多目的直接選好最適化における次元衝突のカリキュラム学習による抑制

西田光甫　西田京介
NTT 株式会社　人間情報研究所
{kosuke.nishida, kyosuke.nishida}@ntt.com

概要

大規模言語モデルを多様な人間にアラインメントさせるためには，選好次元を複数用意し，推論時に選好重みを用いて応答を制御することが重要である．その手段に制御可能多目的選好最適化（MODPO）があるが，本研究は MODPO モデルが仮定する暗黙的報酬が，次元ごとに好ましい応答が異なるデータに対して目的次元に即した応答を選択できない，次元衝突の問題を指摘する．本研究は次元衝突の問題を解決するため，カリキュラム学習と新たな損失関数を提案する．2 次元・5 次元の多目的選好データセットを用いた評価において，提案手法がベースラインに比べて選好重みに条件付属けられた暗黙的報酬を付与することを確認した．

1　はじめに

大規模言語モデル (LLM) のアラインメントは LLM の応答を人間の好みと一致させるための重要な手段であるが，通常単独の目的関数を用いるため，平均的な人間の好みを学習してしまう．制御可能な多目的アラインメントは “有用性” や “無害性” といった複数の次元を用意してアラインメントを行い，かつ推論時に選好重みを用いて次元を跨いだ制御を可能とすることで，LLM が多様な人間の好みを学習し，反映した応答を可能にする技術である．中でも多目的直接選好最適化 (MODPO) [1] は，人間がアノテーションした公開多目的選好データ [2, 3] を用いて実行可能なため，実用性に優れる．

本研究は，既存の MODPO 手法が次元衝突の課題を抱えていることを実験的に示した．ここで，1 つのクエリに 2 つの応答が紐づいたデータセットにおいて，ある次元で好まれる応答と別の次元で好まれる応答が異なる事例である衝突事例に注目した．DPO モデルは各応答に対して暗黙的報酬を与えるが，衝突事例に対して注目したい次元に応じた暗黙的報酬を与えることが MODPO の本質的な目標であると言える．しかし，既存の MODPO 手法では衝突事例に対して選好重みを用いて条件付けた暗黙的報酬を与えることが難しいことを確認した．

我々はこの次元衝突の原因を，MODPO 手法が学習の中で次元非依存の一般的な選好と次元固有の選好を切り分けて学習できていないことにあると考えた．そこで MODPO にカリキュラム学習を導入し，(i) まず次元共有の選好を学習し，(ii) 次に次元固有の選好を学習し，(iii) 最後に選好次元の制御を学習することを提案した．また，カリキュラム学習の特徴を活用した KL ダイバージェンスに基づく損失関数を提案した．本損失関数は，提案手法の特徴である第二段階までに次元固有の暗黙的報酬を学習している点を利用して設計されている．また，本損失関数は DPO の理論と多目的強化学習で一般的な線形スカラー化を組み合わせる形で自然に導出できる．

2 次元・5 次元の多目的選好データセットを用いた実験において，DPO で学習した暗黙的報酬モデルが選好重みで条件付けた報酬を応答に付与しているかを検証した．結果，提案手法は既存手法に比べて，衝突事例に対して選好重みで条件付けられた報酬を付与していることを確認した．

2　準備

2.1　問題設定

データセット要件　サンプル $(x, y_a, y_b, \boldsymbol{s}_a, \boldsymbol{s}_b)$ から構成される k 次元選好データセット $\mathcal{D}$ を利用可能とする．ここで，x はクエリ，y_a, y_b は応答，$\boldsymbol{s}_a, \boldsymbol{s}_b \in \mathbb{R}^k$ は第 i 成分が選好次元 i におけるスコアを示すベクトルである．

目標　本研究の目標は，選好重み $\boldsymbol{w} \in \Delta^k$ で条件付けられた応答を生成する方策 $\pi(y; \boldsymbol{w})$ を学習する

ことである．なお，Δ^k は k 次元単体である．

2.2 直接選好最適化（DPO）

以下に DPO の理論背景を概説する．DPO は人間の選好がオラクル報酬モデル $r^*(y)$ を用いた Bradley-Terry (BT) モデル [4] に従うと仮定した．つまり，応答 y_a が y_b より好まれる確率を

$$\mathcal{P}(y_a \succ y_b) = \sigma\left(r^*(y_a) - r^*(y_b)\right)$$

とした．ここで，$\sigma(\cdot)$ はシグモイド関数である．報酬モデル r と参照方策 π_{ref} に対して，KL 制約強化学習は β をハイパーパラメータとして次の問題を最適化する

$$\max_{\pi} \quad \mathbb{E}_{y\sim\pi}\left[r(y) - \beta \log \frac{\pi(y)}{\pi_{\text{ref}}(y)}\right]. \tag{1}$$

$\hat{\pi}$ を最適化問題 1 の最適解，Z を分配関数とする．最適解 $\hat{\pi}$ を用いて，対応する報酬モデル r は

$$r(y) = \beta \log \frac{\hat{\pi}(y)}{\pi_{\text{ref}}(y)} + \beta Z$$

と書ける．オラクル報酬モデル r^* における最適化問題 1 の解 π^* を BT モデルに代入し，

$$\mathcal{P}(y_a \succ y_b) = \quad \sigma\left(\beta \log \frac{\pi^*(y_a)}{\pi_{\text{ref}}(y_a)} - \beta \log \frac{\pi^*(y_b)}{\pi_{\text{ref}}(y_b)}\right)$$

を得る．以上を背景に，正例と負例のペア (y_c, y_r) を用いて，DPO は方策 π_θ を $\mathcal{P}(y_c \succ y_r)$ の負の対数を損失関数とした最適化によって学習する

$$L_{\text{DPO}}(\theta) = \quad -\log \sigma\left(\beta \log \frac{\pi_\theta(y_c)}{\pi_{\text{ref}}(y_c)} - \beta \log \frac{\pi_\theta(y_r)}{\pi_{\text{ref}}(y_r)}\right).$$

この定式化によって，DPO は選好データセットのみに依存して方策を最適化でき，報酬モデルや強化学習を不要とする．ここで，DPO によって学習した方策の $\beta \log \frac{\pi_\theta(y)}{\pi_{\text{ref}}(y)}$ は暗黙的報酬と呼ばれ，DPO モデルが応答 y に与える報酬と解釈できる．

3 提案手法

提案手法の概要を図 1 に示す．提案手法はモデル構造として Transformer [5] と Panacea [6] に基づく．提案手法の新規性は次元衝突を抑制するためにカリキュラム学習を MODPO へ導入したこと，及びそのために設計したデータ分割と損失関数にある．

Panacea は $k+1$ 個の AdaLoRA [7] モジュールを LLM の各層に導入する．第一のモジュールは次元共有の選好，以降の k モジュールはそれぞれ k 個の選好次元に対応する能力を学習する．h_i を第 i モジュールの出力とし，Panacea は $h_0 + \sum_{0\le i<k} w_i h_{i+1}$

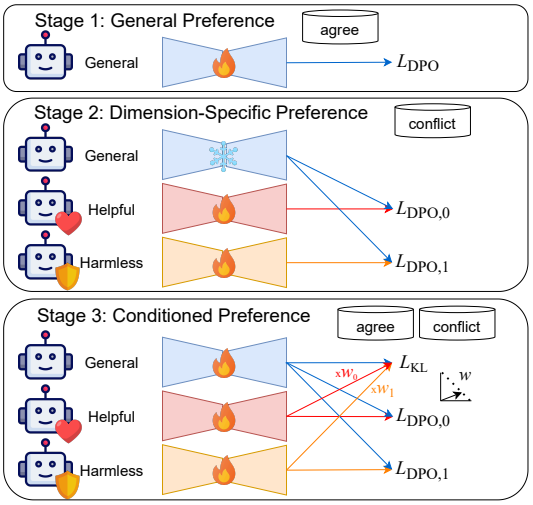

図 1 提案手法の概要．

を上層に伝播することで選好重み $\boldsymbol{w}$ の情報を隠れ状態に埋め込む．ここで，ψ を第 1 モジュールのパラメータ，ϕ_i を第 $i+1$ モジュールのパラメータとする．なお，提案手法は AdaLoRA ではなく LoRA [8] を採用した．

3.1 データ分割

次元衝突の抑制のため，データを 2 つに分割した．$\mathcal{D}_{\text{agree}}$ は次元衝突がないサンプルの集合であり，$\boldsymbol{s}_a - \boldsymbol{s}_b$ の非零要素の正負が一致するサンプルである．$\mathcal{D}_{\text{conflict}}$ は次元衝突があるサンプルの集合であり，$\mathcal{D} \setminus \mathcal{D}_{\text{agree}}$ である．つまり，$\mathcal{D}_{\text{agree}}$ は次元共有の選好を学習するためのサブセット，$\mathcal{D}_{\text{conflict}}$ は次元固有の選好を学習するためのサブセットである．

3.2 カリキュラム学習

一般選好最適化 第一段階では $\mathcal{D}_{\text{agree}}$ を用いて ψ を訓練する．$\boldsymbol{w} = \boldsymbol{0}$ を代入し，次元固有モジュールを無効化する．$\mathcal{D}_{\text{agree}}$ では全ての次元で正例・負例が共通であるため，通常の DPO 損失 $L_{\text{DPO}}(\psi)$ を用いて学習する．全ての段階で，参照方策は第一段階における初期モデルとする．

次元固有選好最適化 第二段階では $\mathcal{D}_{\text{conflict}}$ を用いて $\{\phi_i\}_i$ を訓練し，ψ を固定する．最適化の各ステップにおいて，順伝播・逆伝播計算を k 回行う．i 回目の計算で $\boldsymbol{w} = \mathbb{1}_i$ を代入し，損失関数には第 i 次元のスコアに基づいて正例・負例を定義した DPO 損失 $L_{\text{DPO},i}(\phi_i)$ を用いる．ここで，$\mathbb{1}_i$ は第 i 成分が 1 で他が 0 の指示ベクトルである．

条件つき選好最適化　第三段階では $\psi, \{\phi_i\}_i$ 全てを更新し，方策及び暗黙的報酬の選好重み $\boldsymbol{w}$ による条件付けを学習する．学習データは $\mathcal{D}_{\text{agree}}$ から $|\mathcal{D}_{\text{agree,sub}}| = |\mathcal{D}_{\text{conflict}}|$ を満たすようにダウンサンプリングした $\mathcal{D}_{\text{agree,sub}}$ と $\mathcal{D}_{\text{conflict}}$ とする．最適化の各ステップにおいて，順伝播・逆伝播計算を $k+1$ 回行う．k 回の計算は次元固有の計算であり，第二段階の操作と同様である．このとき，$k+1$ 回目の計算のために勾配情報を捨てた $\{\pi(y_a; \mathbb{1}_i), \pi(y_b; \mathbb{1}_i), \pi_{\text{ref}}(y_a), \pi_{\text{ref}}(y_b)\}_i$ を保存する．$k+1$ 回目の計算では，選好重み $\boldsymbol{w}$ を Δ^k からランダムサンプリングし，$\pi(y_a; \boldsymbol{w}), \pi(y_b; \boldsymbol{w}), \pi_{\text{ref}}(y_a), \pi_{\text{ref}}(y_b)$ を計算する．これらの方策を次小節で説明する損失関数に代入して最適化を行う．

3.3　損失関数

2.2 節で議論したように，DPO の損失は報酬モデル r から導出される．カリキュラム学習を導入した提案手法では，第三段階において，過去の段階で学習した第 i 次元固有の暗黙的報酬モデル r_i を利用できる．提案手法は，r_i を教師モデルとすることで $\boldsymbol{w}$ で条件付けた方策 $\pi_\theta(y; \boldsymbol{w})$ を学習する．

多目的強化学習では，目的次元ごとの報酬 $r_i(y)$ を集計関数 f によって $r_w(y) = f(\{r_i(y)\}_i; \boldsymbol{w}) \in \mathbb{R}$ に変換し，報酬として利用する．そこで提案手法は，方策 $\pi_\theta(y; \boldsymbol{w})$ に相当する暗黙的報酬が $r_w(y)$ と一致するように学習を行う．以下では，集計関数として標準的な線形スカラー化 (LS) $\sum_i w_i r_i$ [9, 10, 11] に即して提案手法を説明する．しかし議論は LS に閉じたものではなく，BT モデルに対応する任意の集計関数に提案手法を適用できる．

導出　選好重み $\boldsymbol{w} \in \{\mathbb{1}_i\}_i$ に対して方策 $\pi_\theta(y; \boldsymbol{w})$ が得られているため，第 i 次元の報酬は

$$r_i(y) = \beta \log \frac{\pi_\theta(y; \mathbb{1}_i)}{\pi_{\text{ref}}(y)} + \beta Z$$

と書ける．今，任意の $\boldsymbol{w} \in \Delta^k$ に対する報酬と対応する BT モデルは集計関数に従って以下で得られる

$$r_w(y) = \sum_i w_i \left(\beta \log \frac{\pi_\theta(y; \mathbb{1}_i)}{\pi_{\text{ref}}(y)} + \beta Z \right),$$

$$\mathcal{P}(y_a \succ y_b; \boldsymbol{w}) = \sigma\left(\beta \sum_i w_i \left(\log \frac{\pi_\theta(y_a; \mathbb{1}_i)}{\pi_{\text{ref}}(y_a)} - \log \frac{\pi_\theta(y_b; \mathbb{1}_i)}{\pi_{\text{ref}}(y_b)} \right) \right).$$

第三段階の目標は $\boldsymbol{w} \in \Delta^k$ で条件付けた方策 $\pi_\theta(y; \boldsymbol{w})$ を得ることである．そこで，方策 $\pi_\theta(y; \boldsymbol{w})$ に基づいて暗黙的報酬を計算する BT モデルと集計関数に従って得られた BT モデルを一致させる問題として損失関数を定式化する

$$L_{\text{KL}}(\theta; \boldsymbol{w}) = \text{KL}\left[\mathcal{P}(y_a \succ y_b; \boldsymbol{w}) \middle\| \sigma\left(\beta \log \frac{\pi_\theta(y_a; \boldsymbol{w})}{\pi_{\text{ref}}(y_a)} - \beta \log \frac{\pi_\theta(y_b; \boldsymbol{w})}{\pi_{\text{ref}}(y_b)} \right) \right].$$

ここで，KL はベルヌーイ分布間の KL ダイバージェンス $\text{KL}[p||q] = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ である．

目的関数の勾配に基づく議論　z を暗黙的報酬の差 $\beta \log \frac{\pi_\theta(y_a; \boldsymbol{w})}{\pi_{\text{ref}}(y_a)} - \beta \log \frac{\pi_\theta(y_b; \boldsymbol{w})}{\pi_{\text{ref}}(y_b)}$ とする．ここで，損失関数の勾配は以下の式で計算できる

$$\frac{\partial L_{\text{KL}}(\theta; \boldsymbol{w})}{\partial z} = \sigma(z) - \mathcal{P}(y_a \succ y_b; \boldsymbol{w}).$$

詳細な導出は付録 A を参照されたい．よって，提案手法では訓練方策に対応する選好確率 $\sigma\left(\beta \log \frac{\pi_\theta(y_a; \boldsymbol{w})}{\pi_{\text{ref}}(y_a)} - \beta \log \frac{\pi_\theta(y_b; \boldsymbol{w})}{\pi_{\text{ref}}(y_b)} \right)$ が次元固有の暗黙的報酬と集計関数から計算された選好確率 $\mathcal{P}(y_a \succ y_b; \boldsymbol{w})$ と一致するように学習が進む．提案手法は，次元ごとの選好性（どちらが正例・負例かのバイナリ）から方策を学習する既存手法に比べ，実数値からなる暗黙的報酬を活用できる点，DPO の理論と整合する点，に関して優れると考える．

4　評価実験

4.1　実験設定

モデル　DPO 研究の慣習に従い，事前学習済みモデルに supervised fine-tuning のみを行ったモデルを初期値として学習した [12]．モデルは Llama3-8B [13] と Mistral-7B-v0.1 [14] から学習した 2 種である．

ベースライン　DPO-Soup は k 個のモデルを独立に訓練する手法である [9]．i 番目のモデルは $L_{\text{DPO},i}(\theta)$ を使って学習する．推論時はそれらのパラメータの選好重み $\boldsymbol{w}$ による重み付き和を用いる．

Panacea もベースライン手法として用いる．Panacea の損失関数として MODPO 損失で最もよく使われる線形スカラー化を採用し [1, 6]，$L_{\text{LS}}(\theta) = \sum_i w_i L_{\text{DPO},i}(\theta)$ とする．

データセット　有用性と無害性の 2 次元を持つ PKU-SafeRLHF [15, 16] と，有用性・正確性・一貫性・複雑性・冗長性の 5 次元を持つ HelpSteer2 [2, 3] の 2 つを用いた．また，次元衝突の問題に焦点を当てるため，評価データは衝突事例のみを用いた．

評価指標　文献 [6] 同様に，DPO によって訓練したモデルの暗黙的報酬が正例を選択できるかを 2 値

表 1 PKU-SafeRLHF における 2 値分類精度．

	Llama3-8B		Mistral-7B	
	有用性	無害性	有用性	無害性
DPO-Soup	**60.6**	61.2	54.7	66.0
Panacea	53.8	66.7	52.9	66.1
Proposed	58.9	**69.3**	**60.7**	**72.5**

表 2 HelpSteer2 における 2 値分類結果．50 個の実験設定における win rate と平均正解確率を示す．win rate は各設定内で 3 手法を比べて最も性能が高かった割合を示す．

	Llama3-8B		Mistral-7B	
	Win	Acc.	Win	Acc.
DPO-Soup	0.02	50.4	0.04	59.6
Panacea	0.46	63.5	0.34	65.3
Proposed	**0.52**	**65.7**	**0.62**	**67.6**

分類問題として評価した．

PKU-SafeRLHF のアノテーションは各次元でどちらの応答が好ましいかを示す．そこで，推論時は選好重み $\boldsymbol{w}$ を指示ベクトル $\mathbb{1}_i$ とする．例えば $\boldsymbol{w} = (1, 0)$ では有用性の次元で好ましい応答に高い報酬を与えた時に正解とする．本設定は，DPO-Soup にとっては目標次元に沿って学習したモデルで目標次元における評価を行う理想的な設定と言える．

HelpSteer2 でも簡潔さのため 2 次元に制限した評価を行った．つまり，5 次元単体上で訓練した後，推論は ${}_5C_2 = 10$ の次元ペアそれぞれで評価した．アノテーションは次元ごとに 0 から 4 のスコアが付いている．そのため $\boldsymbol{w} \in \{(1, 0), (0, 1)\}$ の端点以外での評価ができ，$\{(0.25, 0.75), (0.5, 0.5), (0.75, 0.25)\}$ も用いる．正例・負例は選好重みとスコアの内積 $\boldsymbol{w}^\top \boldsymbol{s}_a$ と $\boldsymbol{w}^\top \boldsymbol{s}_b$ を比較して決定した．

4.2 結果

既存手法は衝突事例に目的次元に即した報酬を与えるか？ 表 1 に PKU-SafeRLHF の結果を示す．MODPO 手法は衝突事例に対する目的次元に即した報酬の付与に課題があると言える．特に無害性が有用性に比べて優先される傾向にあり，Llama3-8B の DPO-Soup 以外で有用性はチャンスレートに近い割合で分類されている．衝突事例における性能の低さは，データの多数を非衝突事例が占めるために次元固有の選好を分離して学習できないことが原因と考えられる．今回は無害性が有用性に比べて指標としてわかりやすく，かつ非衝突事例が無害性だけから分類可能であるため，有用性の学習が進まなかったと考えられる．本観察が提案手法の動機である．

提案手法は衝突事例に目的次元に即した報酬を与えるか？ 提案手法は Llama3-8B の DPO-Soup における有用性以外の設定でベースラインを上回ることを確認した．モデル構造として採用した Panacea と比べると有用性での性能向上が顕著であり，次元共有の選好と有用性・無害性に関する選好を分離して学習したことが示唆される．

5 次元データセットで提案手法は選好重みで条件付けられた報酬を与えるか？ 表 2 に次元ペア 10 通り，選好重み 5 通りの計 50 通りの実験設定における win rate と平均正解率を示す．5 次元のデータセット・端点以外の選好重みにおいても，提案手法が既存手法に比べて選好重みで条件付けられた暗黙的報酬を応答に付与していることが確認できる．

5 関連研究

制御可能多目的直接選好最適化には，プロンプトによる制御 [17, 10] とパラメータによる制御 [9, 6] がある．提案手法はパラメータによる制御に基づいて実装したが，プロンプトによる制御と統合可能である．デコード時の制御の研究も盛んであるが，推論時に次元ごとの報酬モデルを要する [11, 18, 19]．

次元衝突に関しては文献 [20, 21] が訓練データから衝突事例を削除することで LLM の性能が向上することを報告している．また，文献 [22, 23] はパラメータや勾配の部分空間を選好次元に関して直交させることで LLM の性能を向上した．本研究は制御可能多目的直接選好最適化における次元衝突の問題に取り組んだ初めての研究である．

6 おわりに

本研究の独自性 本研究は MODPO において衝突事例に対する暗黙的報酬に注目し，次元衝突の課題が存在することを明らかにした．次元衝突の問題を解決するため，(i) 次元共有の選好，(ii) 次元固有の選好，(iii) 選好重みによる条件付け，を順に学習するカリキュラム学習を提案した．さらに，カリキュラム学習の第二段階で得られた次元固有の暗黙的報酬を線形スカラー化と組み合わせて教師信号とする新たな損失関数を提案した．

本研究の重要性 LLM の社会実装は広く進んでいるが，それらの LLM は平均的な人間の好みにアラインメントされている．LLM が人間の多様性を尊重し，ユーザ・ユースケースに即した応答を行う意義は大きく，本研究はその一助となると考える．

参考文献

[1] Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In **ACL Findings**, pp. 10586–10613, 2024.

[2] Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences, 2024.

[3] Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024.

[4] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. **Biometrika**, Vol. 39, No. 3/4, pp. 324–345, 1952.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS**, pp. 5998–6008, 2017.

[6] Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for llms. In **NeurIPS**, Vol. 37, pp. 75522–75558, 2024.

[7] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In **ICLR**, 2023.

[8] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **ICLR**, 2022.

[9] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In **NeurIPS**, Vol. 36, pp. 71095–71134, 2023.

[10] Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In **ACL**, pp. 8642–8655, 2024.

[11] Xinran Wang, Qi Le, Ammar Ahmed, Enmao Diao, Yi Zhou, Nathalie Baracaldo, Jie Ding, and Ali Anwar. Map: Multi-human-value alignment palette. **CoRR**, Vol. abs/2410.19198, , 2024.

[12] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In **ICLR**, 2024.

[13] AI@Meta. Llama 3 model card. 2024.

[14] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de Las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. corr, abs/2310.06825, 2023. doi: 10.48550. **arXiv preprint ARXIV.2310.06825**, Vol. 10, , 2023.

[15] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. **arXiv preprint arXiv:2307.04657**, 2023.

[16] Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. **arXiv preprint arXiv:2406.15513**, 2024.

[17] Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable preference optimization: Toward controllable multi-objective alignment. In **EMNLP**, pp. 1437–1454, 2024.

[18] Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hannaneh Hajishirzi, Noah A. Smith, and Simon S. Du. Decoding-time language model alignment with multiple objectives. In **NeurIPS**, Vol. 37, pp. 48875–48920, 2024.

[19] Seongho Son, William Bankes, Sangwoong Yoon, Shyam Sundhar Ramesh, Xiaohang Tang, and Ilija Bogunovic. Robust multi-objective controlled decoding of large language models. In **2nd Workshop on Models of Human Feedback for AI Alignment**, 2025.

[20] Yusen Wu, Li Jiang, Junwu Xiong, Jingqing Ruan, Yichuan Ding, Qingpei Guo, zujie wen, JUN ZHOU, and Xiaotie Deng. Hummer: Towards limited competitive preference dataset. In **COLM**, 2024.

[21] Zhihao Xu, Yongqi Tong, Xin Zhang, Jun Zhou, and Xiting Wang. Reward consistency: Improving multi-objective alignment from a data-centric perspective. **arXiv preprint arXiv:2504.11337**, 2025.

[22] Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In **ACL**, pp. 11214–11232, 2025.

[23] Liang Lin, Zhihao Xu, Junhao Dong, Jian Zhao, Yuchen Yuan, Guibin Zhang, Miao Yu, Yiming Zhang, Zhengtao Yao, Huahui Yi, et al. Orthalign: Orthogonal subspace decomposition for non-interfering multi-objective alignment. **arXiv preprint arXiv:2509.24610**, 2025.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **ICLR (Poster)**, 2015.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. **arXiv preprint arXiv:1910.03771**, 2019.

[27] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[28] 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.

[29] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[30] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In **EMNLP**, pp. 3029–3051, 2023.

[31] Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In **NeurIPS**, 2024.

[32] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **ACL Findings**, pp. 4998–5017, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

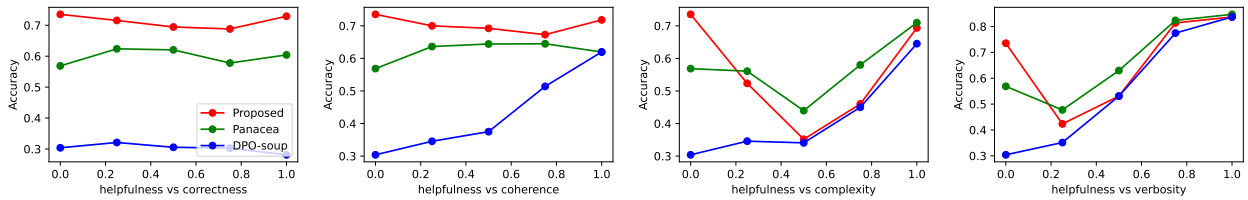

図 2　HelpSteer2 における 2 値分類結果．背後の LLM は Llama3-8B である．横軸の左側ほど有用性の重みが大きい．

A　勾配導出の詳細

提案手法の勾配の導出の詳細を以下に示す．定義 $\mathrm{KL}[p||q] = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$，$q = \sigma(z) = \frac{1}{1+e^{-z}}$ により，

$$\begin{aligned}
\frac{\partial \mathrm{KL}[p||q]}{\partial q} &= -\frac{p}{q} + \frac{1-p}{1-q} \\
&= \frac{-p(1-q)+(1-p)q}{q(1-q)} \\
&= \frac{q-p}{q(1-q)} \\
\frac{\partial q}{\partial z} &= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \sigma(z)(1-\sigma(z)) \\
&= q(1-q), \\
\frac{\partial \mathrm{KL}[p||q]}{\partial z} &= q - p
\end{aligned}$$

である．したがって，$q = \sigma(z)$ と $p = \mathcal{P}(y_a \succ y_b; \boldsymbol{w})$ を代入することで勾配が得られる

$$\frac{\partial L_{\mathrm{KL}}(\theta; \boldsymbol{w})}{\partial z} = \sigma(z) - \mathcal{P}(y_a \succ y_b; \boldsymbol{w}).$$

B　実験設定

表 3 にハイパーパラメータを示す．NVIDIA A100 (80GB) GPU 8 枚上で実験した．最適化手法には Adam [24] を用いた．利用したライブラリは PyTorch (ver. 2.6.0)[1] [25], transformers (ver. 4.51.3)[2] [26], peft (ver. 0.7.1)[3] [27], trl (ver. 0.18.0)[4] [28] である．

モデルの初期値は Llama3-base 8B [13] に Alpaca 手法 [29] を使って学習したモデル[5]と， Mistral-7B-v0.1 [14] を UltraChat 200k [30] を使って学習したモデル[6]の 2 種類とした．

LoRA モジュールは全ての self-attention 層，MLP 層に追加した．提案手法では $r = 8, \alpha = 16$ と設定した．DPO-Soup では，$r = 16, \alpha = 32$ とした．これは，提案手法の次元共有の選好を学習するモジュールと次元固有の選好を学習するモジュールのランクの和に揃えた設定である．AdaLoRA を用いた Panacea でも，$r = 8$ とした．その他の設定は原著論文及びライブラリのデフォルト値を踏襲し，α を 512，学習率を 2e-4，直交ペナルティを 0.5 とした．計算量に関する公平な比較のため，Panacea でも毎ステップ $k+1$ 通りの $\boldsymbol{w}$ を利用した．$\boldsymbol{w}$ は各 $\mathbb{1}_i$ と Δ^k からのサンプリング 1 回で計 $k+1$ 通りとした．

表 3　ハイパーパラメータ．

	PKU-SafeRLHF	HelpSteer2
Batch Size	128	32
Max Token Length	1024	
Learning rate	5e-6	
Warmup Ratio	0.03	
Weight decay	0.	
DPO β	1e-2	

C　実験結果の追加

HelpSteer2 における 5 次元の選好評価において，個々の次元ペアに注目して観察する．有用性・正確性・一貫性は相補的な関係にあるのに対し，それらと複雑性・冗長性は相反する関係にあることが確認できたため，片方のペアを有用性に固定した分類性能を図 3 に示す．win rate に示す通り提案手法は多くの設定で既存手法より高い分類性能を示すが，分類性能が低いケースは，相反する次元のペアであり，かつ $\boldsymbol{w}$ が端点以外である設定であった．この原因は，DPO の暗黙的報酬モデルが学習する長さバイアスにあると考えられる．複雑性・冗長性は長さと強く相関し，これらの次元を学習したモデルは長さバイアスの影響を受ける．提案手法は暗黙的報酬を教師信号として利用するため，長さバイアスが強い次元と弱い次元のペアで評価したとき，長さバイアスの影響について齟齬が生じ性能が低下したと考えられる．長さバイアスの軽減手法 [31, 32] と組み合わせることでこの問題は解決すると考える．

1) https://pytorch.org/
2) https://github.com/huggingface/transformers
3) https://github.com/huggingface/peft
4) https://github.com/huggingface/trl
5) https://huggingface.co/PKU-Alignment/alpaca-8b-reproduced-llama-3
6) https://huggingface.co/alignment-handbook/zephyr-7b-sft-full