BAPO: Base-Anchored Preference Optimization for Personalized Alignment in Large Language Models

Anonymous ACL submission

Abstract

While learning to align Large Language Models (LLMs) with human preferences has shown remarkable success, aligning these models to meet the diverse user preferences presents further challenges in preserving previous knowledge. This paper examines the impact of personalized preference optimization on LLMs, revealing that the extent of knowledge loss varies significantly with preference heterogeneity. Although previous approaches have utilized the KL constraint between the reference model and the policy model, we observe that they fail to maintain general knowledge and alignment when facing personalized preferences. To this end, we introduce Base-Anchored Preference Optimization (BAPO), a simple yet effective approach that utilizes the initial responses of reference model to mitigate forgetting while accommodating personalized alignment. BAPO effectively adapts to diverse user preferences while minimally affecting global knowledge or general alignment. Our experiments demonstrate the efficacy of BAPO in various setups.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023) have been successfully aligned with human preferences across various applications, ranging from summarization tasks to enhancing reasoning capabilities (Ouyang et al., 2022; Stiennon et al., 2020; Tunstall et al., 2023b; Wang et al., 2023a). This alignment process involves collecting human feedback by presenting pairs of responses generated from the same user prompt and asking users to choose their preferred response (Bai et al., 2022; Cui et al., 2023; Lee et al., 2023; Cheng et al., 2023). The LLMs learn from this preference data to produce responses that better match human preferences, effectively addressing the challenge of converting complex human expectations into tangible training objectives (Ouyang et al., 2022; Ji et al., 2023; Xu et al., 2024). Known as preference optimization, this approach has become essential in the final stages of LLM training (Meta, 2024; Abdin et al., 2024; Jiang et al., 2024).

However, the common assumption in preference optimization is that all users share a uniform set of general preferences (Bai et al., 2022; Rafailov et al., 2024; Zheng et al., 2023), leading LLMs to align with an average of these preferences, as derived from collective feedback data (Jafari et al., 2024; Li et al., 2024; Guo et al., 2024). While effective for broadly accepted preferences like helpfulness and harmlessness, this approach does not account for the diversity of individual preferences in real-world scenarios (Jang et al., 2023; Zeng et al., 2023; Cheng et al., 2023; Zhong et al., 2024). For example, given the same context, one user might prefer a humorous response, while another might prefer a concise one. This reliance on averaged preferences often fails to capture the unique preferences of each user. This is known as the *Condorcet Paradox* (Gehrlein, 1983, 2002) in social choice theory, where no single response consistently satisfies all users, leading to non-transitive preferences (Wang et al., 2024a; Munos et al., 2023).

Recent studies have begun to tackle this challenge by fine-tuning instruction-tuned LLMs for personalized alignment (Jang et al., 2023; Zeng et al., 2023; Rame et al., 2024). Although these personalized preference optimization approaches enable support for diverse user preferences (Li et al., 2024; Zhong et al., 2024; Guo et al., 2024), the impact of learning to meet personalized preferences on previously acquired knowledge (Jin and Ren, 2024; Lu et al., 2024), such as global knowledge (Dou et al., 2023) and general alignment (Lin et al., 2023), remains underexplored.

In this work, we systematically analyze how personalizing LLMs according to diverse user preferences impacts their global knowledge and general alignment. Our findings reveal that the extent of knowledge loss heavily depends on these

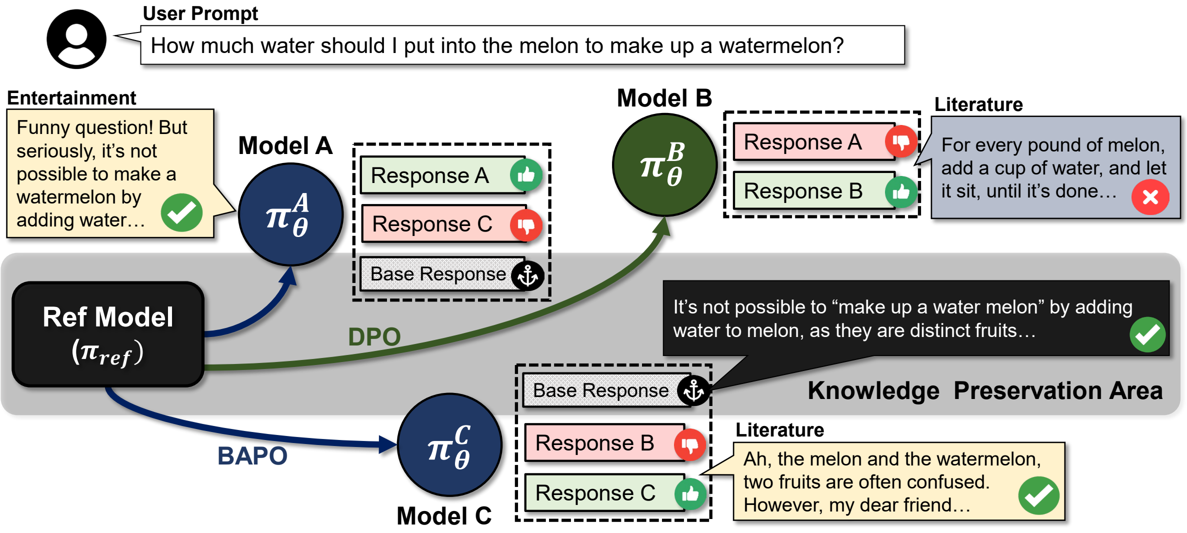

User Prompt
How much water should I put into the melon to make up a watermelon?
Entertainment
Funny question! But seriously, it's not possible to make a watermelon by adding water…
Model A
π_θ^A
Response A
Response C
Base Response
Model B
π_θ^B
Response A
Response B
Literature
For every pound of melon, add a cup of water, and let it sit, until it's done…
Ref Model (π_{ref})
DPO
It's not possible to "make up a water melon" by adding water to melon, as they are distinct fruits…
Knowledge Preservation Area
Base Response
Response B
Response C
Literature
Ah, the melon and the watermelon, two fruits are often confused. However, my dear friend…
BAPO
π_θ^C
Model C

Figure 1: Overview of Base-Anchored Preference Optimization (BAPO): For a given user prompt, the base response achieves general alignment. **Models A** and **Model C**, fine-tuned with BAPO, maintain this alignment by using the base response as an anchor. In contrast, **Model B**, fine-tuned with DPO, fails to preserve the knowledge from the base response, drifting away from the desired knowledge preservation area.

preferences, often inducing significant declines in specific areas of knowledge. This suggests that the conventional Kullback-Leibler (KL) constraints between the policy model and the reference model (Schulman et al., 2017; Rafailov et al., 2024), which is based solely on the tokens appearing in preferred or dispreferred responses (Pal et al., 2024; Azar et al., 2024; Zheng et al., 2023), fail to prevent the forgetting that occurs during personalized preference optimization.

To address this issue, we start by analyzing the initial responses, referred to as base responses, of instruction-tuned models to the given prompts and observe how their likelihood of producing these responses changes over training steps. We discover that personalizing preferences to enhance the distinction between preferred and dispreferred responses not only diminishes the likelihood of producing the dispreferred responses, but also lowers the likelihood of generating these base responses.

We hypothesize that aligning the reference and policy models, especially focusing on the tokens that appear in the base response, is essential for preserving global knowledge and ensuring general alignment. To this end, we introduce a novel preference optimization method named as Base-Anchored Preference Optimization (BAPO). BAPO aims to maintain the likelihood that the policy model will produce a base response originating from the reference model during personalized preference optimization.

Our main contributions are summarized as follows:

- We systematically analyze how diverse user preferences affect the global knowledge and alignment of instruction-tuned LLMs, finding that the extent of forgetting varies significantly with preference type. (Section 2)
- We propose a novel preference optimization method, BAPO, which utilizes the base response from the reference model to preserve existing knowledge in instruction-tuned LLMs during personalized preference optimization. (Section 3)
- We validate the efficacy of BAPO across various setups, demonstrating its effectiveness in preserving global knowledge and general alignment while adapting to diverse personalized preferences. (Section 4)

2 Personalized Preference Optimization

In this section, we first introduce preference optimization for LLM alignment. Next, we examine how personalized preferences impact the existing knowledge of instruction-tuned LLMs. We suggest that the typical KL-constraint in preference optimization is not effective in preventing forgetting.

2.1 Preliminary: Preference Optimization

Consider a dataset of pairwise preferences, denoted as $\mathcal{D} = \{x^i,\ y_w^i,\ y_l^i\}_{i=1}^N$. In this dataset, for each prompt x^i, the responses y_w^i and y_l^i represent the preferred (i.e., chosen) and not preferred (i.e., rejected) responses, respectively. Our goal is to optimize the policy model $\pi_\theta(y|x)$ to maximize the expected value of the ideal reward function $r^*(x,\ y)$ that aligns with human preferences:

$$\pi^* = \arg\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(\,\cdot\,|x)}\left[r^*(x,y)\right]. \quad (1)$$

A common approach to modeling the reward function is using the Bradley-Terry model (Bradley and Terry, 1952), which models the human preference distribution $p^*(y_1 \succ y_2 \mid x)$ as follows:

$$\frac{\exp(r^*(x,y_w))}{\exp(r^*(x,y_w)) + \exp(r^*(x,y_l))}. \quad (2)$$

Note that the Bradley-Terry model assumes that for each prompt x, the paired comparison probabilities $p(y_w \succ y_l \mid x)$ reflect a consistent human preference ordering across all possible responses, depending solely on the reward difference between responses $r^*(x,y_w) - r^*(x,y_l)$.

RLHF Using the reward function defined in Equation 2, Reinforced Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Stiennon et al., 2020) initially trains a reward model $r_\phi(x,y)$ that produces a single scalar prediction for the reward value. In the subsequent RL phase, this reward model guides the LLM to align the learned preference with the reference model π_{ref}, which has undergone supervised fine-tuning (SFT) from a pre-trained LLM as follows:

$$\begin{aligned}\mathcal{L}_{\text{RLHF}} = -\,&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\Big[r_\phi(x,y)\\ &-\beta\mathbb{D}_{\text{KL}}\big[\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x)\big]\Big]. \quad (3)\end{aligned}$$

where β corresponds to the regularization strength of KL-Divergence between the policy model π_θ and the reference model π_{ref}.

DPO By simplifying Equation 3, Direct Preference Optimization (DPO) (Rafailov et al., 2024) optimizes the maximum likelihood of the policy model π_θ without the need to train a separate explicit reward model as follows:

$$\begin{aligned}\mathcal{L}_{\text{DPO}} = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[&\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\\ &-\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Bigg)\Bigg]. \quad (4)\end{aligned}$$

Here, β represents the KL-regularization strength in RLHF. Note that in both RLHF and DPO, this KL-constraint depends only on the tokens appearing in y_w and y_l, the responses directly related to the preference ranking comparison.

2.2 Forgetting from Personalization

To understand how preference heterogeneity affects the extent of forgetting, we conduct an experimental study using heterogeneous preference datasets: P-Soups (Jang et al., 2023) and DSP (Cheng et al., 2023). We fine-tune the instruction-tuned Phi-3-mini (Abdin et al., 2024) model using DPO (Rafailov et al., 2024) with LoRA (Hu et al., 2021). Please refer to the detailed setups provided in Section 4.

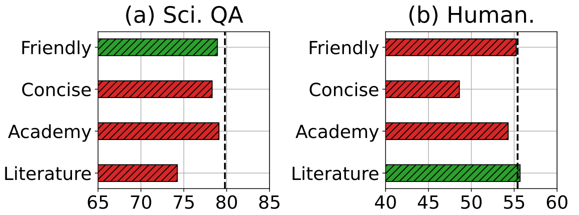


Figure 2: Performance on *Global Knowledge*: (a) Science QA and (b) MMLU - Humanities after personalization on diverse preferences. The black vertical dotted line indicates the base model performance.

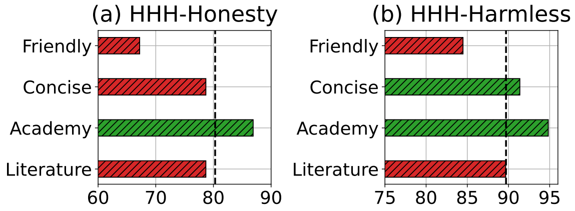


Figure 3: Performance on *General Alignment*: (a) HHH-Honesty and (b) HHH-Harmless after personalization on diverse preferences. The black vertical dotted line indicates the base model performance.

In Figure 2 and Figure 3, we evaluate performance changes after optimizing for specific preference types and present some representative results. The first two rows of each figure depict specific style preferences (e.g., Friendly or Concise) from the P-Soups dataset, while the last two rows showcase specific domain preferences (e.g., Academy or Literature) from the DSP dataset.

Our analysis reveals that the extent of forgetting global knowledge varies significantly with the prioritized preference type. For example, as shown in Figure 2, personalizing for the *Literature* domain preference leads to a notable decrease in performance on Science QA (Lu et al., 2022) datasets, while the *Academy* domain preference shows a lesser decline. Conversely, prioritizing the *Friendly* style preference enhances performance in Social Science within the MMLU (Hendrycks et al., 2020) datasets, whereas the *Concise* style preference causes a substantial drop. This variation is not limited to global knowledge but also extends to general alignment. For example, in Figure 3, the *Friendly* style preference significantly compromises Honesty in the HHH-Alignment (Askell et al., 2021) datasets. On the other hand, favoring the *Academy* domain preference rather improves it.

2.3 Knowledge in Base Response

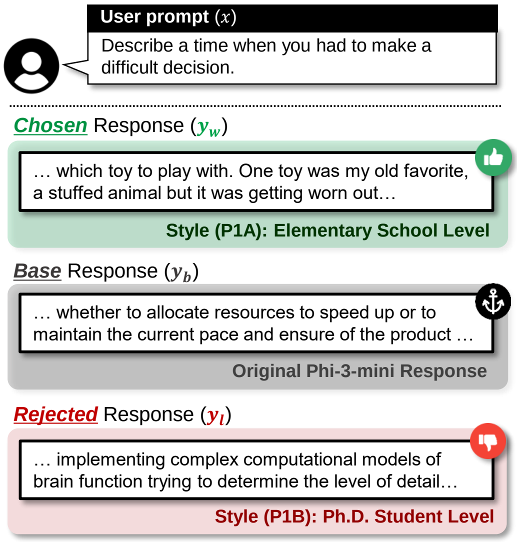
The significant variation in performance after personalized preference optimization suggests that the typical KL-divergence constraints (Ouyang et al., 2022; Rafailov et al., 2024; Zheng et al., 2023) used in general preference optimization (Bai et al., 2022; Tunstall et al., 2023b) still suffer from forgetting induced by preference heterogeneity.

We hypothesize that the original response from the initial reference model π_{ref}, which contains intact global knowledge and aligns with general alignment, is influenced by learning to meet diverse individual preferences. We take a closer look at personalized preference optimization to understand how adapting to heterogeneous preferences affects the likelihood of generating specific responses, represented by $\log[\pi_\theta(y_{(.)}|x) - \pi_{\text{ref}}(y_{(.)}|x)]$.

The observations in Figure 5 verify our conjecture. Personalizing preferences to enhance the distinction between the chosen response y_w (i.e., preferred) and the rejected response y_l (i.e., dispreferred) not only reduces the likelihood of producing the rejected response but also lowers the likelihood of generating base responses y_b. In this context, KL-divergence constraints between the reference and policy models on tokens found in these chosen and rejected responses do not help maintain the likelihood of base responses. Based on our findings, we consider an approach that leverages the tokens appearing in the base responses to encourage knowledge preservation during personalized preference optimization.

3 Proposed Method: BAPO

In this section, we introduce Base-Anchored Preference Optimization (BAPO). Our primary motivation is to use the initial response from the instruction-tuned model before it undergoes personalized preference optimization. By anchoring the policy model to this initial response, the personalized policy model can effectively retain its original global knowledge and general alignment while still accommodating diverse user preferences.


User prompt (x)
Chosen Response (y_w)
Style (P1A): Elementary School Level
Base Response (y_b)
Original Phi-3-mini Response
Rejected Response (y_l)
Style (P1B): Ph.D. Student Level

Describe a time when you had to make a difficult decision.

… which toy to play with. One toy was my old favorite, a stuffed animal but it was getting worn out…

… whether to allocate resources to speed up or to maintain the current pace and ensure of the product …

… implementing complex computational models of brain function trying to determine the level of detail…

Figure 4: An example of *Chosen*, *Base*, and *Rejected* responses to the same user prompt. Note that *Chosen* and *Rejected* in the figure assume that the user has the P1A (elementary school level) style preference.

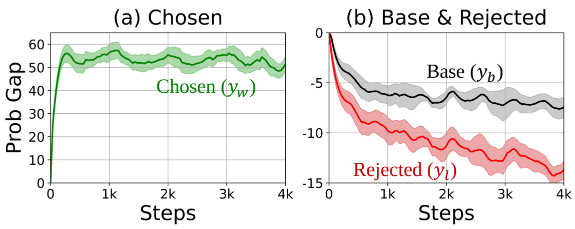


Figure 5: Average difference in reference model (π_{ref}) and policy model (π_θ) log probabilities for **Chosen**, **Base**, and **Rejected** responses during personalization across four domain preferences in DSP datasets.

3.1 Base-Anchored Preference Optimization

Consider the base response y_b from the reference model π_{ref} and the policy model π_θ, which we fine-tune for personalized preferences. The example in Figure 4 showcases how the chosen, base, and rejected responses differ for the same given user prompt. The core concept of BAPO is to preserve the knowledge contained in this base response during the optimization for diverse preferences.

Base Anchor BAPO ensures that the policy model's likelihood of producing the base response y_b ($\pi_\theta(y_b|x)$) remains closely aligned with that of the reference model ($\pi_{\text{ref}}(y_b|x)$):

$$\mathcal{L}_{\text{Anchor}} = \max\left(0,\ \log\frac{\pi_{\text{ref}}(y_b|x)}{\pi_\theta\,(y_b|x)}\right). \qquad (5)$$

Note that the base anchor loss $\mathcal{L}_{\text{Anchor}}$ becomes 0 if the policy model π_θ assigns a higher likelihood to the base response y_b than the reference model does ($\pi_\theta(y_b|x) > \pi_{\text{ref}}(y_b|x)$). Intuitively, if the policy model is already more confident in the base response than the reference model, there's no need to penalize it further. The BAPO objective $\mathcal{L}_{\text{BAPO}}$ is defined as follows:

$$\mathcal{L}_{\text{BAPO}} = \mathcal{L}_{\text{DPO}} + \lambda \cdot \mathcal{L}_{\text{Anchor}} \,. \tag{6}$$

Here, λ controls the strength of the anchoring effect. In our main experiments, we set $\lambda = 5$.

3.2 Theoretical Analysis

We assess the impact of BAPO on personalized preferences by analyzing how information from base responses aids in aligning personal preferences. We assume linear utility and reward functions, with their respective unknown parameters having distinct nonzero components.

Assumption 1. *A utility function $G^\star$ exists for general alignment with global knowledge, and a reward function $L^\star$ measures personal alignment.*

Assumption 2. *For the response y and context x, the functions $G^\star$ and $L^\star$ are linear, defined as: $G^\star(x, y) = \langle \boldsymbol{\phi}(x, y), \boldsymbol{\theta}^G_\star \rangle$ and $L^\star(x, y) = \langle \boldsymbol{\phi}(x, y), \boldsymbol{\theta}^L_\star \rangle$ where $\boldsymbol{\phi}(x, y)$ is a d dimensional feature vector of y and x. Additionally, $\boldsymbol{\theta}^G_\star, \boldsymbol{\theta}^L_\star$ have non-intersecting nonzero components on d-k and the k dimensions respectively.*

Proposition 1. *Given the information of θ_G is known. Then, the sample complexity for estimating θ_L reduces from $O(\sqrt{d})$ to $O(\sqrt{k})$.*

The proof is provided in Appendix C. This proposition suggests that the last k-dimensional subspace, which governs personalized rewards, is typically much smaller than the first $d - k$-dimensional subspace responsible for general alignment. Consequently, this reduction in complexity of the parameter space influencing personalized rewards allows for more efficient estimation with fewer samples. In practice, personalization is often driven by a smaller, critical set of features compared to those affecting broader alignment criteria.

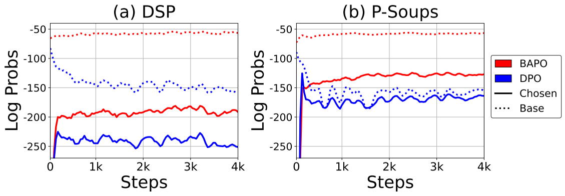


Figure 6: Evolution of log probabilities for *Chosen* and *Base* responses during preference optimization. The results are averaged across different preference types.

The results in Figure 6 empirically support our analysis. We analyze the changes in log probabilities of responses under DPO and BAPO. For experimental details, see Section 4. While log probabilities for the *base* response decrease under DPO, BAPO maintains them, enhancing stability throughout the preference optimization process. This consistency enables the model to assign significantly higher log probabilities to the *chosen* response, thus speeding up the learning process.

4 Experiment

4.1 Experimental Setups

Datasets We use two preference datasets for personalized preference optimization: P-Soups (Jang et al., 2023) and DSP (Cheng et al., 2023).

- **Personalized Soups (P-Soups)** include *Style* preferences , organized into three dimensions: P1, P2, and P3. Each dimension features two contrasting types, A and B. In Table 1, we briefly describe the preference types.
- **Domain Specific Preference (DSP)** includes *Domains* preferences: Academy, *Business*, Entertainment, and Literature & Art.

Each dataset is composed of user queries and a set of responses for each query. In our pairwise preference format, for each user query x, we select a response that aligns with a specific preference as the chosen response y_w. Responses from other preferences are designated as rejected responses y_l. More details are provided in Appendix A.

Table 1: Response Preferences of the P-Soups dataset.

Dimension	**Type**	**Response Preference**
(P1) Expertise	A	Elementary school level.
	B	PhD-level expertise in the field.
(P2) Verbosity	A	Concise, without being verbose.
	B	Informative and fully detailed.
(P3) Style	A	Friendly, witty and humorous.
	B	Answer in an unfriendly manner.

Learning Setups In our main experiments, we primarily use a Phi-3 model (Abdin et al., 2024), specifically its instruction-tuned version referred to as *Phi-3-mini-128k-instruct*, with 3.82 billion parameters. This model has been enhanced with DPO to align with general human preferences and safety guidelines, following the SFT stage. Each personalized model, aimed at aligning with a specific preference type, is fine-tuned using Q-LoRA (Dettmers et al., 2024), a quantized variant of LoRA (Hu et al.,

Table 2: Performance on global knowledge datasets after fine-tuning for personalized preferences. The term 'Base' refers to the initial performance of the Phi-3-mini model before personalization. Values in parentheses represent the standard deviation across different preference types: 6 for the P-Soups datasets and 4 for the DSP datasets.

Preference Dataset: P-Soups (Jang et al., 2023)

Method	**PIQA**	**SIQA**	**ARC-c**	**Sci. QA**	**Comm.**	**MMLU**			
						STEM	Social	Human	Other
Base	78.0	72.0	81.9	79.8	46.0	47.7	62.7	55.4	65.3
DPO	77.0$_{(2.8)}$	69.7$_{(0.1)}$	81.6$_{(0.9)}$	78.5$_{(2.1)}$	68.7$_{(1.6)}$	49.2$_{(1.9)}$	65.7$_{(8.4)}$	53.7$_{(2.5)}$	64.8$_{(5.5)}$
RSO	76.4$_{(2.4)}$	71.1$_{(0.7)}$	82.3$_{(1.3)}$	80.8$_{(1.5)}$	70.5$_{(0.7)}$	50.3$_{(1.2)}$	67.6$_{(6.5)}$	**55.3**$_{(0.5)}$	66.9$_{(1.9)}$
IPO	62.9$_{(11.0)}$	50.2$_{(13.6)}$	51.2$_{(24.2)}$	52.2$_{(20.8)}$	41.8$_{(19.2)}$	38.2$_{(11.3)}$	51.4$_{(19.8)}$	39.8$_{(10.3)}$	48.2$_{(16.6)}$
DPOP	76.1$_{(1.7)}$	70.8$_{(0.8)}$	81.8$_{(1.7)}$	80.6$_{(1.1)}$	70.6$_{(1.4)}$	50.5$_{(1.6)}$	68.3$_{(4.5)}$	54.7$_{(0.8)}$	**67.4**$_{(1.4)}$
ORPO	66.6$_{(6.0)}$	61.5$_{(4.0)}$	71.5$_{(2.9)}$	68.5$_{(6.4)}$	60.7$_{(2.9)}$	**51.2**$_{(1.5)}$	**70.0**$_{(2.7)}$	50.5$_{(0.8)}$	65.1$_{(1.1)}$
BAPO	**78.0**$_{(1.5)}$	**71.6**$_{(0.8)}$	**82.5**$_{(1.1)}$	**81.0**$_{(0.6)}$	**71.5**$_{(0.7)}$	49.7$_{(1.0)}$	66.1$_{(3.0)}$	55.1$_{(0.6)}$	66.4$_{(1.0)}$

Preference Dataset: DSP (Cheng et al., 2023)

Method	**PIQA**	**SIQA**	**ARC-c**	**Sci. QA**	**Comm.**	**MMLU**			
						STEM	Social	Human	Other
Base	78.0	72.0	81.9	79.8	46.0	47.7	62.7	55.4	65.3
DPO	77.6$_{(0.6)}$	70.2$_{(1.3)}$	81.7$_{(1.1)}$	78.5$_{(3.1)}$	69.1$_{(0.9)}$	49.9$_{(3.0)}$	65.2$_{(9.2)}$	54.1$_{(1.9)}$	66.3$_{(2.2)}$
RSO	77.2$_{(0.9)}$	71.2$_{(0.5)}$	81.9$_{(1.0)}$	80.2$_{(0.9)}$	70.4$_{(1.1)}$	50.5$_{(2.0)}$	66.8$_{(6.2)}$	**55.9**$_{(0.6)}$	66.7$_{(1.5)}$
IPO	50.9$_{(1.3)}$	35.5$_{(3.1)}$	19.8$_{(13.1)}$	24.8$_{(9.4)}$	25.8$_{(6.6)}$	28.0$_{(1.9)}$	30.2$_{(4.8)}$	30.3$_{(2.3)}$	30.0$_{(2.5)}$
DPOP	77.9$_{(1.5)}$	70.9$_{(1.0)}$	81.6$_{(2.4)}$	80.1$_{(0.4)}$	69.8$_{(1.5)}$	50.4$_{(1.9)}$	67.4$_{(7.4)}$	55.1$_{(1.4)}$	67.2$_{(1.5)}$
ORPO	66.6$_{(6.0)}$	61.5$_{(4.0)}$	71.5$_{(2.9)}$	68.9$_{(6.4)}$	60.7$_{(2.9)}$	**51.1**$_{(1.5)}$	**69.9**$_{(2.7)}$	50.5$_{(0.7)}$	65.1$_{(1.1)}$
BAPO	**78.0**$_{(1.2)}$	**71.7**$_{(0.3)}$	**83.2**$_{(0.5)}$	**80.9**$_{(1.1)}$	**71.1**$_{(0.7)}$	50.2$_{(0.9)}$	66.1$_{(2.9)}$	55.6$_{(0.4)}$	**67.1**$_{(0.4)}$

2021). We utilize a 4-bit normalized float (nf4) and double quantization with bf-16 to enhance computational efficiency. The LoRA settings, including a rank of $r = 32$ and $\alpha = 64$ with a dropout rate of 0.05, are applied to all linear layer weights of the model. Each personalized model is trained for a single epoch on feedback pairs using an effective batch size of 8. The learning rate, set at 5e-5, follows the conventional training recipe (Tunstall et al., 2023a,b). The learning rate is decayed using a cosine scheduler (Loshchilov and Hutter, 2016).

Evaluation To evaluate the extent of forgetting after personalized preferences optimization, we divide the datasets into two categories: (i) Global Knowledge and (ii) General Alignment. In Global Knowledge, we assess the model's prior knowledge of world understanding through closed-book question-answering tasks. More specifically, we use commonsense datasets such as PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), Arc-Challenge (Clark et al., 2018), Science QA (Lu et al., 2022), Commonsense QA (Talmor et al., 2018), and 5-shot MMLU (Hendrycks et al., 2020). Since the Science QA dataset includes visual tasks, we utilize text-only questions from this dataset. For assessing General Alignment, we conduct evaluations using the HHH-Alignment (Askell et al., 2021) datasets, which consist of categories focused on helpfulness, harmlessness, and honesty.

4.2 Performance on Knowledge Preservation

In Table 2, we evaluate how fine-tuning the Phi-3-mini model affects its performance in the *Global Knowledge* across diverse preferences. This evaluation includes the DPO (Rafailov et al., 2024) and other preference optimization methods such as RSO (Liu et al., 2023), IPO (Azar et al., 2024), DPOP (Pal et al., 2024), and ORPO (Hong et al., 2024). The 'Base' in the table indicates the initial performance of the Phi-3-mini model.

The results show that the baseline methods significantly declines the performance on global knowledge datasets, while our BAPO method effectively maintains consistent performance across various evaluated datasets. We highlight that BAPO has advantageous characteristics, keeping performance variations due to preference heterogeneity remarkably low. In contrast to the baseline method, which exhibits high fluctuation and large variance across varying preferences, BAPO provides more reliable results with minimal variation.

We observe that personalized preferences do not necessarily lead to forgetting. In fact, they can sometimes enhance performance on certain datasets based on the types of preferences involved. For example, fine-tuning with domain-specific preferences in DSP datasets often results in improved performance on MMLU datasets. Notably, the ORPO method, which does not use a reference model during preference optimization, consistently outperforms other approaches in such cases. This implies there exists a natural trade-off in the use of reference models between preserving existing knowledge and acquiring new knowledge.

4.3 Performance on Alignment

General Alignment Table 3 presents the evaluation of general and personalized alignment after the personalized preference optimization. Ideally, the personalized model should maintain general alignment while accommodating its specific personalized preferences. We first evaluate whether the fine-tuned models maintain general alignment after personalization by using the HHH-Alignment datasets (Askell et al., 2021). The results show that BAPO effectively preserves the level of general alignment seen in the initial reference model, with minimal variation across diverse preference types.

Personalized Alignment We evaluate personalized alignment by measuring Reward Accuracy, which assesses whether the policy model prefers the selected response y_w over the dis-preferred response y_l for user prompts x in the validation set. The results show that BAPO more effectively accommodate personalized performance compared to other baselines. This demonstrates that utilizing base responses can boost sample efficiency for reward signals related to personalized preferences.

Table 3: Performance on general/personalized alignment after fine-tuning for personalized preferences.

	Preference Datasets: P-Soups			
Method	**HHH-Alignment**			**Rwd Acc.**
	Helpful	Harmless	Honest	
Base	84.7	89.7	80.3	-
DPO	81.4$_{(6.3)}$	87.9$_{(5.3)}$	76.0$_{(4.8)}$	93.2$_{(3.1)}$
RSO	83.1$_{(3.6)}$	**89.7**$_{(3.6)}$	78.4$_{(3.5)}$	93.0$_{(3.1)}$
IPO	70.3$_{(11.7)}$	73.3$_{(16.5)}$	68.0$_{(14.2)}$	89.6$_{(7.6)}$
DPOP	83.1$_{(4.7)}$	**89.7**$_{(4.4)}$	79.2$_{(2.9)}$	92.2$_{(3.4)}$
ORPO	81.9$_{(2.8)}$	83.6$_{(5.5)}$	74.9$_{(3.1)}$	84.6$_{(2.1)}$
BAPO	**84.0**$_{(1.8)}$	87.9$_{(4.2)}$	**80.6**$_{(1.2)}$	**97.8**$_{(2.3)}$
	Preference Datasets: DSP			
Method	**HHH-Alignment**			**Rwd Acc.**
	Helpful	Harmless	Honest	
Base	84.7	89.7	80.3	-
DPO	82.6$_{(2.9)}$	91.8$_{(3.6)}$	80.7$_{(4.9)}$	87.1$_{(4.0)}$
RSO	81.6$_{(1.3)}$	91.4$_{(2.7)}$	**81.4**$_{(2.4)}$	87.1$_{(4.3)}$
IPO	72.9$_{(7.3)}$	74.6$_{(8.6)}$	68.0$_{(7.6)}$	86.2$_{(3.7)}$
DPOP	72.9$_{(7.3)}$	74.6$_{(8.6)}$	68.0$_{(7.6)}$	87.3$_{(4.1)}$
ORPO	83.1$_{(1.4)}$	92.2$_{(3.3)}$	77.1$_{(3.8)}$	79.2$_{(3.2)}$
BAPO	**85.2**$_{(1.6)}$	**93.1**$_{(1.4)}$	81.2$_{(3.4)}$	**97.0**$_{(0.1)}$

4.4 Ablation Study

Model Architecture We conduct further experiments with the instruction-tuned Gemma-2B model (Team et al., 2024), referred to as *Gemma-2B-it*, which has also undergone RLHF for general alignment after the SFT stage. The results presented in Table 4 validates the robust efficacy of BAPO across different model architectures.

Table 4: Performance of the Gemma-2B-it model after fine-tuning for personalized preferences. Scores for MMLU and HHH are averaged across all categories.

	Preference Datasets: P-Soups				
Method	**Sci. QA**	**Comm.**	**MMLU**	**HHH**	**Rwd Acc**
Base	51.9	46.0	28.8	67.4	-
DPO (P-Soups)	51.3	44.6	28.0	63.9	**95.4**$_{(3.5)}$
BAPO (P-Soups)	**51.5**	**44.9**	**28.1**	**64.2**	**95.4**$_{(3.4)}$
DPO (DSP)	51.5	45.4	28.3	66.1	98.2$_{(0.6)}$
BAPO (DSP)	**51.6**	**45.7**	**28.5**	**66.1**	**98.3**$_{(0.8)}$

Effect of Anchoring Strength The impact of anchoring strength λ on BAPO is illustrated in Figure 7. The performance change is measured against the *Base* model. As depicted in Figure 7(a), increasing anchoring strength generally enhances global knowledge preservation, although the effect plateaus at higher strengths. Additionally, Figure 7(b) shows the results on alignment, indicating that while base anchoring in BAPO improves both types of alignment, its effectiveness also reaches a limit at higher strengths.

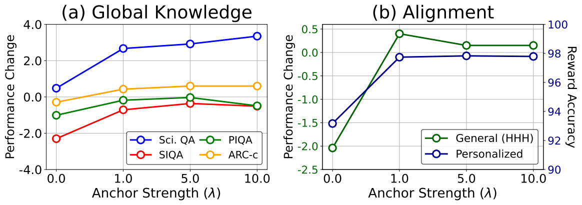


Figure 7: Performance of BAPO on P-Soups datasets with varying anchoring strength values λ. Note that setting $\lambda = 0$ is equivalent to using the vanilla DPO.

Effect of LoRA Rank In Figure 8, we explore the impact of varying LoRA rank r on knowledge preservation and alignment, setting the scaling factor α such that $\frac{\alpha}{r} = 2$. As shown in Figure 8(a) and Figure 8(b), an increase in rank r leads to more pronounced forgetting in both global knowledge and general alignment. This finding supports recent research suggesting that a lower LoRA rank reduces the rate of learning but also minimizes forgetting (Biderman et al., 2024). Nevertheless, BAPO effectively preserves knowledge even as LoRA rank increases and enhances accommodation of personalized preferences at higher ranks, where the increased learning capacity could otherwise detract from learning personalized preferences.

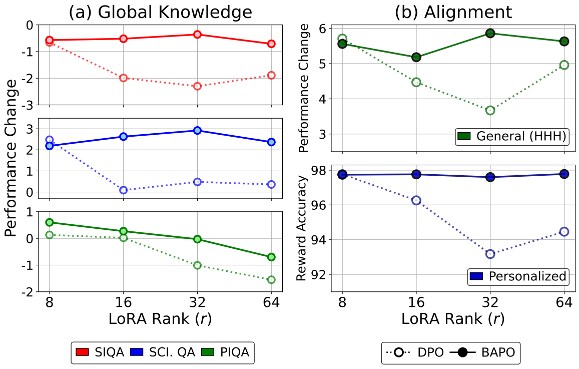


Figure 8: Performance of BAPO on P-Soups datasets with varying LoRA rank values r.

5 Related Work

Learning from Human Feedback Often employed as the final stage of instruction tuning (Tunstall et al., 2023b; Jiang et al., 2024; Meta, 2024), learning from human feedback refines the policy language model to generate responses that align with human preferences (Ji et al., 2023; Zheng et al., 2023). This process usually involves collecting human preferences for pairs of candidate responses to differentiate between those that are preferred and those that are dispreferred. The two main approaches used are Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2024). RLHF involves training a separate reward model that is then utilized in the subsequent reinforcement learning phase (Bai et al., 2022; Tunstall et al., 2023b). In contrast, DPO does not rely on a reward model but directly establishes a mapping between the reward function and the optimization objective (Pal et al., 2024; Liu et al., 2023). Our work specifically addresses the challenge of accommodating personalized preferences within the DPO framework. We focus on maintaining the likelihood of base responses from the reference model during the personalization process, ensuring the preservation of global knowledge and general alignment.

Forgetting in LLM Fine-tuning The issue of forgetting previously acquired knowledge during the fine-tuning on heterogeneous data has been extensively discussed (Wang et al., 2023b, 2024b), highlighting a fundamental trade-off between preserving old knowledge and acquiring new knowledge (Parisi et al., 2019; McCloskey and Cohen, 1989). In the context of LLMs, recent research has shown that the knowledge from pre-training can be compromised by supervised fine-tuning (SFT) on instruction data (Dou et al., 2023; Dong et al., 2023). A similar issue, known as *alignment tax*, occurs in preference optimization (Lin et al., 2023; Lu et al., 2024). While prior research has addressed the forgetting induced by the LLM fine-tuning (Biderman et al., 2024; Luo et al., 2023), the effects of accommodating specific user preferences and the impact of their heterogeneity remain largely unexplored. In our study, we investigate how personalized preference optimization affects both global knowledge and general alignment.

Personalized Preference in LLM The use of a single scalar reward to represent user preferences presents a significant limitation when users have diverse and conflicting preferences (Ji et al., 2023; Zheng et al., 2023). To address this issue, some studies have explored clustering users who presumably share the same reward (Chakraborty et al., 2024; Park et al., 2024). Others have considered defining a reward function with multiple objective dimensions (Jafari et al., 2024; Yang et al., 2024) to achieve Pareto optimality among them (Guo et al., 2024; Zhong et al., 2024; Wang et al., 2024a). Additionally, some approaches involve merging model parameters trained for each dimension to accommodate the diverse combinations expressed by those dimension (Jang et al., 2023; Rame et al., 2024). In our study, we focus on the impact of preference heterogeneity on forgetting during personalized preference optimization, assuming that each user has a specific, definitive type of preference.

6 Conclusion

This study explores the degree and nature of forgetting caused by personalized preference optimization in instruction-tuned LLMs. Our findings indicate a reduced likelihood of generating original responses, alongside a decrease in the generation of dispreferred responses. To address this, we introduce Base-Anchored Preference Optimization (BAPO), a method that anchors the likelihood of base responses during the preference optimization process. This approach effectively preserves global knowledge and general alignment while successfully accommodating personalized preferences. We have conducted extensive experiments to validate the efficacy of BAPO and its benefits.

Limitations While BAPO effectively preserves existing knowledge by leveraging the base response, it is important to note that if the base model is biased, the fine-tuned personalized model may also exhibit a similar bias. This consideration is crucial for machine learning practitioners. In our experiments, we utilized Q-LoRA for fine-tuning. Although we conducted an ablation study varying model capacity by adjusting the LoRA rank, a full-finetune might show different tendencies. Nonetheless, using LoRA fine-tuning for personalization is a common approach in LLM context. Regarding computational costs, although BAPO requires the use of a base response, potentially increasing memory and computational demands during training, most of these costs can be mitigated by pre-generating the base responses and caching them as offline datasets. Concerning the anchoring strength hyperparameter, λ, the extent and scope of forgetting may vary based on the type of knowledge and the personalized preference types. Ideally, the λ value should be adaptively assigned, but this aspect is left for future research.

Ethical Considerations In developing and implementing our approach to align LLMs with personalized preferences, we must consider the potential implications. While BAPO aims to accommodate diverse user preferences, it is crucial to ensure this customization does not unintentionally reinforce harmful biases or perpetuate discrimination. Additionally, we must protect user privacy and data security, ensuring that personalization does not expose sensitive information or compromise user anonymity. Finally, maintaining a balance between personalized alignment and the integrity of general knowledge is essential to avoid scenarios where excessive personalization might result in misinformation or a loss of objective truth.

Use of AI Assistants We utilize Copilot[1] for the development of our code pipeline, primarily for its auto-completion capabilities. For drafting the paper, we employ ChatGPT[2] to review the content, focusing on identifying grammatical errors and awkward expressions. However, the core content of the paper is original. We do not rely on AI assistants to generate any specific content that is closely related to the main claims of our research.

[1] https://github.com/features/copilot

[2] https://chatgpt.com/

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves a reward: Learning customized human preferences. *arXiv preprint arXiv:2309.03126*.

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

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. 2023. The art of balancing: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*.

William V Gehrlein. 1983. Condorcet's paradox. *Theory and decision*, 15(2):161–197.

William V Gehrlein. 2002. Condorcet's paradox and the likelihood of its occurrence: different perspectives on balanced preferences. *Theory and decision*, 52(2):171–199.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Jiexin Wang, Huimin Chen, Bowen Sun, Ruobing Xie, Jie Zhou, Yankai Lin, et al. 2024. Controllable preference optimization: Toward controllable multi-objective alignment. *arXiv preprint arXiv:2402.19085*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yasaman Jafari, Dheeraj Mekala, Rose Yu, and Taylor Berg-Kirkpatrick. 2024. Morl-prompt: An empirical analysis of multi-objective reinforcement learning for discrete prompt optimization. *arXiv preprint arXiv:2402.11711*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Xisen Jin and Xiang Ren. 2024. What will my model forget? forecasting forgotten examples in language model refinement. *arXiv preprint arXiv:2402.01865*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Xinyu Li, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. 2024. Online merging optimizers for boosting rewards and mitigating tax in alignment. *arXiv preprint arXiv:2405.17931*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. 2023. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. 2024. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Shengyi Huang, Kashif Rasul, Alexander M. Rush, and Thomas Wolf. 2023a. The alignment handbook. `https://github.com/huggingface/alignment-handbook`.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023b. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024b. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023a. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*.

Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. 2023b. A comprehensive survey of forgetting in deep learning beyond continual learning. *arXiv preprint arXiv:2307.09218*.

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

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.

Dun Zeng, Yong Dai, Pengyu Cheng, Tianhao Hu, Wanshun Chen, Nan Du, and Zenglin Xu. 2023. On diversified preferences of large language model alignment. *arXiv preprint arXiv:2312.07401*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for llms. *arXiv preprint arXiv:2402.02030*.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR.

A Dataset details

In our experiments, we utilize two datasets written in English for personalized preference optimization: the DSP (Domain-Specific Preference) and P-Soups datasets. The DSP dataset, proposed by (Cheng et al., 2023), comprises 13,000 prompts selected from the 52,000 Alpaca datasets (Taori et al., 2023). It features preferred responses tailored to specific queries across four practical domains: Academy, Business, Entertainment, and Literature & Art. Each prompt includes five responses: one from each of the four domains and the original Alpaca response. In our pairwise preference setup, the response from the corresponding domain is selected as the preferred one, while the others are considered rejected. This results in 52,000 pairs per dataset, with 2,000 pairs designated for the validation split. The P-Soups dataset (Jang et al., 2023), simulated by GPT-4, consists of pairwise feedback data where the AI is instructed to choose the better of two candidate responses. This dataset builds on prompt instances from Alpaca-GPT4, with additional prompts provided to ensure consistency in preference criteria. To facilitate experiments that require the same prompt across different preference types, we exclude the additional prompts introduced by P-Soups, keeping only the original prompts from the Alpaca dataset. The P-Soups dataset categorizes six conflicting preferences into three dimensions: expertise, informativeness, and friendliness, resulting in six preference combinations. Out of the 10,000 Alpaca-GPT4 prompts (Peng et al., 2023), we derive between 47,000 to 49,000 pairwise feedback entries per preference type, allocating 45,000 samples for the training split and the remainder for the validation split. Note that while response lengths may vary among preferences in the P-Soups dataset, they remain fairly consistent across different domains in the DSP dataset.

B Resources & Others

We use six RTX A6000 48GB GPU cards for our experiments, although we do not employ multi-GPU training. The GPU hours required for each run vary, but typically, running one epoch on the Phi-3-mini model with a LoRA rank of 32, including time for evaluation, takes about 10 hours. To reproduce all the results in this paper, approximately 1,600 GPU hours are required. We employ the Hugging Face Transformers library (Wolf et al., 2019) for the overall code.

C Proof of Proposition 1

First, observe that under the linear reward model assumption, maximum likelihood estimation (MLE) of unknown parameter $\theta_\star$ is obtained as follows:

$$\hat{\theta}_{\text{MLE}} = \text{argmin}_{\theta \in \theta_B} \mathcal{L}_{\text{BT}}(\theta) = \text{argmin}_{\theta \in \theta_B} \sum_{i=1}^{n} -\log\left(\sigma\left(\langle \theta, \phi(x^i, y_w^i) - \phi(x^i, y_l^i) \rangle\right)\right), \quad (7)$$

where, σ is a sigmoid function and (x^i, y_w^i, y_l^i) denotes i-th (out of n) preference sample with context x^i, winning and losing response y_w^i, y_l^i respectively. Also, and $\theta_B = \{\theta \in \mathbb{R}^d : \|\theta\| \leqslant B\}$.

In order to prove the **Proposition 1**, we bring the latest result for the sample complexity bound for the linear preference model which is provided in the **Lemma 3.1** of (Zhu et al., 2023). We restate the lemma here for the completeness.

Lemma 1. *Assume, $\phi(y, x) \leqslant L$ for all possible response, context pairs (y, x) and $\theta_\star \leqslant B$ is unknown parameter for linear model in eq. 7. Then, for $\lambda > 0$, constant $C' > 0$ and estimator $\hat{\theta}_{MLE}$ for the Bradley-Terry model loss (eq. 7), the following confidence bound holds with probability $1 - \delta$:*

$$\|\hat{\theta}_{MLE} - \theta_\star\|_{\Sigma_D + \lambda I} \leqslant C' \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n} + \lambda B^2}. \quad (8)$$

Here, $\Sigma_D = \frac{1}{n}\sum_{i=1}^{n}(\phi(x^i, y_w^i) - \phi(x^i, y_l^i))(\phi(x^i, y_w^i) - \phi(x^i, y_l^i))^\top$, $\gamma = 1/(2 + \exp(-LB) + \exp(LB))$.

Now, suppose we have full information of $\theta_\star^G$ and without loss of generality, $\theta_\star^G$ has nonzero values on the first d-k dimensions. (Note that according to **Assumption. 2**, this automatically implies that $\theta_\star^L$ has nonzero components only on the last k dimensions.) Here, for the ease of analysis, we truncate only nonzero parts of $\theta_\star^L$ to make it a k-dimensional vector. Also, denote $\phi^L(y, x) = \phi(y, x)_{d-k+1:d}$ be the last k dimensional part of feature vector that governs personalization reward L. With this, we can calculate MLE of BT model only for the last k dimensional components in the following way:

$$\hat{\theta}^L_{\text{MLE}} = \text{argmin}_{\theta^L \in \theta_{B'}} \mathcal{L}_{\text{BT}}(\theta^L) = \text{argmin}_{\theta^L \in \theta_{B'}} \sum_{i=1}^{n} -\log\left(\sigma\left(\langle \theta^L, \phi^L(x^i, y_w^i) - \phi^L(x^i, y_l^i) \rangle\right)\right), \quad (9)$$

where $\theta_{B'} = \{\theta \in \mathbb{R}^k : \|\theta\| \leqslant B'\}$.

Now, with **Lemma. 1** and **Assumptions 1, 2**, it is easy to see the following confidence bound for the $\theta_\star^L$ holds by the following lemma.

Lemma 2. *With $\phi(y, x)_{d-k+1:d} \leqslant L'$ for all possible response, context pairs (y, x) and $\theta_\star^L \leqslant B'$. Then, for $\lambda' > 0$, constant $C'' > 0$ and unknown parameter $\hat{\theta}^L_{MLE}$ from the modified Bradley-Terry model loss (eq. 9), the following confidence bound holds with probability $1 - \delta$:*

$$\|\hat{\theta}^L_{MLE} - \theta^L_\star\|_{\Sigma^L_D + \lambda' I} \leqslant C'' \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma'^2 n} + \lambda' B'^2}. \quad (10)$$

Here, $\Sigma_D^L = \frac{1}{n}\sum_{i=1}^{n}(\phi^L(x^i, y_w^i) - \phi^L(x^i, y_l^i))(\phi^L(x^i, y_w^i) - \phi^L(x^i, y_l^i))^\top$, $\gamma' = 1/(2 + \exp(-L'B') + \exp(L'B'))$.

Combining **Lemma 1** and **Lemma 2**, we see that **Proposition. 1** holds.