089139**
model.layers.26	**16169**	**6.9381e-05**	**0.05695**	**0.0010247**	**0.0084593**
model.layers.27	**22055**	**9.4638e-05**	**0.12671**	**0.0014846**	**0.02676**

表 6.35: 実験 2 における $M_{\mathrm{DPO}_{\mathrm{D1}}}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$ 間の各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	4.6711e-06	3.8147e-06	2.1798e-06	5.2939e-06	**9.8800e-03**	1.0737e-05	0.0000e+00
model.layer.1	9.8092e-06	2.5068e-05	6.5395e-06	8.0186e-06	7.3840e-02	**1.1468e-01**	0.0000e+00
model.layer.2	6.7730e-06	8.1744e-06	7.0844e-06	2.2577e-05	2.9970e-02	**3.2900e-02**	0.0000e+00
model.layer.3	9.8871e-06	2.6703e-05	4.3597e-06	1.0121e-05	5.5939e-05	**6.8252e-05**	0.0000e+00
model.layer.4	1.1833e-05	2.7248e-05	1.1989e-05	1.4636e-05	3.8589e-05	**7.5204e-05**	0.0000e+00
model.layer.5	1.7361e-05	**7.0844e-05**	1.2534e-05	3.4722e-05	4.2153e-05	1.9339e-05	0.0000e+00
model.layer.6	3.4566e-05	1.4714e-05	1.9073e-05	1.3702e-05	2.2491e-05	**3.5378e-05**	0.0000e+00
model.layer.7	1.2223e-05	**4.4686e-05**	1.0899e-05	1.3390e-05	2.5112e-05	2.2166e-05	0.0000e+00
model.layer.8	6.8509e-06	2.3433e-05	7.0844e-06	**3.8692e-05**	1.5995e-05	1.9589e-05	0.0000e+00
model.layer.9	1.6894e-05	**4.6866e-05**	4.9046e-06	2.9739e-05	2.0046e-05	3.3066e-05	0.0000e+00
model.layer.10	7.6294e-06	**3.5422e-05**	1.7439e-05	1.7828e-05	2.0208e-05	3.1843e-05	0.0000e+00
model.layer.11	3.2464e-05	4.7956e-05	4.0872e-05	**5.4184e-05**	2.5039e-05	3.5025e-05	0.0000e+00
model.layer.12	1.3780e-05	**4.8501e-05**	1.7439e-05	4.7333e-05	2.0119e-05	3.4200e-05	1.4729e-08
model.layer.13	1.6972e-05	6.2670e-05	1.5804e-05	**7.4426e-05**	2.4287e-05	4.0238e-05	0.0000e+00
model.layer.14	5.6442e-05	7.1934e-05	1.8529e-05	**2.3316e-04**	3.6748e-05	4.0386e-05	0.0000e+00
model.layer.15	2.0008e-05	3.7057e-05	2.2343e-05	**8.0109e-05**	3.2904e-05	5.5718e-05	0.0000e+00
model.layer.16	2.1020e-05	6.1035e-05	1.0899e-05	**1.3453e-04**	5.2846e-05	7.4320e-05	0.0000e+00
model.layer.17	1.3935e-05	1.5095e-04	2.0163e-05	**2.0467e-04**	4.7735e-05	7.3761e-05	0.0000e+00
model.layer.18	3.3865e-05	**1.7057e-04**	4.6321e-05	1.6520e-04	8.7517e-05	1.1481e-04	0.0000e+00
model.layer.19	3.3865e-05	**2.8774e-04**	1.5804e-05	2.7123e-04	5.6396e-05	1.1325e-04	7.3643e-08
model.layer.20	1.6582e-05	**1.9782e-04**	9.2643e-06	7.7695e-05	4.9208e-05	8.8695e-05	0.0000e+00
model.layer.21	4.1806e-05	**1.5913e-04**	3.5967e-05	4.8345e-05	3.6836e-05	6.0829e-05	0.0000e+00
model.layer.22	3.0829e-05	**2.2997e-04**	3.1607e-05	1.7314e-04	7.9210e-05	1.2309e-04	0.0000e+00
model.layer.23	3.6979e-05	**2.6757e-04**	2.8338e-05	1.8529e-04	6.1035e-05	1.3228e-04	0.0000e+00
model.layer.24	1.3118e-04	**1.7221e-04**	1.3297e-04	1.2059e-04	4.7779e-05	1.2104e-04	0.0000e+00
model.layer.25	6.0334e-05	**3.2752e-04**	6.5395e-06	2.0623e-04	2.3109e-05	1.6272e-04	1.4729e-08
model.layer.26	1.7548e-04	**3.6676e-04**	7.9019e-05	2.0615e-04	7.8135e-05	7.5764e-05	0.0000e+00
model.layer.27	3.7524e-05	**4.3706e-04**	3.1607e-05	2.2826e-04	4.9002e-05	2.1289e-04	0.0000e+00

スモデルから離れないような制約のもとで最適化するため, 同じ日本語タスクにおいても LLM は 1 層でも 1 億個以上のパラメータをもつため冗長性により日本語と識別する働きをしていると考えられる浅い層において DPO と SFT では重みの変化量が異なりコンフリクトが発生していると考えられる.

表 6.36: 実験 2 における ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, M_{base}), ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$), ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D0}}}$) 間の L1 ノルム, L2 ノルム

モデル	L1 ノルム	L2 ノルム
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$	1.63094×10^5	4.61360
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, M_{base}	3.29579×10^5	6.86015
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D0}}}$	3.50282×10^5	7.56997
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$	3.85041×10^5	8.09529

表 6.36 には, 実験 2 における ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, M_{base}), ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D1}}}$), ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D0}}}$) 間の L1 ノルム, L2 ノルム を示している. 実験 1 で偉得た結果と同様に, 似た出力をすると考えられるモデルへの回帰現象は見られなかった.

学習後のモデルの出力結果から LLM の性能を評価する. 表 6.37, 6.38 にそれぞれ実験 2 における学習後のモデルの D_0, D_1 との BLEU, BERTScore の値を示す.

表 6.37: 実験 2 における各モデルの生成結果とデータセット D_0 との BLEU, BERTScore

モデル	BLEU (D_0 train)	BLEU (D_0 test)	BERTScore (D_0 train)	BERTScore (D_0 test)
M_{base}	0.01223	0.01851	0.74054	0.72492
$M_{\mathrm{SFT}_{\mathrm{D0}}}$	0.04710	0.04150	0.76229	0.74681
$M_{\mathrm{SFT}_{\mathrm{D1}}}$	0.02466	0.04392	0.73408	0.0.72332
$M_{\mathrm{DPO}_{\mathrm{D0}}}$	0.02883	0.04292	0.73196	0.72512
$M_{\mathrm{DPO}_{\mathrm{D1}}}$	0.00716	0.01265	0.72710	0.71840
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	0.03303	0.03293	0.73624	0.72237
$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D2}}})}$	0.00391	0.0	0.68725	0.67552
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$	0.01077	0.02196	0.71055	0.70108
$M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$	0.00686	0.00745	0.70311	0.69022

表 6.39 に LLM の評価結果を示す.

実験 1 と同様に学習データセットにおいて D_1 を最後に用いたモデルは M_{base} より明らかなロールプレイ性能の向上が見られた. また, LLM が最も高いスコアをつけた

表 6.38: 実験 2 における各モデルの生成結果とデータセット D_1 との BLEU, BERTScore

モデル	BLEU (D_1 train)	BLEU (D_1 test)	BERTScore (D_1 train)	BERTScore (D_1 test)
M_{base}	0.01515	0.01860	0.74367	0.71784
$M_{\text{SFT}_{\text{D0}}}$	0.02737	0.03569	0.74437	0.72917
$M_{\text{SFT}_{\text{D1}}}$	0.05206	0.04202	0.76334	0.75507
$M_{\text{DPO}_{\text{D0}}}$	0.017180	0.024115	0.72177	0.71072
$M_{\text{DPO}_{\text{D1}}}$	0.029059	0.018137	0.75078	0.72947
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.01825	0.01904	0.72229	0.0.70622
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	0.02416	0.01750	0.72868	0.70621
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	0.00431	0.00744	0.69145	0.68640
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	0.04022	**0.03838**	0.74317	0.72441

のは $M_{\text{DPO}_{\text{D1}}}$ となった.

表 6.39: 実験 2 における評価者 LLM による各モデルの生成結果の評価

モデル	train データのクエリ (162 件)	test データのクエリ (40 件)
dataset	7.79	8.1
M_{base}	5.49	6.29
$M_{\text{SFT}_{\text{D0}}}$	3.96	3.65
$M_{\text{SFT}_{\text{D1}}}$	8.58	8.67
$M_{\text{DPO}_{\text{D0}}}$	3.23	3.77
$M_{\text{DPO}_{\text{D1}}}$	**8.70**	**8.90**
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	2.50	1.85
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D2}}})}$	8.58	8.42
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	2.72	2.31
$M_{\text{DPO}_{\text{D1}}(\text{M}_{\text{SFT}_{\text{D1}}})}$	8.52	8.49

また, 表 6.40 に「このカフェ素敵ですね」に対する LLM の応答を示す.

実験 1 と同様に, 学習データ内の「雰囲気」という単語や D_1 を用いた場合の「わたくし」や語尾の「わ」,「わね」などキャラクターの特徴的な表現を抑えてロールプレイが実現できていることがわかる. また, 実験 1 では見られなかった「ありがとうございます。」という単語の出現頻度が多くなっている. これは実験 1 の結果と比較するとベースモデルの違いによるものと考慮でき, Qwen/Qwen2.5-7B-Instruct の大規模な事前学習と instruction tuning の影響がファインチューニング後にも残っているものと考えられる.

表 6.40: 実験 2 で学習した LLM の「このカフェ素敵ですね」に対する応答

モデル	「このカフェ素敵ですね」に対する LLM の応答
D_0	私もこのお店が気に入っています。静かな雰囲気がいいですね
D_1	はい私もこちらのお店が大好きですわ。落ち着いた雰囲気がいいですわね
M_{base}	ありがとうございます！このカフェは、内装もおしゃれで、雰囲気もとても良く友達と来るにはぴったりなお店です。
$M_{\text{SFT}_{D0}}$	ありがとうございます。落ち着いた雰囲気のカフェです
$M_{\text{SFT}_{D1}}$	ありがとうございます。お気に入りの場所ですわ
$M_{\text{DPO}_{D0}}$	店内が清潔で雰囲気も良いです。
$M_{\text{DPO}_{D1}}$	こちらのカフェは落ち着いた雰囲気でとても素敵ですわ。
$M_{\text{DPO}_{D0}(M_{\text{SFT}_{D2}})}$	ありがとうございます, よくここに来ます
$M_{\text{DPO}_{D1}(M_{\text{SFT}_{D2}})}$	わたくしも存じ上げておりませぬわ。お気に入りの場所でございますわ。
$M_{\text{DPO}_{D0}(M_{\text{SFT}_{D1}})}$	ありがとうございます。
$M_{\text{DPO}_{D1}(M_{\text{SFT}_{D1}})}$	わたくしも存じ上げませんわ！

6.3 実験 3

(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 6.41 に結果を示す. 実験 1, 2 と同様に, SFT は DPO と比較して重みの変化量が大きい手法であるといえる.

表 6.42, 6.43 に, それぞれ SFT を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値, DPO を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値を示す.

表 6.42, 6.43 に共通して DPO, SFT D_0, $D\prime_1$ 問わず 8 から 16 層付近の中間層と深い層で変化量が大きくなっている. これは D_1 と $D'_{1]}$ の違いが現れていると考える. D_0 は D_1 から応答の内容を変えずに口調だけ一般的な応答に変換したデータセットだが, D'_1 は D_1 を参考に生成しているが男子大学生のペルソナに従って応答を生成するようにプロンプトで指示しているため口調だけでなくデータセット内に登場する固有名詞なども D_0 とは異なる. このため D_0, $D\prime_1$ に共通して, データセット内に存在する語彙の情報を強く保持するために 7 から 14 層の中間層の変化量が大きくなっており, 応答のスタイルを学習するため深い層でも変化量が大きくなっていると考えられる.

同一でデータセットが異なるモデル間の Conflict Limited L2 ノルムを計算する. 表 6.44 に $M_{\text{SFT}_{D0}}$, $M_{\text{SFT}_{D'1}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.45 に各層の各パラメータのコンフリクト率の値を示す.

表 6.41: 実験 3 における (学習後のモデル, 学習前のモデル) の関係にある 2 モデル間の L1 ノルム, L2 ノルム

モデル組	L1 ノルム	L2 ノルム
$M_{\mathrm{SFT_{D0}}}$, M_{base}	2.36883 $\times 10^5$	1.57609
$M_{\mathrm{SFT_{D'1}}}$, M_{base}	2.34641 $\times 10^5$	1.56648
$M_{\mathrm{SFT_{D'2}}}$, M_{base}	2.34641 $\times 10^5$	1.56648
$M_{\mathrm{DPO_{D0}}}$, M_{base}	1.33416 $\times 10^5$	0.27663
$M_{\mathrm{DPO_{D'1}}}$, M_{base}	1.28637 $\times 10^5$	0.24942
$M_{\mathrm{DPO_{D0}(M_{SFT_{D'2}})}}$, M_{base}	2.97099 $\times 10^5$	2.45277
$M_{\mathrm{DPO_{D'1}(M_{SFT_{D'2}})}}$, M_{base}	2.94181 $\times 10^5$	2.41816
$M_{\mathrm{DPO_{D0}(M_{SFT_{D'2}})}}$, $M_{\mathrm{SFT_{D'2}}}$	1.45960 $\times 10^5$	0.36411
$M_{\mathrm{DPO_{D'1}(M_{SFT_{D'2}})}}$, $M_{\mathrm{SFT_{D'2}}}$	141160 $\times 10^5$	0.33912
$M_{\mathrm{DPO_{D0}(M_{SFT_{D'1}})}}$, M_{base}	2.93887 $\times 10^5$	2.41596
$M_{\mathrm{DPO_{D'1}(M_{SFT_{D'1}})}}$, M_{base}	2.94542 $\times 10^5$	2.44918
$M_{\mathrm{DPO_{D0}(M_{SFT_{D'1}})}}$, $M_{\mathrm{SFT_{D'1}}}$	1.39841 $\times 10^5$	0.32770
$M_{\mathrm{DPO_{D'1}(M_{SFT_{D'1}})}}$, $M_{\mathrm{SFT_{D'1}}}$	140739 $\times 10^5$	0.37868

同様に, 表 6.46 に $M_{\mathrm{DPO_{D0}}}$, $M_{\mathrm{DPO_{D'1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.47 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.48 に $M_{\mathrm{DPO_{D0}(M_{SFT_{D'2}})}}$, $M_{\mathrm{DPO_{D'1}(M_{SFT_{D'2}})}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.49 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.50 に $M_{\mathrm{DPO_{D0}(M_{SFT_{D'1}})}}$, $M_{\mathrm{DPO_{D'1}(M_{SFT_{D'1}})}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.49 に各層の各パラメータのコンフリクト率の値を示す.

表 6.44 , 表 6.46, 6.48, 表 6.50 では 実験 1 と同様に, 中間層と深い層でコンフリクトが大きくなっている. また 31 層は上位 15 層の中に全て入っておりロールプレイで出力のスタイルを決めるにあたって重要な層であると考えられる.

表 6.42: 実験 3 における SFT を適用したモデルとベースモデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値 (太字は上位 15 層)

	$M_{\mathrm{SFT_{D0}}}$	$M_{\mathrm{SFT_{D'1}}}$	$M_{\mathrm{SFT_{D'2}}}$	$M_{\mathrm{DPO_{D0}(M_{SFT_{D'2}})}}$	$M_{\mathrm{DPO_{D'1}(M_{SFT_{D'2}})}}$	$M_{\mathrm{DPO_{D0}(M_{SFT_{D'1}})}}$	$M_{\mathrm{DPO_{D'1}(M_{SFT_{D'1}})}}$
model.layers.0	0.00000	0.00000	0.02826	0.00000	0.00000	0.00000	0.00000
model.layers.1	0.07586	0.06351	0.05410	0.12474	0.08994	0.09558	0.04653
model.layers.2	0.04013	0.03030	0.00383	0.15858	0.09399	0.05838	0.04788
model.layers.3	0.05578	0.04025	0.00000	0.19519	0.09610	0.09241	0.06366
model.layers.4	0.08896	0.10557	0.05658	0.29127	0.17524	0.20339	0.15289
model.layers.5	0.13757	0.20152	0.11545	0.40297	0.27180	0.24551	0.21944
model.layers.6	0.20961	0.23267	0.20543	0.37640	0.28955	0.29036	0.27865
model.layers.7	0.23742	0.32739	0.26823	**0.53254**	0.38434	0.40243	0.38982
model.layers.8	0.30986	0.39383	0.29661	**0.56507**	0.45836	0.42257	0.44468
model.layers.9	**0.39776**	**0.47791**	0.37487	**0.62850**	**0.53339**	**0.47049**	**0.52902**
model.layers.10	**0.35173**	0.37379	0.32012	**0.53177**	0.49406	0.43289	**0.51549**
model.layers.11	**0.59310**	**0.67629**	**0.61235**	**0.84069**	**0.73934**	**0.75244**	**0.79164**
model.layers.12	**0.49295**	**0.52818**	**0.51827**	**0.69386**	**0.61568**	**0.65563**	**0.69849**
model.layers.13	**0.61135**	**0.59379**	**0.59754**	**0.73172**	**0.64630**	**0.67229**	**0.73625**
model.layers.14	**0.64909**	**0.56592**	**0.62223**	**0.68055**	**0.60919**	**0.65659**	**0.73449**
model.layers.15	**0.59408**	**0.57232**	**0.56745**	**0.57602**	**0.58062**	**0.54386**	**0.60450**
model.layers.16	**0.49195**	**0.50698**	**0.50410**	**0.54497**	**0.55813**	**0.51145**	**0.58448**
model.layers.17	**0.40870**	0.39459	**0.44576**	0.44394	0.47496	**0.46643**	**0.53680**
model.layers.18	0.31287	0.34541	0.32673	0.33006	0.41923	0.31676	0.37552
model.layers.19	0.34318	0.40426	**0.44280**	0.40720	0.49759	**0.48056**	**0.52211**
model.layers.20	0.30156	0.33066	0.27795	0.29570	0.44599	0.28819	0.34366
model.layers.21	0.25738	0.33722	0.29134	0.33258	0.45555	0.31730	0.37224
model.layers.22	0.21964	0.27235	0.25935	0.22403	0.43035	0.22073	0.32004
model.layers.23	0.25197	0.26270	0.28585	0.24086	0.41956	0.27936	0.34752
model.layers.24	0.27793	**0.42211**	0.42450	0.39520	**0.55669**	**0.46367**	0.51483
model.layers.25	**0.40323**	**0.52440**	**0.52030**	**0.47802**	**0.58527**	**0.54204**	**0.57365**
model.layers.26	**0.45556**	**0.51614**	**0.43971**	**0.47392**	**0.62961**	0.41743	0.47955
model.layers.27	**0.47141**	**0.48843**	**0.50290**	0.44578	**0.63247**	**0.44702**	**0.52253**
model.layers.28	0.31217	**0.49352**	**0.43223**	0.41580	**0.57266**	0.37286	0.43531
model.layers.29	**0.50521**	**0.55143**	**0.51241**	**0.52688**	**0.69684**	**0.51699**	**0.53784**
model.layers.30	**0.66863**	**0.67597**	**0.69767**	**0.63553**	**0.74269**	**0.65667**	**0.68248**
model.layers.31	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**	**1.00000**

表 6.43: 実験 3 における DPO を適用後モデルと DPO を適用前モデルとのタスクベクトルにおける絶対値上位 20 % の要素の層ごとの L2 ノルムを全層について最大値を 1, 最小値を 0 として正規化した値 (太字は上位 15 層)

	$M_{\mathrm{DPO}_{\mathrm{D0}}}$	$M_{\mathrm{DPO}_{\mathrm{D'1}}}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$	$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$	$M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$
model.layers.0	0.00000	0.00000	0.01321	0.00000	0.21019	0.00000
model.layers.1	**0.58690**	0.32721	**1.00000**	0.27454	**0.81233**	0.25190
model.layers.2	0.27981	0.21593	0.29394	0.18687	**0.60170**	0.15986
model.layers.3	0.49027	0.55722	**0.44027**	0.28840	**0.78035**	0.12156
model.layers.4	0.51686	0.54939	**0.67563**	0.33085	**0.70940**	0.21447
model.layers.5	0.48598	0.47646	**0.63211**	0.42111	**0.89664**	0.29611
model.layers.6	0.24181	0.21474	0.33177	0.20247	0.55728	0.21159
model.layers.7	0.57848	0.60455	**0.71852**	**0.49169**	**0.97901**	0.33551
model.layers.8	**0.58430**	0.57565	**0.73333**	**0.64929**	**0.85640**	0.42715
model.layers.9	**0.73110**	**0.77614**	**0.76162**	**0.88972**	**1.00000**	0.61209
model.layers.10	0.51179	0.55743	**0.60850**	**0.79518**	**0.71747**	0.54746
model.layers.11	**0.86726**	**0.66790**	**0.86799**	**1.00000**	**0.92057**	**0.75139**
model.layers.12	**0.78262**	**0.80457**	**0.65700**	**0.86098**	**0.78073**	**0.69810**
model.layers.13	**0.90105**	**0.76544**	**0.59012**	**0.84150**	**0.89406**	0.60749
model.layers.14	**0.92708**	**1.00000**	**0.63730**	**0.99779**	**0.98059**	**0.67863**
model.layers.15	0.41444	**0.89393**	0.25123	**0.62008**	0.43797	0.57817
model.layers.16	0.49249	**0.96616**	0.14430	**0.60659**	0.48910	**0.65907**
model.layers.17	**0.61214**	**0.89074**	**0.43332**	**0.65697**	**0.56594**	0.60864
model.layers.18	0.12567	0.38245	0.06916	0.33712	0.16408	0.44311
model.layers.19	**0.65860**	**0.87631**	0.38718	**0.57269**	0.28817	0.63574
model.layers.20	0.23981	**0.68862**	0.21674	0.38859	0.08034	0.62382
model.layers.21	0.51388	0.52155	0.22389	0.46474	0.23246	**0.67918**
model.layers.22	0.22013	0.61318	0.00000	0.29447	0.00000	**0.65867**
model.layers.23	0.10537	0.44222	0.26438	0.37416	0.17827	**0.68682**
model.layers.24	0.42151	0.52456	**0.41504**	**0.59197**	0.20783	**0.93005**
model.layers.25	**0.76882**	0.58854	**0.65796**	**0.65121**	0.27432	**0.75676**
model.layers.26	**0.90236**	**0.76250**	0.28745	**0.56193**	0.18183	**1.00000**
model.layers.27	**1.00000**	**0.77132**	0.14292	0.36566	0.10277	**0.97628**
model.layers.28	0.32216	**0.66483**	0.09721	0.20637	0.02925	**0.63730**
model.layers.29	**0.78327**	0.51881	0.21168	0.33332	0.14813	**0.96217**
model.layers.30	**0.70572**	**0.69028**	0.11260	0.22756	0.20110	**0.79977**
model.layers.31	**0.76258**	**0.65641**	0.18787	0.29973	**0.58197**	**0.85815**
0.49060						

表 6.44: 実験 3 における $M_{\mathrm{SFT}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}'1}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	534	2.4483e-06	0.0091022	0.00000	0.00000
model.layers.1	1852	8.4911e-06	0.01767	0.05389	0.14815
model.layers.2	1107	5.0754e-06	0.01296	0.02343	0.06677
model.layers.3	1422	6.5196e-06	0.01517	0.03631	0.10485
model.layers.4	1185	5.433e-06	0.01380	0.02662	0.08117
model.layers.5	2604	1.1939e-05	0.02031	0.08464	0.19375
model.layers.6	2613	1.198e-05	0.02031	0.08501	0.19377
model.layers.7	4788	2.1952e-05	0.02801	0.17395	0.32691
model.layers.8	5320	2.4391e-05	0.02927	0.19570	0.34862
model.layers.9	**9840**	**4.5114e-05**	0.03999	**0.38052**	0.53388
model.layers.10	3287	1.507e-05	0.02294	0.11257	0.23913
model.layers.11	**20417**	**9.3608e-05**	**0.05854**	**0.81301**	**0.85457**
model.layers.12	8004	3.6697e-05	0.03573	0.30545	0.46036
model.layers.13	**15461**	**7.0886e-05**	**0.05051**	**0.61036**	**0.71573**
model.layers.14	**19567**	**8.9711e-05**	**0.05751**	**0.77825**	**0.83677**
model.layers.15	**22658**	**0.00010388**	**0.06223**	**0.90465**	**0.91842**
model.layers.16	**20035**	**9.1856e-05**	**0.05777**	**0.79739**	**0.84118**
model.layers.17	**12410**	**5.6897e-05**	**0.04506**	**0.48561**	**0.62153**
model.layers.18	9646	4.4225e-05	0.04022	0.37259	0.53791
model.layers.19	**10778**	**4.9415e-05**	**0.04247**	**0.41887**	**0.57677**
model.layers.20	9778	4.483e-05	**0.04079**	0.37798	**0.54773**
model.layers.21	5682	2.6051e-05	0.03051	0.21050	0.37013
model.layers.22	5624	2.5785e-05	0.02993	0.20813	0.36009
model.layers.23	4195	1.9233e-05	0.02600	0.14970	0.29204
model.layers.24	7531	3.4528e-05	0.03490	0.28611	0.44598
model.layers.25	**16765**	**7.6864e-05**	**0.05351**	**0.66368**	**0.76755**
model.layers.26	**17515**	**8.0303e-05**	**0.05454**	**0.69435**	**0.78539**
model.layers.27	**17466**	**8.0078e-05**	**0.05360**	**0.69235**	**0.76921**
model.layers.28	**12784**	**5.8612e-05**	**0.04607**	**0.50090**	**0.63902**
model.layers.29	**12416**	**5.6925e-05**	**0.04530**	**0.48585**	**0.62565**
model.layers.30	**18063**	**8.2815e-05**	**0.05535**	**0.71676**	**0.79947**
model.layers.31	**24990**	**0.00011457**	**0.06695**	**1.00000**	**1.00000**

表 6.45: 実験 3 における $M_{\mathrm{SFT}_{\mathrm{D}0}}$, $M_{\mathrm{SFT}_{\mathrm{D}'1}}$ 間の各パラメータのコンフリクト率の値

パラメータ	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	4.9472e-06	**2.0266e-05**	2.3842e-06	7.3314e-06	2.0095e-06	1.9584e-06	0.0000e+00
model.layer.1	**4.8637e-05**	1.6451e-05	2.5511e-05	7.0333e-06	1.5668e-06	3.8317e-06	7.2377e-06
model.layer.2	7.5698e-06	**1.0467e-04**	1.1921e-05	5.6028e-06	4.1383e-06	2.6226e-06	0.0000e+00
model.layer.3	8.2850e-06	4.6253e-05	**7.2241e-05**	7.5102e-06	6.2159e-06	5.0238e-06	0.0000e+00
model.layer.4	9.2983e-06	**6.0320e-05**	7.6294e-06	2.0266e-06	8.3787e-06	3.7125e-06	0.0000e+00
model.layer.5	5.1856e-06	**1.1826e-04**	8.5831e-05	4.6730e-05	7.3910e-06	7.5442e-06	0.0000e+00
model.layer.6	2.9802e-06	**8.0824e-05**	1.7166e-05	1.8954e-05	1.8086e-05	1.3147e-05	0.0000e+00
model.layer.7	6.8545e-06	**2.6655e-04**	1.0252e-05	5.7161e-05	2.4983e-05	1.8494e-05	0.0000e+00
model.layer.8	1.5259e-05	4.0054e-05	6.2943e-05	**6.7830e-05**	2.5306e-05	3.4111e-05	8.5149e-08
model.layer.9	1.5676e-05	1.4877e-04	3.5524e-05	**1.5253e-04**	2.0521e-05	8.5524e-05	3.0654e-07
model.layer.10	2.1100e-05	**2.3603e-05**	2.0266e-05	9.1791e-06	2.0095e-05	2.3212e-05	8.8555e-07
model.layer.11	1.7047e-05	9.2745e-05	1.6212e-05	1.4150e-04	9.8177e-05	**1.9644e-04**	0.0000e+00
model.layer.12	3.4869e-05	3.0041e-05	3.9816e-05	**1.6868e-04**	4.1349e-05	3.1795e-05	1.7030e-08
model.layer.13	1.9312e-05	1.1396e-04	9.2983e-05	**1.5461e-04**	1.3726e-04	5.8413e-05	3.1505e-06
model.layer.14	1.3292e-05	**3.4046e-04**	3.6001e-05	2.2006e-04	5.5194e-05	1.8431e-04	1.5327e-07
model.layer.15	2.1994e-05	**2.5177e-04**	5.4598e-05	1.6034e-04	2.2844e-04	8.3344e-05	1.0218e-07
model.layer.16	1.5736e-05	**1.9121e-04**	1.0800e-04	4.2439e-05	1.3312e-04	1.6877e-04	1.3113e-06
model.layer.17	2.3484e-05	**4.6253e-04**	4.1723e-05	1.3864e-04	3.0960e-05	9.8041e-05	0.0000e+00
model.layer.18	2.9087e-05	**9.1815e-04**	2.7895e-05	5.1320e-05	4.6985e-05	2.6720e-05	1.7030e-08
model.layer.19	9.0241e-05	**5.8985e-04**	7.8678e-05	1.0586e-04	4.2081e-05	3.7687e-05	0.0000e+00
model.layer.20	2.8074e-05	**1.0700e-03**	5.8413e-05	8.0824e-05	2.7980e-05	2.6345e-05	2.7248e-07
model.layer.21	2.0146e-05	**3.5906e-04**	6.5327e-05	2.0623e-05	2.3076e-05	3.1727e-05	0.0000e+00
model.layer.22	3.6776e-05	**3.2091e-04**	1.9550e-05	3.7968e-05	2.2650e-05	2.7384e-05	6.8120e-08
model.layer.23	1.8179e-05	**1.3947e-04**	7.0095e-05	1.4067e-05	1.5974e-05	3.1284e-05	0.0000e+00
model.layer.24	4.0531e-05	**1.7524e-04**	8.2016e-05	5.4061e-05	5.0698e-05	3.2101e-05	5.1090e-08
model.layer.25	1.0610e-04	1.1253e-04	1.0538e-04	**1.7780e-04**	7.0282e-05	1.1855e-04	0.0000e+00
model.layer.26	3.6001e-05	**2.8610e-04**	2.5201e-04	6.8963e-05	2.0691e-05	2.0911e-04	5.1090e-08
model.layer.27	4.2439e-05	**3.9148e-04**	3.0994e-05	1.8215e-04	3.2800e-05	1.7030e-04	0.0000e+00
model.layer.28	3.0160e-05	**1.9622e-04**	1.3351e-05	1.1086e-04	1.3278e-04	2.9666e-05	0.0000e+00
model.layer.29	5.3048e-06	6.4373e-05	4.7207e-05	7.1347e-05	5.9298e-05	**1.2226e-04**	1.7030e-08
model.layer.30	3.9279e-05	**1.7309e-04**	5.6982e-05	1.0878e-04	1.6493e-04	7.4897e-05	9.0429e-06
model.layer.31	9.2447e-05	**4.7541e-04**	9.2030e-05	1.4389e-04	1.9859e-04	1.1740e-04	1.5327e-06

表 6.46: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{DPO}_{\mathrm{D}'1}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	587	2.6913e-06	0.0098477	0.09418	0.00000
model.layers.1	2990	1.3709e-05	0.02273	0.47970	0.59544
model.layers.2	1862	8.5369e-06	0.01719	0.29873	0.33951
model.layers.3	**3870**	**1.7743e-05**	0.02441	**0.62089**	0.67343
model.layers.4	**5365**	**2.4597e-05**	**0.02976**	**0.86074**	**0.92066**
model.layers.5	**4107**	**1.883e-05**	**0.02530**	**0.65891**	**0.71423**
model.layers.6	1824	8.3627e-06	0.01777	0.29264	0.36631
model.layers.7	**3983**	**1.8261e-05**	0.02402	**0.63902**	0.65510
model.layers.8	3414	1.5653e-05	0.02269	0.54773	0.59392
model.layers.9	**6017**	**2.7587e-05**	**0.03039**	**0.96535**	**0.94966**
model.layers.10	2846	1.3048e-05	0.02139	0.45660	0.53359
model.layers.11	3783	1.7344e-05	0.02432	0.60693	0.66913
model.layers.12	**4817**	**2.2085e-05**	**0.02727**	**0.77282**	**0.80540**
model.layers.13	**4392**	**2.0136e-05**	**0.02644**	**0.70464**	**0.76707**
model.layers.14	**6053**	**2.7752e-05**	**0.02950**	**0.97112**	**0.90857**
model.layers.15	**4109**	**1.8839e-05**	**0.02536**	**0.65923**	**0.71711**
model.layers.16	**5266**	**2.4144e-05**	**0.02840**	**0.84486**	**0.85764**
model.layers.17	**6233**	**2.8577e-05**	**0.03116**	**1.00000**	**0.98539**
model.layers.18	1803	8.2664e-06	0.01632	0.28927	0.29938
model.layers.19	3833	1.7574e-05	**0.02525**	0.61495	**0.71227**
model.layers.20	2797	1.2824e-05	0.02153	0.44874	0.53995
model.layers.21	2763	1.2668e-05	0.02053	0.44329	0.49375
model.layers.22	2140	9.8115e-06	0.01844	0.34333	0.39728
model.layers.23	1355	6.2124e-06	0.01519	0.21739	0.24690
model.layers.24	1509	6.9185e-06	0.01579	0.24210	0.27484
model.layers.25	**4487**	**2.0572e-05**	**0.02803**	**0.71988**	**0.84044**
model.layers.26	**5245**	**2.4047e-05**	**0.03148**	**0.84149**	**1.00000**
model.layers.27	3632	1.6652e-05	**0.02463**	0.58270	**0.68332**
model.layers.28	2991	1.3713e-05	0.02223	0.47987	0.57245
model.layers.29	**4311**	**1.9765e-05**	**0.02668**	**0.69164**	**0.77818**
model.layers.30	2635	1.2081e-05	0.02170	0.42275	0.54779
model.layers.31	**5777**	**2.6486e-05**	**0.03093**	**0.92684**	**0.97496**

表 6.47: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D}0}}$, $M_{\mathrm{DPO}_{\mathrm{D}'1}}$ 間の各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	1.0729e-05	**2.6703e-05**	8.8215e-06	4.2915e-06	1.4986e-06	1.6689e-06	0.0000e+00
model.layer.1	**3.9220e-05**	3.5048e-05	2.7895e-05	1.6093e-06	1.5327e-06	4.0191e-06	2.9206e-05
model.layer.2	2.8491e-05	**6.9857e-05**	2.7418e-05	1.7285e-05	5.1771e-06	6.5054e-06	0.0000e+00
model.layer.3	5.8055e-05	**2.7537e-04**	3.5524e-05	5.3644e-06	3.5422e-06	2.2037e-05	0.0000e+00
model.layer.4	**1.3059e-04**	5.0545e-05	5.2452e-05	5.5432e-06	2.3041e-05	2.2071e-05	0.0000e+00
model.layer.5	4.3571e-05	**1.2350e-04**	4.5061e-05	1.0967e-05	2.5766e-05	1.6553e-05	0.0000e+00
model.layer.6	1.8179e-05	6.9857e-05	**1.2445e-04**	7.8678e-06	6.1819e-06	3.5592e-06	0.0000e+00
model.layer.7	7.1526e-06	1.6689e-05	3.5524e-05	**4.6968e-05**	1.3181e-05	3.5456e-05	0.0000e+00
model.layer.8	7.2122e-06	**6.9857e-05**	3.0279e-05	1.9550e-05	2.6464e-05	1.6877e-05	0.0000e+00
model.layer.9	8.9407e-06	9.4891e-05	8.0109e-05	**2.0164e-04**	2.6481e-05	3.3208e-06	0.0000e+00
model.layer.10	2.8551e-05	**1.3661e-04**	5.3883e-05	2.0027e-05	1.4322e-05	6.6587e-06	0.0000e+00
model.layer.11	3.3021e-05	**6.5804e-05**	4.2677e-05	3.0100e-05	2.6771e-05	1.1870e-05	0.0000e+00
model.layer.12	2.8610e-05	**5.8174e-05**	2.7895e-05	2.2531e-05	4.8007e-05	1.3266e-05	0.0000e+00
model.layer.13	6.3837e-05	**1.1754e-04**	1.0705e-04	2.7716e-05	1.2585e-05	2.0010e-05	0.0000e+00
model.layer.14	1.1981e-05	**1.5640e-04**	5.5790e-05	8.4996e-05	1.1751e-05	4.8467e-05	0.0000e+00
model.layer.15	1.8656e-05	**1.8048e-04**	1.1468e-04	6.4850e-05	1.4169e-05	1.0865e-05	0.0000e+00
model.layer.16	3.5107e-05	**5.8174e-05**	1.7643e-05	5.3644e-06	2.2531e-05	5.0170e-05	0.0000e+00
model.layer.17	8.3029e-05	**2.9516e-04**	2.1219e-05	6.2346e-05	2.5306e-05	1.6706e-05	0.0000e+00
model.layer.18	2.0266e-06	**2.9302e-04**	1.4782e-05	2.4438e-06	4.4107e-06	3.0313e-06	0.0000e+00
model.layer.19	3.8087e-05	1.0800e-04	**2.1863e-04**	2.3186e-05	5.0068e-06	1.9431e-05	0.0000e+00
model.layer.20	5.8055e-05	**3.3259e-04**	2.0504e-05	7.1526e-07	2.8610e-06	2.7588e-06	0.0000e+00
model.layer.21	1.5378e-05	**1.1683e-04**	2.3603e-05	5.6446e-05	1.1103e-05	5.3985e-06	0.0000e+00
model.layer.22	2.8551e-05	**8.4877e-05**	5.9605e-05	1.7643e-05	4.1383e-06	8.7874e-06	0.0000e+00
model.layer.23	1.6510e-05	**1.2064e-04**	5.5313e-05	2.6226e-06	4.2064e-06	8.3447e-07	0.0000e+00
model.layer.24	7.9274e-06	2.1458e-05	**9.2983e-05**	1.1206e-05	9.8773e-07	1.1069e-05	0.0000e+00
model.layer.25	5.6446e-05	**2.8110e-04**	2.3556e-04	1.1027e-05	1.4271e-05	5.9605e-06	0.0000e+00
model.layer.26	2.8253e-05	6.6757e-05	**6.8808e-04**	6.6757e-06	2.0725e-05	4.7003e-06	0.0000e+00
model.layer.27	1.0216e-04	**1.9646e-04**	4.3631e-05	8.3447e-06	8.4979e-06	4.6321e-06	0.0000e+00
model.layer.28	3.3557e-05	2.5988e-05	**2.6226e-04**	9.8348e-06	4.1383e-06	1.3811e-05	0.0000e+00
model.layer.29	5.9307e-05	**2.3389e-04**	4.2200e-05	1.1921e-06	1.7217e-05	1.9193e-05	0.0000e+00
model.layer.30	2.0206e-05	**2.2984e-04**	6.3419e-05	2.0504e-05	9.0258e-06	3.2697e-06	0.0000e+00
model.layer.31	2.1636e-05	**3.0088e-04**	2.4128e-04	1.4174e-04	5.4325e-06	7.5442e-06	0.0000e+00

表 6.48: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$ 間の M_{base} から計算したタスクベクトルの コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	2548	1.1682e-05	0.02152	0.01585	0.02270
model.layers.1	**13101**	**6.0065e-05**	**0.05281**	**0.89592**	**1.00000**
model.layers.2	3697	1.695e-05	0.02473	0.11167	0.12307
model.layers.3	7292	3.3432e-05	0.03482	0.41148	0.43808
model.layers.4	2701	1.2384e-05	0.02273	0.02860	0.06035
model.layers.5	2989	1.3704e-05	0.02373	0.05262	0.09173
model.layers.6	3847	1.7638e-05	0.02549	0.12418	0.14661
model.layers.7	**8092**	**3.71e-05**	**0.03606**	**0.47819**	**0.47685**
model.layers.8	**8721**	**3.9984e-05**	**0.03675**	**0.53065**	**0.49842**
model.layers.9	**14349**	**6.5787e-05**	**0.04849**	**1.00000**	**0.86507**
model.layers.10	**9233**	**4.2331e-05**	**0.03943**	**0.57335**	**0.58199**
model.layers.11	**11723**	**5.3748e-05**	**0.04369**	**0.78100**	**0.71494**
model.layers.12	**9239**	**4.2359e-05**	**0.03911**	**0.57385**	**0.57208**
model.layers.13	**9302**	**4.2648e-05**	**0.03986**	**0.57910**	**0.59542**
model.layers.14	**11188**	**5.1295e-05**	**0.04251**	**0.73639**	**0.67812**
model.layers.15	4529	2.0765e-05	0.02744	0.18105	0.20763
model.layers.16	3945	1.8087e-05	0.02564	0.13235	0.15141
model.layers.17	6972	3.1965e-05	0.03338	0.38479	0.39316
model.layers.18	4389	2.0123e-05	0.02756	0.16938	0.21125
model.layers.19	**8547**	**3.9186e-05**	**0.03898**	**0.51614**	**0.56798**
model.layers.20	3842	1.7615e-05	0.02520	0.12376	0.13768
model.layers.21	5914	2.7115e-05	0.03117	0.29656	0.32421
model.layers.22	2592	1.1884e-05	0.02180	0.01951	0.03151
model.layers.23	3540	1.623e-05	0.02463	0.09857	0.11994
model.layers.24	**7823**	**3.5867e-05**	**0.03653**	**0.45576**	**0.49136**
model.layers.25	**14070**	**6.4508e-05**	**0.04997**	**0.97673**	**0.91109**
model.layers.26	**8486**	**3.8907e-05**	**0.03937**	**0.51105**	**0.58024**
model.layers.27	3478	1.5946e-05	0.02490	0.09340	0.12814
model.layers.28	2358	1.0811e-05	0.02079	0.00000	0.00000
model.layers.29	2700	1.2379e-05	0.02201	0.02852	0.03801
model.layers.30	**8739**	**4.0067e-05**	**0.03857**	**0.53215**	**0.55521**
model.layers.31	**9564**	**4.3849e-05**	**0.04002**	**0.60095**	**0.60062**

表 6.49: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$, $M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$ 間の M_{base} から計算したタスクベクトルの各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	4.9233e-05	6.1989e-05	**0.00019646**	1.7405e-05	4.1383e-06	1.7541e-06	0.00000
model.layer.1	0.00011539	4.3392e-05	7.2479e-05	1.8477e-06	3.0313e-06	6.3522e-06	**0.00017195**
model.layer.2	5.4538e-05	**8.3208e-05**	5.722e-05	5.579e-05	8.8726e-06	1.2534e-05	0.00000
model.layer.3	8.6665e-05	**0.00045681**	0.0001142	2.1517e-05	1.1683e-05	4.0804e-05	0.00000
model.layer.4	0.00012153	0.00013375	**0.0001905**	1.651e-05	4.0395e-05	4.5827e-05	0.00000
model.layer.5	0.00014281	**0.00026846**	9.4652e-05	4.065e-05	5.3338e-05	3.1182e-05	0.00000
model.layer.6	2.3842e-05	**0.00017381**	0.00012326	2.6405e-05	2.0657e-05	9.2813e-06	0.00000
model.layer.7	2.5034e-05	7.6294e-05	**0.00017881**	0.00012714	2.7146e-05	4.8961e-05	0.00000
model.layer.8	1.4782e-05	**0.0001204**	4.9829e-05	0.00010031	4.3426e-05	6.0047e-05	0.00000
model.layer.9	2.3365e-05	0.00026608	0.00020409	**0.00042737**	6.0643e-05	2.1355e-05	0.00000
model.layer.10	7.54e-05	**0.00030017**	0.00026941	0.00015265	2.8321e-05	2.3058e-05	1.703e-08
model.layer.11	4.9412e-05	**0.00021052**	0.00012016	8.4281e-05	9.2234e-05	4.5572e-05	1.703e-08
model.layer.12	9.1553e-05	**0.00018454**	7.4625e-05	3.7611e-05	7.6447e-05	2.5477e-05	0.00000
model.layer.13	9.9778e-05	0.00027204	**0.00031948**	7.5042e-05	3.062e-05	3.5592e-05	0.00000
model.layer.14	0.00014734	**0.00019622**	7.3433e-05	0.00011694	1.9261e-05	7.6498e-05	0.00000
model.layer.15	6.5088e-05	**0.00026202**	4.673e-05	3.2902e-05	9.1961e-06	1.7881e-05	0.00000
model.layer.16	2.3901e-05	**9.6321e-05**	4.673e-05	6.0797e-06	3.2493e-05	1.5906e-05	0.00000
model.layer.17	5.0128e-05	**0.0004015**	1.3113e-05	4.6194e-05	2.5528e-05	3.6052e-05	1.703e-08
model.layer.18	1.5378e-05	**0.0002737**	0.00018692	6.3777e-06	2.7571e-05	8.0551e-06	0.00000
model.layer.19	9.197e-05	0.00035691	**0.00056672**	6.9141e-06	1.6723e-05	3.4605e-05	0.00000
model.layer.20	5.6267e-05	**0.00017905**	3.8147e-06	5.126e-06	2.5749e-05	9.0769e-06	0.00000
model.layer.21	2.3365e-05	**0.00015283**	1.3828e-05	4.7088e-06	6.112e-05	1.9584e-05	8.5149e-08
model.layer.22	2.7895e-05	5.0545e-05	**0.00015044**	7.2122e-06	1.5957e-05	3.7977e-06	0.00000
model.layer.23	4.5538e-05	**0.00019026**	0.00011611	3.0398e-05	3.968e-06	1.2738e-05	0.00000
model.layer.24	0.00019056	3.5048e-05	**0.00023508**	0.00013012	2.861e-06	1.9448e-05	0.00000
model.layer.25	**0.00047141**	0.00032806	0.00026679	0.00010729	2.58e-05	5.9775e-06	0.00000
model.layer.26	**0.00030684**	0.00021887	0.00020003	2.3842e-06	2.3331e-05	2.9121e-06	0.00000
model.layer.27	7.0632e-05	**0.00037217**	6.0558e-05	2.2054e-06	4.2745e-06	3.2357e-06	0.00000
model.layer.28	2.9027e-05	1.9312e-05	**0.00025153**	2.0266e-06	6.1137e-06	5.8242e-06	0.00000
model.layer.29	**7.4089e-05**	7.2956e-05	1.9789e-05	1.0312e-05	9.0599e-06	6.1819e-06	0.00000
model.layer.30	5.1975e-05	7.9155e-05	**0.00012064**	2.8312e-05	5.5177e-06	3.2697e-06	0.00000
model.layer.31	1.7107e-05	0.00014663	**0.00015974**	3.7849e-05	8.0381e-06	5.2282e-06	5.109e-08

表 6.50: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$ 間の M_{base} から計算したタスクベクトルの コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 正規化したコンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	3225	1.4786e-05	0.02208	0.00000	0.00000
model.layers.1	13804	6.3289e-05	0.04975	0.27228	0.45729
model.layers.2	10130	4.6444e-05	0.03984	0.17772	0.29359
model.layers.3	7113	3.2612e-05	0.03291	0.10007	0.17898
model.layers.4	10519	4.8228e-05	0.04001	0.18773	0.29634
model.layers.5	**20641**	**9.4635e-05**	**0.05714**	**0.44825**	**0.57941**
model.layers.6	11023	5.0538e-05	0.04100	0.20071	0.31277
model.layers.7	19797	9.0765e-05	0.05525	0.42653	0.54820
model.layers.8	**22968**	**0.0001053**	**0.05951**	**0.50815**	**0.61867**
model.layers.9	**29092**	**0.00013338**	**0.06814**	**0.66577**	**0.76126**
model.layers.10	**26048**	**0.00011942**	**0.06413**	**0.58742**	**0.69503**
model.layers.11	**34067**	**0.00015619**	**0.07330**	**0.79381**	**0.84661**
model.layers.12	**28933**	**0.00013265**	**0.06843**	**0.66167**	**0.76609**
model.layers.13	**31545**	**0.00014463**	**0.07116**	**0.72890**	**0.81124**
model.layers.14	**42078**	**0.00019292**	**0.08258**	**1.00000**	**1.00000**
model.layers.15	**21395**	**9.8092e-05**	**0.05863**	**0.46766**	**0.60418**
model.layers.16	**26676**	**0.0001223**	**0.06587**	**0.60358**	**0.72368**
model.layers.17	**25242**	**0.00011573**	**0.06349**	**0.56667**	**0.68442**
model.layers.18	**24024**	**0.00011015**	**0.06226**	**0.53533**	**0.66409**
model.layers.19	19526	8.9523e-05	0.05625	0.41956	0.56469
model.layers.20	11881	5.4472e-05	0.04375	0.22279	0.35811
model.layers.21	18308	8.3939e-05	0.05409	0.38821	0.52901
model.layers.22	9020	4.1355e-05	0.03761	0.14915	0.25665
model.layers.23	15878	7.2797e-05	0.05097	0.32566	0.47748
model.layers.24	**20818**	**9.5446e-05**	**0.05769**	**0.45281**	**0.58865**
model.layers.25	17286	7.9253e-05	0.05302	0.36190	0.51131
model.layers.26	18651	8.5511e-05	0.05549	0.39703	0.55222
model.layers.27	16112	7.387e-05	0.05091	0.33169	0.47658
model.layers.28	8893	4.0773e-05	0.03752	0.14588	0.25529
model.layers.29	16136	7.398e-05	0.05136	0.33230	0.48400
model.layers.30	**19997**	**9.1682e-05**	**0.05762**	**0.43168**	**0.58737**
model.layers.31	**20384**	**9.3457e-05**	**0.06594**	**0.44164**	**0.72484**

表 6.51: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$ 間の M_{base} から計算したタスクベクトルの各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	4.4644e-05	**1.2374e-04**	5.0306e-05	3.2723e-05	1.0252e-05	1.0133e-05	0.0000e+00
model.layer.1	**2.1225e-04**	1.4949e-04	1.8096e-04	2.1160e-05	1.2517e-05	1.7047e-05	1.1522e-04
model.layer.2	1.3179e-04	1.8263e-04	1.0586e-04	**1.8948e-04**	2.4131e-05	3.5984e-05	0.0000e+00
model.layer.3	5.3227e-05	**3.6740e-04**	6.8188e-05	6.4731e-05	2.5136e-05	3.1182e-05	0.0000e+00
model.layer.4	5.8472e-05	**2.1887e-04**	1.0252e-04	4.1306e-05	5.0681e-05	7.6992e-05	0.0000e+00
model.layer.5	1.6445e-04	**3.7003e-04**	1.2350e-04	1.6010e-04	1.3517e-04	8.8334e-05	3.4060e-08
model.layer.6	3.0935e-05	**3.1447e-04**	3.6716e-05	8.4341e-05	7.5953e-05	5.3746e-05	0.0000e+00
model.layer.7	6.0856e-05	**2.4557e-04**	1.2779e-04	2.2465e-04	9.3034e-05	1.3586e-04	0.0000e+00
model.layer.8	6.2406e-05	**3.9530e-04**	8.9884e-05	1.5390e-04	1.1652e-04	1.7813e-04	3.4060e-08
model.layer.9	5.0962e-05	7.2336e-04	1.4210e-04	**8.3482e-04**	1.0468e-04	7.5834e-05	1.7030e-08
model.layer.10	8.4281e-05	**8.3590e-04**	1.7524e-04	3.3557e-04	1.3084e-04	1.1906e-04	1.5157e-06
model.layer.11	1.3006e-04	**6.0344e-04**	3.7384e-04	3.5959e-04	1.5119e-04	2.1906e-04	2.0436e-07
model.layer.12	2.2370e-04	**7.4315e-04**	2.1386e-04	1.7059e-04	2.0368e-04	1.0794e-04	1.0218e-07
model.layer.13	2.7496e-04	**7.4983e-04**	5.0044e-04	2.8610e-04	7.4727e-05	2.1281e-04	6.8120e-08
model.layer.14	**4.9871e-04**	4.2439e-04	1.3280e-04	3.9232e-04	9.3409e-05	3.2861e-04	1.8733e-07
model.layer.15	1.1712e-04	**6.9213e-04**	2.1791e-04	3.3474e-04	7.7912e-05	9.2234e-05	1.0218e-07
model.layer.16	7.2777e-05	2.3627e-04	9.3699e-05	2.0283e-04	**2.6092e-04**	9.1008e-05	5.1090e-08
model.layer.17	1.4222e-04	**1.0500e-03**	2.6155e-04	5.5850e-04	3.2169e-05	1.0267e-04	1.4816e-06
model.layer.18	7.0870e-05	**5.2309e-04**	4.3559e-04	4.5365e-04	6.5565e-05	1.2480e-04	4.2575e-07
model.layer.19	2.2000e-04	3.6001e-04	**3.7098e-04**	4.4107e-05	6.0133e-05	1.4472e-04	0.0000e+00
model.layer.20	9.6917e-05	**5.5265e-04**	7.2241e-05	1.5616e-05	8.5916e-05	3.9629e-05	0.0000e+00
model.layer.21	3.4153e-05	**2.8849e-04**	6.6757e-05	6.6221e-05	1.9911e-04	5.8532e-05	8.5149e-08
model.layer.22	8.6367e-05	**1.6952e-04**	7.5340e-05	4.9889e-05	6.7234e-05	2.9956e-05	0.0000e+00
model.layer.23	1.2612e-04	2.7084e-04	**6.6257e-04**	9.9659e-05	4.9744e-05	8.9441e-05	3.4060e-08
model.layer.24	1.2481e-04	3.2234e-04	2.0742e-04	**5.2625e-04**	2.4182e-05	1.0647e-04	1.7030e-08
model.layer.25	2.3550e-04	**4.6301e-04**	2.8038e-04	2.9814e-04	5.8464e-05	3.0330e-05	1.7030e-08
model.layer.26	1.7637e-04	**1.2700e-03**	1.1015e-04	9.1612e-05	1.1303e-04	2.9428e-05	0.0000e+00
model.layer.27	6.7770e-05	**5.7340e-04**	2.3389e-04	1.4341e-04	6.2312e-05	9.3903e-05	1.7030e-07
model.layer.28	9.6083e-05	**2.6131e-04**	9.4652e-05	9.2983e-05	2.8661e-05	4.3341e-05	0.0000e+00
model.layer.29	2.9182e-04	**4.0245e-04**	1.5640e-04	1.4663e-04	3.6052e-05	7.3075e-05	4.7684e-07
model.layer.30	3.1251e-04	3.9029e-04	2.0623e-04	**5.7036e-04**	1.7967e-05	2.7725e-05	0.0000e+00
model.layer.31	1.7518e-04	5.3048e-04	**1.0100e-03**	1.0067e-04	2.9751e-05	1.2784e-04	8.5149e-07

また, 表 6.45 , 表 6.47, 6.49, 表 6.51 において, 実験 1, 2 と共通してほとんどの層において v_proj のコンフリクト率が高くなっている. ロールプレイの演じ分けをするにあたって v_proj の重みが大きな意味をなしているといえる.

同じデータセットで手法が異なる $M_{\mathrm{DPO_{D0}}}$ と $M_{\mathrm{SFT_{D0}}}$, $M_{\mathrm{DPO_{D'1}}}$ と $M_{\mathrm{SFT_{D'1}}}$ に関する Conflict Limited L2 ノルム を調べる. 表 6.52 に $M_{\mathrm{DPO_{D0}}}$, $M_{\mathrm{SFT_{D0}}}$ 間の コンフリクト数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.53 に各層の各パラメータのコンフリクト率の値を示す.

同様に, 表 6.54 に $M_{\mathrm{DPO_{D'1}}}$, $M_{\mathrm{SFT_{D'1}}}$ 間の コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム, 全層に関して最大値が 1, 最小値が 0 となるように正規化したコンフリクト率, Conflict Limited L2 ノルム を示す. また, 表 6.55 に各層の各パラメータのコンフリクト率の値を示す.

表 6.52, 表 6.54 に共通して 25 から 31 層の深い層に加えて 9 から 17 層の中間層において Conflict Limited L2 ノルム の値が大きくなっている. データセットの違いに関して, 上記で述べたように D_0, $D_{'1}$ は一人称や語尾だけでなく文全体が違うスタイルになっている. 実験 1 の D_0, D_1 はデータセットの文は一人称や口調のみが異なり文中に出てくる固有名詞などはほとんど同じであったため実験 1 のような差は見られず, D_0, $D_{'1}$ どちらも一般的な応答をするロールプレイタスク, 男子大学生の応答をするロールプレイタスクと言うように独立した新しいスタイルの応答を学習するというタスク差となったため表 6.52, 表 6.54 が同じような特徴を示したと考えられる.

表 6.56 には, 実験 3 における $(M_{\mathrm{DPO_{D0}(M_{SFT_{D'1}})}}, M_{\mathrm{base}})$, $(M_{\mathrm{DPO_{D0}(M_{SFT_{D'1}})}}, M_{\mathrm{SFT_{D'1}}})$, $(M_{\mathrm{DPO_{D0}(M_{SFT_{D'1}})}}, M_{\mathrm{DPO_{D0}}})$ 間の L1 ノルム, L2 ノルム を示している. 実験 1, 2 により LLM の冗長性により似たようなタスクでも重みの回帰現象は発生しないと考えられる.

学習後のモデルの出力結果から LLM の性能を評価する. 表 6.57, 6.58 にそれぞれ実験 3 における学習後のモデルの D_0, D'_1 との BLEU, BERTScore の値を示す.

表 6.59 に LLM の評価結果を示す.

実験 1, 2 と同様に, 最後に D'_1 を学習に用いたモデルはロールプレイ性能が高くなり, 最後に D_0 を学習に用いたモデルはロールプレイ性能が低くなる傾向が見られ, $M_{\mathrm{DPO_{D'1}}}$ が最もロールプレイの性能が高くなった.

また, 表 6.60 に「このカフェ素敵ですね」に対する LLM の応答を示す.

実験 1 と同様に, 学習データ内の「雰囲気」という単語が多く見られる. また, 「ほ

表 6.52: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間 コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	241	1.1049e-06	0.00658	0.00000	0.0003839
model.layers.1	**4028**	**1.8468e-05**	**0.03061**	**0.42488**	**0.67841**
model.layers.0	534	2.4483e-06	0.0091022	0.00000	0.00000
model.layers.1	1852	8.4911e-06	0.01767	0.05389	0.14815
model.layers.2	1107	5.0754e-06	0.01296	0.02343	0.06677
model.layers.3	1422	6.5196e-06	0.01517	0.03631	0.10485
model.layers.4	1185	5.433e-06	0.01380	0.02662	0.08117
model.layers.5	2604	1.1939e-05	0.02031	0.08464	0.19375
model.layers.6	2613	1.198e-05	0.02031	0.08501	0.19377
model.layers.7	4788	2.1952e-05	0.02801	0.17395	0.32691
model.layers.8	5320	2.4391e-05	0.02927	0.19570	0.34862
model.layers.9	**9840**	**4.5114e-05**	0.03999	**0.38052**	0.53388
model.layers.10	3287	1.507e-05	0.02294	0.11257	0.23913
model.layers.11	**20417**	**9.3608e-05**	**0.05854**	**0.81301**	**0.85457**
model.layers.12	8004	3.6697e-05	0.03573	0.30545	0.46036
model.layers.13	**15461**	**7.0886e-05**	**0.05051**	**0.61036**	**0.71573**
model.layers.14	**19567**	**8.9711e-05**	**0.05751**	**0.77825**	**0.83677**
model.layers.15	**22658**	**0.00010388**	**0.06223**	**0.90465**	**0.91842**
model.layers.16	**20035**	**9.1856e-05**	**0.05777**	**0.79739**	**0.84118**
model.layers.17	**12410**	**5.6897e-05**	**0.04506**	**0.48561**	**0.62153**
model.layers.18	9646	4.4225e-05	0.04022	0.37259	0.53791
model.layers.19	**10778**	**4.9415e-05**	**0.04247**	**0.41887**	**0.57677**
model.layers.20	9778	4.483e-05	**0.04079**	0.37798	**0.54773**
model.layers.21	5682	2.6051e-05	0.03051	0.21050	0.37013
model.layers.22	5624	2.5785e-05	0.02993	0.20813	0.36009
model.layers.23	4195	1.9233e-05	0.02600	0.14970	0.29204
model.layers.24	7531	3.4528e-05	0.03490	0.28611	0.44598
model.layers.25	**16765**	**7.6864e-05**	**0.05351**	**0.66368**	**0.76755**
model.layers.26	**17515**	**8.0303e-05**	**0.05454**	**0.69435**	**0.78539**
model.layers.27	**17466**	**8.0078e-05**	**0.05360**	**0.69235**	**0.76921**
model.layers.28	**12784**	**5.8612e-05**	**0.04607**	**0.50090**	**0.63902**
model.layers.29	**12416**	**5.6925e-05**	**0.04530**	**0.48585**	**0.62565**
model.layers.30	**18063**	**8.2815e-05**	**0.05535**	**0.71676**	**0.79947**
model.layers.31	**24990**	**0.00011457**	**0.06695**	**1.00000**	**1.00000**

表 6.53: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D0}}}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$ 間の各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	4.9472e-06	**2.0266e-05**	2.3842e-06	7.3314e-06	2.0095e-06	1.9584e-06	0.0000e+00
model.layer.1	**4.8637e-05**	1.6451e-05	2.5511e-05	7.0333e-06	1.5668e-06	3.8317e-06	7.2377e-06
model.layer.2	7.5698e-06	**1.0467e-04**	1.1921e-05	5.6028e-06	4.1383e-06	2.6226e-06	0.0000e+00
model.layer.3	8.2850e-06	4.6253e-05	**7.2241e-05**	7.5102e-06	6.2159e-06	5.0238e-06	0.0000e+00
model.layer.4	9.2983e-06	**6.0320e-05**	7.6294e-06	2.0266e-06	8.3787e-06	3.7125e-06	0.0000e+00
model.layer.5	5.1856e-06	**1.1826e-04**	8.5831e-05	4.6730e-05	7.3910e-06	7.5442e-06	0.0000e+00
model.layer.6	2.9802e-06	**8.0824e-05**	1.7166e-05	1.8954e-05	1.8086e-05	1.3147e-05	0.0000e+00
model.layer.7	6.8545e-06	**2.6655e-04**	1.0252e-05	5.7161e-05	2.4983e-05	1.8494e-05	0.0000e+00
model.layer.8	1.5259e-05	4.0054e-05	6.2943e-05	**6.7830e-05**	2.5306e-05	3.4111e-05	8.5149e-08
model.layer.9	1.5676e-05	1.4877e-04	3.5524e-05	**1.5253e-04**	2.0521e-05	8.5524e-05	3.0654e-07
model.layer.10	2.1100e-05	**2.3603e-05**	2.0266e-05	9.1791e-06	2.0095e-05	2.3212e-05	8.8555e-07
model.layer.11	1.7047e-05	9.2745e-05	1.6212e-05	1.4150e-04	9.8177e-05	**1.9644e-04**	0.0000e+00
model.layer.12	3.4869e-05	3.0041e-05	3.9816e-05	**1.6868e-04**	4.1349e-05	3.1795e-05	1.7030e-08
model.layer.13	1.9312e-05	1.1396e-04	9.2983e-05	**1.5461e-04**	1.3726e-04	5.8413e-05	3.1505e-06
model.layer.14	1.3292e-05	**3.4046e-04**	3.6001e-05	2.2006e-04	5.5194e-05	1.8431e-04	1.5327e-07
model.layer.15	2.1994e-05	**2.5177e-04**	5.4598e-05	1.6034e-04	2.2844e-04	8.3344e-05	1.0218e-07
model.layer.16	1.5736e-05	**1.9121e-04**	1.0800e-04	4.2439e-05	1.3312e-04	1.6877e-04	1.3113e-06
model.layer.17	2.3484e-05	**4.6253e-04**	4.1723e-05	1.3864e-04	3.0960e-05	9.8041e-05	0.0000e+00
model.layer.18	2.9087e-05	**9.1815e-04**	2.7895e-05	5.1320e-05	4.6985e-05	2.6720e-05	1.7030e-08
model.layer.19	9.0241e-05	**5.8985e-04**	7.8678e-05	1.0586e-04	4.2081e-05	3.7687e-05	0.0000e+00
model.layer.20	2.8074e-05	**1.0700e-03**	5.8413e-05	8.0824e-05	2.7980e-05	2.6345e-05	2.7248e-07
model.layer.21	2.0146e-05	**3.5906e-04**	6.5327e-05	2.0623e-05	2.3076e-05	3.1727e-05	0.0000e+00
model.layer.22	3.6776e-05	**3.2091e-04**	1.9550e-05	3.7968e-05	2.2650e-05	2.7384e-05	6.8120e-08
model.layer.23	1.8179e-05	**1.3947e-04**	7.0095e-05	1.4067e-05	1.5974e-05	3.1284e-05	0.0000e+00
model.layer.24	4.0531e-05	**1.7524e-04**	8.2016e-05	5.4061e-05	5.0698e-05	3.2101e-05	5.1090e-08
model.layer.25	1.0610e-04	1.1253e-04	1.0538e-04	**1.7780e-04**	7.0282e-05	1.1855e-04	0.0000e+00
model.layer.26	3.6001e-05	**2.8610e-04**	2.5201e-04	6.8963e-05	2.0691e-05	2.0911e-04	5.1090e-08
model.layer.27	4.2439e-05	**3.9148e-04**	3.0994e-05	1.8215e-04	3.2800e-05	1.7030e-04	0.0000e+00
model.layer.28	3.0160e-05	**1.9622e-04**	1.3351e-05	1.1086e-04	1.3278e-04	2.9666e-05	0.0000e+00
model.layer.29	5.3048e-06	6.4373e-05	4.7207e-05	7.1347e-05	5.9298e-05	**1.2226e-04**	1.7030e-08
model.layer.30	3.9279e-05	**1.7309e-04**	5.6982e-05	1.0878e-04	1.6493e-04	7.4897e-05	9.0429e-06
model.layer.31	9.2447e-05	**4.7541e-04**	9.2030e-05	1.4389e-04	1.9859e-04	1.1740e-04	1.5327e-06

表 6.54: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D}'1}}$, $M_{\mathrm{SFT}_{\mathrm{D}'1}}$ 間 コンフリクト 数, コンフリクト率, Conflict Limited L2 ノルム

モデル	C	C_rate	C Limited L2	C_rate (正規化)	C Limited L2 (正規化)
model.layers.0	37	1.6964e-07	0.0025837	0.00000	0.00000
model.layers.1	82	3.7595e-07	0.0040191	0.01647	0.06931
model.layers.2	80	3.6678e-07	0.0038292	0.01573	0.06014
model.layers.3	168	7.7025e-07	0.0055081	0.04793	0.14121
model.layers.4	157	7.1981e-07	0.0053489	0.04391	0.13352
model.layers.5	298	1.3663e-06	0.0072875	0.09550	0.22713
model.layers.6	189	8.6653e-07	0.0058949	0.05562	0.15988
model.layers.7	701	3.2139e-06	0.01121	0.24296	0.41637
model.layers.8	630	2.8884e-06	0.01065	0.21698	0.38937
model.layers.9	**901**	**4.1309e-06**	**0.01272**	**0.31614**	**0.48961**
model.layers.10	576	2.6408e-06	0.01007	0.19722	0.36130
model.layers.11	**2121**	**9.7244e-06**	**0.02025**	**0.76253**	**0.85314**
model.layers.12	**1310**	**6.0061e-06**	**0.01531**	**0.46579**	**0.61465**
model.layers.13	**1273**	**5.8365e-06**	**0.01509**	**0.45225**	**0.60386**
model.layers.14	**2770**	**1.27e-05**	**0.02208**	**1.00000**	**0.94165**
model.layers.15	**1901**	**8.7157e-06**	**0.01836**	**0.68203**	**0.76167**
model.layers.16	**1518**	**6.9597e-06**	**0.01650**	**0.54190**	**0.67201**
model.layers.17	**900**	**4.1263e-06**	**0.01233**	**0.31577**	**0.47078**
model.layers.18	592	2.7142e-06	0.01033	0.20307	0.37414
model.layers.19	**1289**	**5.9098e-06**	**0.01549**	**0.45810**	**0.62315**
model.layers.20	**1271**	**5.8273e-06**	**0.01530**	**0.45152**	**0.61418**
model.layers.21	546	2.5033e-06	0.0099321	0.18624	0.35483
model.layers.22	434	1.9898e-06	0.0086936	0.14526	0.29503
model.layers.23	245	1.1233e-06	0.0067305	0.07611	0.20023
model.layers.24	501	2.297e-06	0.0094366	0.16978	0.33090
model.layers.25	**1003**	**4.5986e-06**	**0.01368**	**0.35346**	**0.53573**
model.layers.26	**862**	**3.9521e-06**	**0.01267**	**0.30187**	**0.48701**
model.layers.27	**1204**	**5.5201e-06**	**0.01482**	**0.42700**	**0.59067**
model.layers.28	834	3.8237e-06	0.01231	0.29162	0.46961
model.layers.29	641	2.9389e-06	0.01096	0.22100	0.40442
model.layers.30	**1551**	**7.111e-06**	**0.01793**	**0.55397**	**0.74105**
model.layers.31	**2612**	**1.1975e-05**	**0.02329**	**0.94219**	**1.00000**

表 6.55: 実験 3 における $M_{\mathrm{DPO}_{\mathrm{D}'_1}}$, $M_{\mathrm{SFT}_{\mathrm{D}'_1}}$ 間の各パラメータのコンフリクト率の値

layer	q_proj	v_proj	k_proj	o_proj	gate_proj	up_proj	down_proj
model.layer.0	5.9605e-07	**3.0994e-06**	2.3842e-07	1.7881e-07	8.5149e-08	8.5149e-08	0.0000e+00
model.layer.1	**2.5630e-06**	1.6689e-06	7.1526e-07	1.1921e-07	3.4060e-08	1.5327e-07	2.7248e-07
model.layer.2	1.4305e-06	**3.0994e-06**	2.6226e-06	2.9802e-07	1.7030e-07	2.8951e-07	0.0000e+00
model.layer.3	**6.2585e-06**	1.6689e-06	3.3379e-06	4.1723e-07	1.1921e-07	4.7684e-07	0.0000e+00
model.layer.4	**3.6359e-06**	4.7684e-07	1.9073e-06	1.1921e-07	6.3011e-07	8.0041e-07	0.0000e+00
model.layer.5	8.9407e-07	**1.5020e-05**	5.4836e-06	6.0201e-06	1.1240e-06	5.1090e-07	0.0000e+00
model.layer.6	8.3447e-07	**6.6757e-06**	3.0994e-06	8.9407e-07	1.6008e-06	4.2575e-07	0.0000e+00
model.layer.7	1.2517e-06	**2.8372e-05**	2.1458e-06	1.5557e-05	2.8781e-06	2.0776e-06	0.0000e+00
model.layer.8	4.1723e-07	5.2452e-06	2.6226e-06	**1.1504e-05**	2.5715e-06	4.1894e-06	0.0000e+00
model.layer.9	9.5367e-07	**1.7881e-05**	1.4305e-06	1.7345e-05	3.4060e-06	5.3304e-06	0.0000e+00
model.layer.10	2.3842e-06	**2.4080e-05**	8.5831e-06	4.8280e-06	2.2479e-06	3.1676e-06	0.0000e+00
model.layer.11	2.3246e-06	2.0742e-05	4.2915e-06	1.3173e-05	8.8385e-06	**2.1066e-05**	0.0000e+00
model.layer.12	6.6161e-06	**3.5048e-05**	8.1062e-06	2.2471e-05	7.6635e-06	3.2527e-06	0.0000e+00
model.layer.13	5.8413e-06	**3.3855e-05**	7.1526e-06	1.6630e-05	6.1478e-06	6.1478e-06	3.4060e-08
model.layer.14	1.8477e-06	**6.1035e-05**	1.0967e-05	5.2810e-05	3.8658e-06	2.2548e-05	0.0000e+00
model.layer.15	9.7752e-06	**4.6730e-05**	1.7881e-05	3.7551e-05	8.9237e-06	5.3133e-06	0.0000e+00
model.layer.16	4.7684e-06	**2.1458e-05**	7.6294e-06	5.1260e-06	9.2983e-06	1.1648e-05	0.0000e+00
model.layer.17	2.6226e-06	**5.6267e-05**	3.8147e-06	1.2279e-05	1.9755e-06	4.8024e-06	0.0000e+00
model.layer.18	2.2650e-06	**4.0054e-05**	6.6757e-06	4.4107e-06	1.4646e-06	3.3719e-06	0.0000e+00
model.layer.19	3.6061e-05	**5.6982e-05**	2.7180e-05	2.3842e-06	1.5838e-06	3.3719e-06	0.0000e+00
model.layer.20	6.0797e-06	**2.1982e-04**	1.6928e-05	8.9407e-07	8.8555e-07	1.8563e-06	0.0000e+00
model.layer.21	1.4901e-06	**7.6056e-05**	5.7220e-06	2.3842e-06	1.5497e-06	8.0041e-07	0.0000e+00
model.layer.22	4.3511e-06	**3.9577e-05**	2.6226e-06	2.9802e-06	9.0258e-07	1.3794e-06	0.0000e+00
model.layer.23	4.6492e-06	1.1206e-05	**1.1921e-05**	1.0729e-06	4.7684e-07	4.0872e-07	0.0000e+00
model.layer.24	1.2517e-06	**2.0266e-05**	1.2636e-05	4.5300e-06	9.7070e-07	3.5592e-06	0.0000e+00
model.layer.25	3.2187e-06	**5.2929e-05**	1.0014e-05	9.2387e-06	3.3038e-06	5.7220e-06	0.0000e+00
model.layer.26	3.8147e-06	**4.9353e-05**	2.4557e-05	4.2319e-06	3.6103e-06	3.4911e-06	0.0000e+00
model.layer.27	1.5259e-05	**8.2493e-05**	1.4067e-05	9.2387e-06	3.3379e-06	3.2697e-06	0.0000e+00
model.layer.28	5.3644e-06	1.6212e-05	9.2983e-06	**1.9491e-05**	2.7078e-06	2.5715e-06	0.0000e+00
model.layer.29	5.9009e-06	1.0252e-05	**1.8120e-05**	3.1590e-06	3.0654e-06	3.2357e-06	0.0000e+00
model.layer.30	7.2718e-06	**1.6284e-04**	1.9312e-05	8.3447e-06	4.9727e-06	3.9509e-06	1.7030e-08
model.layer.31	1.8597e-05	**6.8188e-05**	4.5776e-05	3.7432e-05	1.1819e-05	8.4128e-06	1.0218e-07

表 6.56: (実験 3 における $M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, M_{base}), ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D'1}}}$), ($M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D0}}}$) 間の L1 ノルム, L2 ノルム

モデル	L1 ノルム	L2 ノルム
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D'1}}}$	1.39841×10^5	0.32770
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, M_{base}	2.93887×10^5	2.41596
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, $M_{\mathrm{DPO}_{\mathrm{D0}}}$	3.10921×10^5	2.75366
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$, $M_{\mathrm{SFT}_{\mathrm{D0}}}$	3.55587×10^5	3.58559

表 6.57: 実験 3 における各モデルの生成結果とデータセット D_0 との BLEU, BERTScore

モデル	BLEU (D_0 train)	BLEU (D_0 test)	BERTScore (D_0 train)	BERTScore (D_0 test)
M_{base}	0.00551	0.00124	0.71494	0.70432
$M_{\mathrm{SFT}_{\mathrm{D0}}}$	0.05884	0.04047	0.76336	0.74730
$M_{\mathrm{SFT}_{\mathrm{D'1}}}$	0.01188	0.00794	0.72741	0.71770
$M_{\mathrm{DPO}_{\mathrm{D0}}}$	0.030986	0.03412	0.73157	0.72027
$M_{\mathrm{DPO}_{\mathrm{D'1}}}$	0.00386	0.00326	0.70879	0.69791
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$	0.02996	0.02777	0.74932	0.73694
$M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'2}}})}$	0.1424	0.00256	0.70277	0.68461
$M_{\mathrm{DPO}_{\mathrm{D0}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$	0.04344	0.03214	0.75932	0.73677
$M_{\mathrm{DPO}_{\mathrm{D'1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D'1}}})}$	0.1534	0.00209	0.70451	0.69735

んまやな！」といった男子大学生の関西弁のスタイルを反映してロールプレイしていることがわかる. また $M_{\mathrm{DPO}_{\mathrm{D1}}(\mathrm{M}_{\mathrm{SFT}_{\mathrm{D1}}})}$ の生成例では, 一人称が「俺」となっている. これは D_0, D'_1 の一人称の違いが学習で反映された結果だと考えられる.

表 6.58: 実験 3 における各モデルの生成結果とデータセット D'_1 との BLEU, BERTScore

モデル	BLEU (D'_1 train)	BLEU (D'_1 test)	BERTScore (D'_1 train)	BERTScore (D'_1 test)
M_{base}	0.01013	0.00153	0.71704	0.71367
$M_{\text{SFT}_{\text{D0}}}$	0.02000	0.01555	0.72668	0.72545
$M_{\text{SFT}_{\text{D}'1}}$	0.06373	0.03176	0.76366	0.75154
$M_{\text{DPO}_{\text{D0}}}$	0.00707	0.01175	0.70401	0.70297
$M_{\text{DPO}_{\text{D}'1}}$	0.02481	0.01892	0.74651	0.73052
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D}'2}})}$	0.01994	0.01201	0.72920	0.72781
$M_{\text{DPO}_{\text{D}'1}(\text{M}_{\text{SFT}_{\text{D}'2}})}$	0.04932	0.02237	0.74694	0.73018
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D}'1}})}$	0.01734	0.01559	0.72810	0.72585
$M_{\text{DPO}_{\text{D}'1}(\text{M}_{\text{SFT}_{\text{D}'1}})}$	0.04252	0.01809	0.74170	0.73809

表 6.59: 実験 3 における評価者 LLM による各モデルの生成結果の評価

モデル	train データのクエリ (162 件)	test データのクエリ (40 件)
dataset	6.56	6.85
M_{base}	5.40	5.35
$M_{\text{SFT}_{\text{D0}}}$	2.35	2.38
$M_{\text{SFT}_{\text{D}'1}}$	7.6	7.8
$M_{\text{DPO}_{\text{D0}}}$	7	3.77
$M_{\text{DPO}_{\text{D}'1}}$	**7.70**	**7.93**
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D}'2}})}$	2.50	1.85
$M_{\text{DPO}_{\text{D}'1}(\text{M}_{\text{SFT}_{\text{D}'2}})}$	7.5	7.48
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D}'1}})}$	1.92	1.78
$M_{\text{DPO}_{\text{D}'1}(\text{M}_{\text{SFT}_{\text{D}'1}})}$	7.56	7.65

表 6.60: 「このカフェ素敵ですね」に対する LLM の応答

モデル	「このカフェ素敵ですね」に対する LLM の応答
D_0	私もこのお店が気に入っています。静かな雰囲気がいいですね
D_1	せやな、俺もここよう来るで。落ち着いた雰囲気がええ感じや
M_{base}	まあ、普通のカフェだな。
$M_{\text{SFT}_{\text{D0}}}$	落ち着いた雰囲気が気に入っています。勉強にも集中できそうです。
$M_{\text{SFT}_{\text{D}'1}}$	おお、ほんまやな！落ち着いた雰囲気でええわ
$M_{\text{DPO}_{\text{D0}}}$	店内が清潔で雰囲気も良いです。
$M_{\text{DPO}_{\text{D}'1}}$	ほんまやな！めっちゃお洒落やし、雰囲気ええやん！
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D}'2}})}$	落ち着いた雰囲気で気に入っています。
$M_{\text{DPO}_{\text{D}'1}(\text{M}_{\text{SFT}_{\text{D}'2}})}$	ほんまやな！なんか落ち着くわ。
$M_{\text{DPO}_{\text{D0}}(\text{M}_{\text{SFT}_{\text{D}'1}})}$	ありがとうございます。落ち着いた雰囲気で気に入っています。
$M_{\text{DPO}_{\text{D}'1}(\text{M}_{\text{SFT}_{\text{D}'1}})}$	おお、わかるわ！俺も好きやで！

7　まとめと今後の課題

本研究では Conflict Limited ノルムを用いてユーザー嗜好学習について重みに基づく解析をした. 具体的には事前学習済みモデルを 2 種類のデータセットを用いて様々な条件下でファインチューニングし, データセット間の違いおよび SFT と DPO の違いを分析した. 3 種類の実験の結果以下の知見が得られた.

- DPO と SFT の学習時の重み変化量の違いを定量化して示した.
- Conflict Limited L2 ノルムにより, LLM の線形層のパラメータの中で v_proj がロールプレイのタスクに置いて重要な役割を果たしている可能性を示唆した.
- LLM の冗長性により, 特定タスクを実現した後のファインチューニングにおける重み分布が一意に定まらないことを示した.
- 日本語向けに大規模な事前学習を施したモデルと, 多言語対応の汎用性の高いデータセットを用いて事前学習したモデルとの間で, 日本語ロールプレイタスクにおける重み変化の局所的な違いを明らかにした.
- 多言語に対応する汎用性の高いデータセットで事前学習したモデルにおいて, 浅い層で言語情報やタスクの識別に役立つ情報を保持している可能性を示唆した.
- 日本語向けに大規模な事前学習を施したモデルについて, 異なるデータセット間で層ごとの重み変化を確認した結果, 中間層がデータセット内の固有名詞や語尾など特徴的な言語表現を保持している可能性, また最終層が出力のスタイルを決定する上で重要な役割を担っている可能性を示唆した.
- 小規模なデータセットでは DPO と SFT を組み合わせることによる性能の強化は見られなかった.

本研究における今後の課題としては以下が挙げられる.

- より大規模で異なるドメインのデータセット

  本研究は訓練データが 162 件という非常に少数のデータでファインチューニングをした. 定性的な評価ではロールプレイは実現できたが,ChatHaruhi で公開されている 52 万件の大規模データセットを用いてロールプレイを行う LLM を作成することで, ユーザー嗜好学習における重み変動がより顕著に現れる可能性がある.

- 特定の層のパラメータを固定した学習

  今回の実験では,elyza/Llama-3-ELYZA-JP-8B において, 日本語ロールプレイタスクで重要な役割を果たす層が最終層付近および中間層であることが明らかになったが, 各層がどのレベルの言語情報を保持しているかの詳細な解析は行えていない. 今後は, 中間層以降のパラメータを固定してファインチューニングすることで, 浅い層の役割や, 最もベースモデルからの変動が大きい層の影響を段階的に評価し, 層ごとの役割を定量的に明らかにする可能性がある.

- モデルマージへの応用

  Conflict Limited L2 ノルムは, タスクベクトル間のコンフリクトの大きさを定量的に示す指標である. モデルマージにおいて, マージに使用するタスクベクトルの構築時に, コンフリクトが激しい層についてはベースモデル側またはどちらか一方の重みを用いるといった新たなマージ手法を,Conflict Limited L2 ノルムを活用して実現できる可能性がある.

- 様々なベースモデルの検討実験 1,2 の比較から,SFT や DPO による重み変動はベースモデルの事前学習内容に大きく依存することが示された. 本研究では 2 種類のモデルしか試していないが, 今後は他の事前学習済み日本語特化モデルや, 日本語以外の言語に特化した LLM に日本語タスクを学習させることで, ロールプレイタスクにおいてどの層が特に寄与しているかを明らかにする可能性がある.

謝辞

あとで先輩の参考に書く.

2025 年 2 月 21 日

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, Vol. abs/1706.03762, , 2017.

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[3] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond, 2023.

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[5] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models, 2023.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, Vol. 9, No. 8, p. 1735–1780, nov 1997.

[7] F.A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: continual prediction with lstm. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, Vol. 2, pp. 850–855 vol.2, 1999.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[10] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

[11] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, Vol. abs/1810.04805, , 2018.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, Vol. abs/2010.11929, , 2020.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, Vol. abs/1910.10683, , 2019.

[16] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, Vol. abs/2005.14165, , 2020.

[17] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho,

Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John

Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[20] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira

Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco

Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe,

Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *CoRR*, Vol. abs/2009.03300, , 2020.

[22] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

[23] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano,

Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, Vol. abs/2110.14168, , 2021.

[24] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023.

[25] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[26] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report, 2024.

[27] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[29] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2024.

[30] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2023.

[31] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, Vol. abs/1707.06347, , 2017.

[33] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86, 1951.

[34] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, Vol. abs/2106.09685, , 2021.

[35] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[37] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, Vol. abs/1904.09675, , 2019.

[38] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 150–157, 2003.

[39] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. *CoRR*, Vol. abs/2005.00661, , 2020.

[40] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025.

[41] LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

[42] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing, 2023.

[43] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. Chatharuhi: Reviving anime character in reality via large language model, 2023.

[44] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models, 2024.

[45] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations, 2019.

[46] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors, 2018.

[47] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[48] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019.

[49] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.