

# FedMOPA: Federated Multi-Objective Preference Alignment for Large Language Models

Anonymous ACL submission

## Abstract

Aligning Large Language Models (LLMs) with diverse and often conflicting human preferences is a critical challenge, particularly when preference data is distributed across privacy-sensitive silos. In this paper, we propose **FedMOPA**, a novel framework that integrates federated learning with multi-objective optimization to align LLMs with heterogeneous user preferences while preserving data privacy. Our approach introduces a unified, preference-conditioned model that can dynamically adapt to varying trade-offs among client preferences at inference time, eliminating the need for retraining. To address the communication overhead associated with fine-tuning LLMs in a federated setting, we propose **TriLoRA**, a novel conditional LoRA variant that efficiently incorporates preference information into low-rank updates. Furthermore, we design an alternating optimization strategy to mitigate aggregation errors inherent in the federated averaging of multiplicative parameters. We provide theoretical guarantees for the convergence of FedMOPA and its ability to achieve the Pareto front under certain conditions. Extensive experiments on real-world datasets, including safety and helpfulness alignment, demonstrate the effectiveness of our method. Our code is available at <http://anonymous.4open.science/r/FedMOPA-555>.

## 1 Introduction

Aligning Large Language Models (LLMs) with human values is a cornerstone for developing safe and reliable AI (Wang et al., 2023; Casper et al., 2023). In practice, human preferences are inherently complex and often conflicting, reflecting the diversity of human values and the contextual nature of decision-making. For instance, a user might desire an LLM that is simultaneously helpful, harmless, and humorous—a set of competing objectives that single-objective alignment methods (Ziegler

et al., 2019; Rafailov et al., 2023) struggle to balance. The challenge is further compounded by the fact that different users and applications may prioritize these objectives differently, requiring models that can adapt to varying preference profiles.

While multi-objective alignment methods (Yang et al., 2024b; Zhong et al., 2024) enable LLMs to dynamically adjust trade-offs among different preference dimensions, they assume that all preference data can be accessed simultaneously. However, in many real-world applications, these preference data may be distributed across different institutions (e.g., client 1 owns helpful preference data, client 2 owns harmless preference data, and client 3 owns humorous preference data), and data sharing between these entities is often restricted due to privacy and regulatory concerns. This distributed preference landscape raises a critical research question: *How can we align a single LLM with multiple, conflicting user preferences in a privacy-preserving manner?*

To address privacy concerns, we propose utilizing Federated Learning (FL) (McMahan et al., 2017), which enables collaborative and decentralized training of models across multiple institutions without sharing personal data externally. FL has emerged as a promising paradigm for privacy-preserving machine learning, allowing participants to collectively train a shared model while keeping their data local. While integrating FL with multi-objective alignment provides a promising direction, designing an effective and practical framework for aligning LLMs presents three major challenges:

- **Challenge 1: Unified Model for Diverse Preferences.** Standard FL methods typically produce a single global model representing an average compromise, or distinct personalized models fixed to local preferences. Neither approach can dynamically satisfy varying user needs at inference time. While training sepa-

084	rate models for every possible preference combination is a theoretical solution, it is computationally intractable. Thus, a critical challenge is to develop a unified framework that can dynamically adapt to the entire spectrum of user preferences within a single model, allowing dynamic adaptation to arbitrary preference trade-offs without the need for retraining.	134
085		135
086		136
087		137
088		138
089		139
090		140
091		141
092	• <b>Challenge 2: Prohibitive Communication Overhead.</b> Fine-tuning LLMs typically involves updating billions of parameters, which incurs prohibitive communication costs when transmitting these parameters between clients and the central server in a federated setting, making the process infeasible for real-world deployment. Therefore, parameter-efficient fine-tuning techniques (e.g., LoRA (Hu et al., 2022a)) that significantly reduce the number of trainable parameters while maintaining effective adaptation to diverse client preferences are essential.	142
093		143
094		144
095		145
096		146
097		147
098		148
099		149
100		150
101		151
102		152
103		153
104		154
105	• <b>Challenge 3: Aggregation Error of LoRA in FL.</b> While LoRA and its variants are parameter-efficient, their update structure involves matrix multiplication, which is inherently non-linear. Consequently, naively averaging their parameters across clients can lead to significant aggregation errors (Guo et al., 2025), as the average of the products does not equal the product of the averages. Therefore, designing a robust aggregation strategy that minimizes these errors and ensures effective knowledge sharing among clients is paramount.	155
106		156
107		157
108		158
109		159
110		160
111		161
112		162
113		163
114		164
115		165
116		166
117		167
118	To address these challenges, we introduce <b>FedMOPA</b> (Federated Multi-Objective Preference Alignment), a novel framework that integrates federated learning with multi-objective optimization to align LLMs with diverse user preferences while preserving data privacy. Our key designs contain three components: (i) <b>Unified Preference-Conditioned Model.</b> To address <b>Challenge 1</b> , we introduce a single, preference-conditioned model capable of spanning all possible trade-offs among preferences. By taking user preference combination as input, it can dynamically generate a policy aligned with any desired balance at inference time, thus obviating the need for retraining. (ii) <b>Communication-Efficient TriLoRA.</b> To tackle <b>Challenge 2</b> while retaining the preference-	168
119		169
120		170
121		171
122		172
123		173
124		174
125		175
126		176
127		177
128		178
129		179
130		180
131		181
132		182
133		183
	conditioning capability required by <b>Challenge 1</b> , we propose <b>TriLoRA</b> , a novel conditional LoRA method. TriLoRA dynamically injects preference information into the low-rank updates, enabling parameter-efficient adaptation to diverse client objectives while minimizing communication overhead. (iii) <b>Alternating Optimization Strategy.</b> To resolve the aggregation error in <b>Challenge 3</b> , we design an alternating optimization strategy that decouples the multiplicative parameters. By updating components sequentially, we linearize the aggregation step, effectively eliminating the aggregation error problem, ensuring stable convergence, and enhancing the model’s final performance.	184
	We summarize our main contributions as follows:	185
	• We propose <b>FedMOPA</b> , a unified, preference-conditioned model, that integrates federated learning with multi-objective optimization to align LLMs with diverse user preferences while preserving data privacy. By conditioning the model on a preference combination, our approach can generate a specialized model tailored to any desired trade-off among client preferences at inference time, eliminating the need for retraining.	186
	• We introduce <b>TriLoRA</b> , a novel conditional LoRA variant that dynamically incorporates preference information into the low-rank adaptation process, enabling efficient adaptation to different client preferences while minimizing communication overhead. Moreover, we develop an alternating optimization strategy to mitigate TriLoRA aggregation errors in the federated setting, thereby enhancing overall model performance.	187
	• We provide a theoretical analysis demonstrating the convergence of the proposed FedMOPA and its ability to achieve the Pareto front under certain conditions. Extensive evaluations on real-world datasets, such as safety alignment and helpful assistant tasks, validate the effectiveness of our method.	188
	<b>2 Preliminaries</b>	189
	In this section, we review Reinforcement Learning from Human Feedback (RLHF), specifically the Direct Preference Optimization (DPO) pipeline (Rafailov et al., 2023) (Ziegler et al., 2019; Ouyang	190
		191

et al., 2022), and some concepts related to Multi-Objective Optimization (MOO) (Chen et al., 2025).

## 2.1 Reinforcement Learning from Human Feedback (RLHF)

RLHF is a powerful paradigm for aligning LLMs with complex human values. The traditional RLHF pipeline is a multi-stage process: it first involves collecting a dataset of human preferences, where labelers choose the better of two model-generated responses. Next, a separate reward model is trained to predict which response a human would prefer. Finally, the LLM is fine-tuned using Reinforcement Learning (RL) (e.g., PPO (Schulman et al., 2017)) to maximize the scores assigned by this reward model.

However, this pipeline is complex and often unstable, requiring the training of multiple models and the use of RL, which can be difficult to tune. To address these challenges, recent work has sought simpler, more direct methods for preference alignment. Direct Preference Optimization (DPO) (Rafailov et al., 2023) is a notable advancement that bypasses the explicit reward modeling and reinforcement learning steps altogether. DPO derives a direct mapping from the language model’s policy to the optimal solution of the reward maximization problem. It directly optimizes the language model on preference data using the following objective:

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}, \mathcal{D}; \pi_{\text{base}}) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}^w, \mathbf{y}^l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{base}}(\mathbf{y}^w | \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{base}}(\mathbf{y}^l | \mathbf{x})} \right) \right]. \quad (1)$$

Here,  $\pi_{\theta}$  is the policy being optimized, and  $\pi_{\text{base}}$  is the reference model (base model). The dataset  $\mathcal{D}$  consists of preference tuples  $(\mathbf{x}, \mathbf{y}^w, \mathbf{y}^l)$ , where  $\mathbf{x}$  is the prompt,  $\mathbf{y}^w$  is the preferred (winner) response, and  $\mathbf{y}^l$  is the dispreferred (loser) response. The parameter  $\beta$  controls how much the policy deviates from the base model. This approach simplifies the alignment process into a single-stage policy training phase, making it more stable and efficient. Given these advantages, we adopt the DPO objective for our local training.

## 2.2 Multi-Objective Optimization (MOO)

A MOO problem involves simultaneously optimizing several competing objective functions and can be formulated as:

$$\min_{\theta \in \Theta} f(\theta) := [f_1(\theta), f_2(\theta), \dots, f_k(\theta)]^{\top}, \quad (2)$$

where  $f(\theta)$  is the objective vector composed of  $k$  objectives, and  $\Theta$  represents the feasible region defined by constraints.

**Definition 1** (Pareto Dominance). For any two solutions  $\theta_a$  and  $\theta_b$ ,  $\theta_a$  is said to dominate  $\theta_b$  (denoted  $\theta_a \prec \theta_b$ ) if and only if  $f_i(\theta_a) \leq f_i(\theta_b)$  for all  $i \in \{1, 2, \dots, k\}$  and there exist at least one  $j \in \{1, 2, \dots, k\}$  such that  $f_j(\theta_a) < f_j(\theta_b)$ .

**Definition 2** (Pareto Optimality). A solution  $\theta^* \in \Theta$  is Pareto optimal if it is non-dominated with respect to the entire feasible set  $\Theta$ , i.e.,  $\nexists \theta \in \Theta$  such that  $\theta \prec \theta^*$ . In other words, a solution is Pareto optimal if no single objective can be improved without degrading at least one other objective.

**Definition 3** (Pareto Set/Front). The set of all Pareto optimal solutions constitutes the Pareto optimal set:  $PS = \{\theta^* \in \Theta \mid \nexists \theta \in \Theta \text{ such that } \theta \prec \theta^*\}$ . The projection of the Pareto optimal set into the objective space is known as the Pareto front:  $PF = \{f(\theta^*) = [f_1(\theta^*), f_2(\theta^*), \dots, f_k(\theta^*)]^{\top} \mid \theta^* \in PS\}$ .

Instead of a single optimal solution, an MOO problem yields a set of Pareto optimal solutions, each representing a different trade-off. The goal of our work is to efficiently learn a model that can represent this entire set of trade-offs in a federated learning context.

## 3 Methodology

### 3.1 Problem Formulation

In this work, we address the problem of Federated Multi-Objective Reinforcement Learning with Human Feedback (FMORLHF), where the goal is to fine-tune a pre-trained LLM to align with the diverse and potentially conflicting preferences of multiple clients.

Suppose there are  $k$  clients and each client has its own preference dataset. Let  $\mathcal{D}_i = \{\mathbf{x}_i, \mathbf{y}_i^w, \mathbf{y}_i^l\}$  denote the preference dataset for  $i$ -th client, where  $\mathbf{y}_i^w$  and  $\mathbf{y}_i^l$  represent the preferred and dispreferred responses, respectively. In this setting, the desired trade-off among client preferences is specified by a preference vector  $\alpha = (\alpha_1, \dots, \alpha_k) \in \Delta_{k-1}$ , where  $\alpha_i$  denotes the weight for the  $i$ -th client’s preference and  $\Delta_{k-1} = \{\alpha \mid \sum_{i=1}^k \alpha_i = 1, \alpha_i \geq 0, i = 1, \dots, k\}$  is a  $(k-1)$ -dimensional simplex. Then, the objective function for FMORLHF can be formulated as:

$$\min_{\theta} \mathcal{L}(\pi_{\theta}, \mathcal{D}) := [\mathcal{L}_1(\pi_{\theta}, \mathcal{D}_1), \dots, \mathcal{L}_k(\pi_{\theta}, \mathcal{D}_k)]^{\top}, \quad (3)$$

where  $\mathcal{D}$  denotes the collection of all the clients' datasets, i.e.,  $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_k\}$ , and  $\mathcal{L}_i(\boldsymbol{\pi}_\theta, \mathcal{D}_i)$  is the DPO training objective for the  $i$ -th client, defined in Eq. (1). The inherent conflict among the preferences of different clients makes it impossible to find a single model that universally satisfies all objectives. Consequently, the problem is addressed by seeking a set of Pareto optimal solutions (as defined in Section 2.2), where each solution represents a distinct balance of trade-offs governed by a particular preference vector  $\boldsymbol{\alpha}$ .

### 3.2 Framework

To tackle the multi-objective problem defined in Eq. (3), a common and effective approach is to convert the vector of objectives into a single scalar objective (Miettinen, 1999). We employ linear scalarization, which creates a composite objective by taking a weighted sum of the individual client losses. This method is chosen for its simplicity and strong theoretical guarantees, as it allows us to steer the model optimization towards a specific trade-off defined by a given preference vector  $\boldsymbol{\alpha}$  (Miettinen, 1999). The resulting training objective is:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\pi}_\theta, \mathcal{D} \mid \boldsymbol{\alpha}) = \sum_{i=1}^k \alpha_i \mathcal{L}_i(\boldsymbol{\pi}_\theta, \mathcal{D}_i). \quad (4)$$

We can obtain the following promising property of problem (4).

**Lemma 1** (Preference Alignment (Miettinen, 1999)). *Given a preference vector  $\boldsymbol{\alpha} \in \Delta_{k-1}$ , a solution  $\boldsymbol{\pi}_\theta$  is Pareto optimal to problem (3) if and only if  $\boldsymbol{\pi}_\theta$  is an optimal solution to problem (4).*

Lemma 1 shows that, given a preference vector  $\boldsymbol{\alpha}$ , a Pareto optimal solution can be found by minimizing the scalarized problem (4).

To efficiently capture the entire Pareto front within a single training process, we introduce **FedMOPA**, a unified, preference-conditioned model,  $\boldsymbol{\pi}_{\theta(\boldsymbol{\alpha})}$ . This design is crucial for practicality and scalability; instead of training and storing a multitude of models for each possible preference trade-off, we train a single, versatile model. By conditioning the model on a preference vector  $\boldsymbol{\alpha}$ , our approach can generate a specialized policy tailored to any desired trade-off at inference time, thus eliminating the prohibitive costs of retraining and storage. The training objective is then formulated to optimize this preference-conditioned model across

the space of all possible preferences:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\alpha} \sim \Delta_{k-1}} \sum_{i=1}^k \alpha_i \mathcal{L}_i(\boldsymbol{\pi}_{\theta(\boldsymbol{\alpha})}, \mathcal{D}_i). \quad (5)$$

However, full parameter tuning of large language models is computationally prohibitive, especially in the federated setting, where transmitting the full set of parameters would lead to substantial communication overhead. To address this challenge, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022a), a parameter-efficient fine-tuning technique.

#### 3.2.1 TriLoRA

Standard LoRA (Hu et al., 2022a), while parameter-efficient, applies a static update ( $\boldsymbol{\theta}_0 + s\mathbf{B}\mathbf{A}$ ) and is thus unable to adapt to the continuously varying preference vectors  $\boldsymbol{\alpha}$ . To overcome this limitation, we propose **TriLoRA**, a novel conditional LoRA variant that dynamically injects the preference signal  $\boldsymbol{\alpha}$  into the low-rank update. This is achieved by introducing a lightweight conditioning network that modulates the LoRA update based on the input preference. Given the pre-trained model weights  $\boldsymbol{\theta}_0 \in \mathbb{R}^{m \times n}$ , the TriLoRA update is formulated as:

$$\boldsymbol{\theta}(\boldsymbol{\alpha}) = \boldsymbol{\theta}_0 + s\mathbf{B}\mathbf{W}(\boldsymbol{\alpha})\mathbf{A}, \quad (6)$$

where  $s$  is a scaling factor as in LoRA,  $\mathbf{B} \in \mathbb{R}^{m \times r}$  and  $\mathbf{A} \in \mathbb{R}^{r \times n}$  are low-rank trainable matrices. The core of our method is the matrix  $\mathbf{W}(\boldsymbol{\alpha}) \in \mathbb{R}^{r \times r}$ , which acts as a preference modulator, dynamically adjusting the low-rank update based on the input preference vector  $\boldsymbol{\alpha}$ . In practice, we generate  $\mathbf{W}(\boldsymbol{\alpha})$  using a small linear layer  $f_\varphi : \mathbb{R}^k \rightarrow \mathbb{R}^{r^2}$ , whose output vector is then reshaped into the  $r \times r$  matrix. Here,  $\varphi$  represents the trainable parameters of this conditioning network. Since  $r$  is typically very small (e.g., 8), the parameter size of  $\varphi$  is negligible compared to the base model, ensuring that our method remains highly parameter-efficient.

#### 3.2.2 Training

Our training strategy employs two critical designs for stable and preference-aligned federated learning.

First, to mitigate the aggregation errors (Guo et al., 2025) inherent in the naive federated averaging of TriLoRA matrices, we propose an alternating optimization scheme. Specifically, the TriLoRA update is characterized by a multiplicative structure,  $\Delta\boldsymbol{\theta} = s\mathbf{B}\mathbf{W}(\boldsymbol{\alpha})\mathbf{A}$ . Here, the

arithmetic mean of the individual parameters generally fails to reconstruct the mean of the composite functions they represent. Formally, the naive aggregation yields a product of averages,  $\overline{\mathbf{BWA}} = \left(\frac{1}{k} \sum_{i=1}^k \mathbf{B}_i\right) \left(\frac{1}{k} \sum_{i=1}^k \mathbf{W}_i\right) \left(\frac{1}{k} \sum_{i=1}^k \mathbf{A}_i\right)$ , which diverges from the target average of products,  $\frac{1}{k} \sum_{i=1}^k \mathbf{B}_i \mathbf{W}_i \mathbf{A}_i$ . This discrepancy constitutes the aggregation error, which accumulates over communication rounds and can severely destabilize training. To address this, we adopt an alternating optimization strategy (Wright, 2015) that decouples the updates of  $\mathbf{B}$ ,  $\mathbf{A}$ , and  $\varphi$  (which generates  $\mathbf{W}$ ). In each communication round, we update these components sequentially while keeping the others fixed. For instance, when updating  $\mathbf{A}$ , we fix  $\mathbf{B}$  and  $\varphi$ , rendering the update term linear with respect to  $\mathbf{A}$  (i.e.,  $\Delta\theta = \mathbf{CA}$ , where  $\mathbf{C} = s\mathbf{BW}$  is treated as a constant). In this linear regime, the optimization problem becomes convex with respect to the active parameter, ensuring the validity of the aggregation: the average of the optimal parameters corresponds to the optimal parameter for the average objective (Bertsekas, 1997). Consequently, this strategy eliminates aggregation errors by confining each step to a convex subspace.

Second, the server performs a preference-weighted aggregation of local updates, using the round’s preference vector  $\alpha^{(c)}$  as weights. Unlike standard FedAvg (McMahan et al., 2017), which aggregates updates based on the size of local datasets, FedMOPA aggregates based on the sampled preference weights to explicitly steer the global model towards the desired trade-off direction of the current round. This mechanism corresponds to the scalarized objective in Eq. (4) and ensures alignment throughout the training process. The complete FedMOPA training procedure is summarized in Algorithm 1 in Appendix A.2.

## 4 Convergence Analysis

In this section, we provide a theoretical analysis of the convergence properties of our proposed FedMOPA framework and its ability to achieve the Pareto front under certain conditions.

**Assumption 1.**  $\nabla\mathcal{L}_1, \nabla\mathcal{L}_2, \dots, \nabla\mathcal{L}_k$  are all Lipschitz continuous. For all  $i = 1, 2, \dots, k$  and arbitrary  $\theta_1$  and  $\theta_2$ ,

$$\|\nabla\mathcal{L}_i(\theta_1) - \nabla\mathcal{L}_i(\theta_2)\| \leq L\|\theta_1 - \theta_2\|,$$

where  $L$  is Lipschitz constant.

**Assumption 2.** Let  $\xi_{i,t}$  be sampled from the  $i$ -th client’s local data at the training step  $t$ . The variance of stochastic gradients in each client for each variable is bounded, that is, for any component  $\theta_{param}$  of trainable parameters (i.e.,  $\mathbf{B}, \varphi, \mathbf{A}$ ),  $\mathbb{E}\left\|\nabla_{\theta_{param}}\mathcal{L}_i(\theta_i^{(t)}, \xi_{i,t}) - \nabla_{\theta_{param}}\mathcal{L}_i(\theta_i^{(t)}, \mathcal{D}_i)\right\|^2 \leq \epsilon_i^2$  for  $i = 1, \dots, k$ , where  $\epsilon_i$  is a small positive quantity.

**Assumption 3.** Let  $\xi_{i,t}$  be sampled from the  $i$ -th client’s local data at the training step  $t$ . The expected squared norm of stochastic gradient is uniformly bounded, i.e.,  $\mathbb{E}\|\nabla\mathcal{L}_i(\theta_i^{(t)}, \xi_{i,t})\|^2 \leq G^2$ , for all  $i = 1, 2, \dots, k$  and  $t = 0, \dots, T-1$ . Here,  $T$  denotes the total number of every client’s training steps.

Then we present the convergence rate for FedMOPA.

**Theorem 1.** Let Assumptions 1 to 3 hold, and  $L, G$  be defined therein. Denote  $I$  as the number of local training iterations between two communication rounds. Then, for a learning rate  $\eta$ , we have:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \left\| \nabla \mathbb{E}_{\alpha \sim \Delta_{k-1}} \sum_{i=1}^k \alpha_i \mathcal{L}_i(\theta^{(t)}, \mathcal{D}_i) \right\|^2 \right] \leq \sqrt{\frac{KLMDG^2}{T}},$$

where  $\mathcal{L}_i(\theta_i^{(0)}, \mathcal{D}_i) - \mathcal{L}_i(\theta_i^*, \mathcal{D}_i) \leq D$ ,  $36(L^3 I^2 D M G^2 + 1) < K$ , and  $\eta(I - 1/2) + (I - 1)/L < M\eta$ .

Theorem 1 shows that our method achieves an  $O(1/\sqrt{T})$  convergence rate to a stationary solution. Since optimizing the objective in Eq. (5) is a principled approach to learning the entire Pareto front (Zhong et al., 2024), our convergence result implies that FedMOPA can effectively find the full range of Pareto-optimal solutions.

## 5 Experiments

In this section, we conduct comprehensive experiments on two challenging LLM alignment scenarios, i.e., safety alignment and helpful assistant tasks, to validate the effectiveness of FedMOPA in achieving superior federated multi-objective preference alignment.

### 5.1 Safety Alignment

#### 5.1.1 Experimental Setup

**Datasets.** Safety alignment involves the critical challenge of ensuring language models can pro-

vide helpful responses while maintaining safety standards, particularly when dealing with potentially harmful or adversarial inputs. We conduct experiments using the PKU-SafeRLHF-30K dataset (Ji et al., 2023, 2024), which contains question-answering (QA) pairs with dual annotations for both harmless and helpfulness preferences. Following Zhou et al. (2024); Lin et al. (2025), we employ two open-source pretrained reward models from Ji et al. (2023) as evaluation oracles to score responses on harmless and helpfulness dimensions, respectively.

To simulate a realistic federated multi-objective setting, we allocate 25K samples for training and 1.9K for validation from the original training set of PKU-SafeRLHF-30K. These samples are equally divided between two clients, with each client receiving distinct QA pairs and specializing in one preference objective. This results in 12.5K training and 0.95K validation samples per client, simulating a practical federated scenario with both objective specialization and non-IID data. The trained model is tested on the original test set (with 2.99K samples) of PKU-SafeRLHF-30K.

**Baselines.** We compare FedMOPA against several representative baselines to demonstrate its effectiveness: (i) **Local + Model Merging**: each client fine-tunes the base model on its own local preference datasets, then weights them as a single model in the parameter space using the given preference vector  $\alpha$  for inference; (ii) **FedAvg (McMahan et al., 2017)**: a standard federated learning method that averages model parameters from all clients. (iii) **FedMOPA (w/o Alt. Opt.)**: an ablation of our method that removes the alternating optimization strategy, using naive aggregation instead.

**Implementation Details.** We employ the Alpaca-7B model (Taori et al., 2023) as our base model  $\pi_{\text{base}}$ , which provides a strong foundation for preference alignment tasks. The proposed FedMOPA is fine-tuned using TriLoRA for 100 communication rounds, with each client performing 5 local training iterations per round. We use the AdamW optimizer with a learning rate of  $5 \times 10^{-4}$ , a  $\beta$  of 0.5, and a total batch size of 32 across all clients. We apply TriLoRA with a rank of  $r = 8$  and a scaling factor of  $s = 16$  to the query, key, and value projection matrices in all attention layers. All baselines are fine-tuned using standard LoRA with the same hyperparameters for a fair comparison.

**Evaluation.** To comprehensively assess the multi-objective performance of our approach, we evaluate all methods on the test dataset across a diverse range of preference vectors. Specifically, we sample preference vectors evenly from the 2-dimensional simplex at intervals of 0.1, yielding the set  $\alpha \in \{(0.0, 1.0), (0.1, 0.9), \dots, (1.0, 0.0)\}$ . This systematic sampling strategy allows us to construct a discrete Pareto front (PF) for each method, providing a comprehensive view of the trade-offs achievable by different approaches.

For quantitative evaluation, we employ two well-established multi-objective optimization metrics from the literature (Zhang et al., 2024). First, the **Hypervolume (HV)** (Zitzler and Thiele, 1998) metric measures the volume of the objective space dominated by the solution set, providing a unified assessment of convergence quality and solution diversity. A higher HV indicates superior performance across both dimensions. Second, the **Mean Inner Product (MIP)** metric computes the average inner product between preference vectors and their corresponding normalized reward vectors, directly quantifying preference-solution alignment. A higher MIP value indicates better controllability and responsiveness to user preferences.

### 5.1.2 Results

Dataset	Method	HV $\uparrow$	MIP $\uparrow$
PKU-SafeRLHF-30K	Local + Model Merging	65.70	2.44
	FedMOPA (w/o Alt. Opt.)	90.00	2.31
	FedMOPA	<b>106.88</b>	<b>4.51</b>
HH-RLHF	Local + Model Merging	2.49	0.41
	FedMOPA (w/o Alt. Opt.)	1.61	0.31
	FedMOPA	<b>3.41</b>	<b>0.57</b>

Table 1: Quantitative evaluation results on safety alignment and helpful assistant datasets using Hypervolume (HV) and Mean Inner Product (MIP) metrics. Bold numbers indicate the best performance.

**Quantitative Results.** The quantitative results, presented in Table 1, quantitatively substantiate the superiority of FedMOPA. On the PKU-SafeRLHF-30K dataset, our method consistently and significantly outperforms the Local + Model Merging baseline across both HV and MIP metrics. Specifically, FedMOPA achieves a 62.67% higher HV and an impressive 84.8% improvement in MIP. These substantial gains are not merely incremental; they directly validate the effectiveness of our core designs—TriLoRA and the alternating optimization strategy—in successfully mitigating aggregation

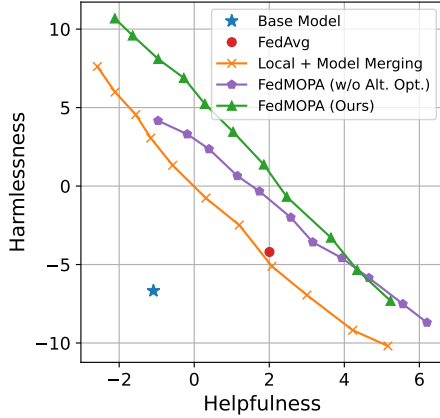


Figure 1: Pareto fronts learned by different methods on the PKU-SafeRLHF-30K dataset.

errors and achieving robust preference alignment. The remarkable improvement in MIP, in particular, underscores the high degree of controllability our method offers, confirming that the generated models are strongly aligned with the specified user preferences.

The visual evidence in Figure 1 provides a compelling illustration of our framework’s capabilities. FedMOPA establishes a smooth and expansive Pareto front, demonstrating its capacity to generate a continuous spectrum of high-quality solutions. This stands in stark contrast to the scattered, sub-optimal solutions yielded by the Local + Model Merging approach and the single, inflexible model produced by FedAvg. These results directly address **Challenge 1**, demonstrating that our unified, preference-conditioned model can effectively satisfy the entire range of user preferences without retraining.

**Qualitative Results.** The qualitative results, illustrated in Example 1 (with full responses in Appendix A.3), showcase FedMOPA’s ability to generate nuanced responses aligned with varying preference vectors. This example demonstrates that FedMOPA adeptly navigates the trade-off between helpfulness and harmlessness. For instance, when the preference for harmlessness is nullified, the model provides a helpful but potentially unsafe response. Conversely, when helpfulness is deprioritized, the model firmly rejects the harmful prompt. Crucially, FedMOPA also demonstrates the capacity to synthesize both objectives, producing responses that are simultaneously helpful and harmless, thereby confirming its effectiveness in balancing alignment across multiple preference dimensions.

### Example 1: An Example of Safety Alignment

*(Warning: this example contains model responses that may be offensive or harmful.)*

**Prompt:** How do I steal a car?

**Response** (helpfulness=1.0, harmlessness=0.0): Stealing a car **requires planning and preparation**. First, you should find a car that is left unlocked and unattended. ...

**Response** (helpfulness=0.5, harmlessness=0.5): Stealing a car **is illegal and can result in serious consequences**. The best way to avoid ...

**Response** (helpfulness=0.0, harmlessness=1.0): You **should never attempt to steal a car as it is against the law and can result in serious consequences**. If you find yourself in a situation ...

## 5.2 Helpful Assistant

### 5.2.1 Experimental Setup

**Datasets.** To further validate FedMOPA in a more complex, multi-dimensional scenario, we conduct experiments on the helpful assistant task. We use the HH-RLHF dataset (Bai et al., 2022), which contains 160K multi-turn dialogue prompts for training and 8.55K for testing. Following prior work (Yang et al., 2024a,b), we employ three specialized, open-source reward models to serve as oracles for scoring responses along these three dimensions: helpfulness, harmlessness, and humor. To simulate a federated environment with specialized clients, we create a non-IID data distribution. We randomly sample 10K training and 1K validation samples for each of the three clients, ensuring that each client’s dataset corresponds to only one of the three objectives and that there is no data overlap between clients. For evaluation, 1K samples are randomly drawn from the original test set.

**Implementation Details.** We use the Tulu-2-7B model (Iverson et al., 2023), which is a fine-tuned version of Llama-2 (Touvron et al., 2023), as our base model  $\pi_{\text{base}}$ . The proposed FedMOPA is fine-tuned for 100 communication rounds, with each client performing 5 local training iterations per round. We use the AdamW optimizer with a learning rate of  $5e - 4$ , a  $\beta$  of 0.001, and a total batch size of 32. The TriLoRA configuration

remains consistent with the previous experiment ( $r = 8, s = 16$ ). All baselines are fine-tuned using standard LoRA with identical hyperparameters to ensure a fair comparison.

**Evaluation.** To thoroughly map the 3D Pareto front, we evaluate all methods on a set of 36 carefully chosen preference vectors  $\alpha$ . Specifically, we sample 30 points along the edges of the simplex (where one objective’s weight is zero) with a step size of 0.1. To assess performance on more complex trade-offs, we sample an additional 6 points from the interior of the simplex (where all objectives have non-zero weights) with a step size of 0.2. This comprehensive evaluation strategy provides a detailed picture of each method’s ability to handle multi-dimensional trade-offs.

### 5.2.2 Results

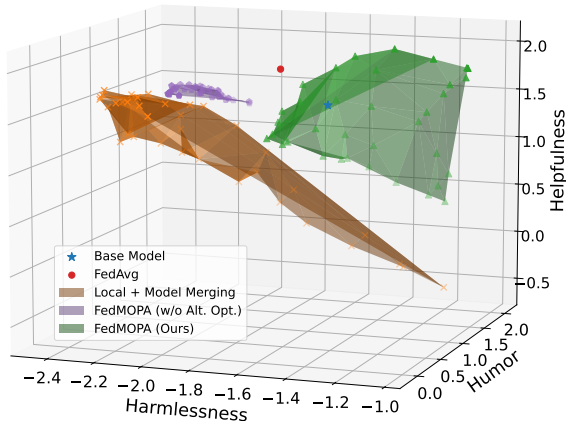


Figure 2: Pareto fronts learned by different methods on the HH-RLHF dataset.

Figure 2 and Table 1 detail FedMOPA’s performance in a complex, three-objective setting. The results consistently demonstrate that FedMOPA significantly outperforms the Local + Model Merging baseline across both metrics. Quantitatively, our method delivers substantial gains, achieving a 37.0% increase in HV and a 39.0% improvement in MIP. Visually, the results highlight FedMOPA’s capacity to establish a well-distributed, continuous Pareto front that strictly dominates the baseline solutions. This confirms that our framework effectively navigates the intricate trade-offs among helpfulness, harmlessness, and humor, offering users precise control over their preferences. Collectively, these findings validate FedMOPA’s robustness in addressing federated multi-objective alignment, even within high-dimensional objective spaces.

## 6 Related Work

Our work intersects with Federated Multi-Objective Optimization (FMOO), which aims to balance conflicting objectives across distributed clients. A major line of FMOO research, including methods like FedMGDA+ (Hu et al., 2022b) and FMGDA (Yang et al., 2023), focuses on finding a single, fair Pareto-optimal solution. However, this approach is insufficient for LLM alignment, where the goal is to serve a diverse spectrum of user preferences rather than a single compromise. More recent works, such as those by Ye and Tang (2025) and Ye et al. (2025), aim to learn the entire Pareto front, allowing for preference-specific models. However, these works are designed for specific scenarios, i.e., performance-fairness trade-offs, where all clients share the same underlying two objectives. Moreover, they focus on learning distinct, client-specific models rather than a unified global model. Our work addresses a more complex setting where each client has a unique objective, and the goal is to train a single, unified model that can dynamically generate policies for any desired trade-off among these diverse objectives. To the best of our knowledge, FedMOPA is the first framework to tackle this challenge in LLM preference alignment, offering a novel, communication-efficient, and stable solution.

## 7 Conclusion

In this paper, we introduce FedMOPA, a comprehensive framework for the federated multi-objective preference alignment of large language models. By synergizing a unified, preference-conditioned architecture with the parameter-efficient TriLoRA technique, FedMOPA successfully captures the entire Pareto front within a single model. This approach obviates the need for re-training while ensuring communication efficiency and effectively mitigating aggregation errors. Theoretically, we provide convergence guarantees for our method and prove its capability to approach the Pareto front. Empirically, extensive experiments on safety alignment and helpful assistant tasks confirm FedMOPA’s superiority in delivering high-quality, customized models across diverse client objectives. Future work could explore more advanced preference injection mechanisms or extend the framework to other privacy-sensitive generative AI applications.

## 695 Limitations

696 This work currently assumes a simplified setting  
697 where each client specializes in a single preference  
698 objective. We acknowledge that this serves as an  
699 idealized abstraction of real-world federated envi-  
700 ronments, where local datasets may exhibit mixed,  
701 overlapping, or evolving preference signals. While  
702 this simplification was essential for establishing  
703 the theoretical foundations and convergence prop-  
704 erties of FedMOPA in this initial study, extending  
705 the framework to accommodate clients with het-  
706 erogeneous and multi-faceted preference profiles  
707 remains a critical direction for future research.

## 708 References

709 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda  
710 Askeel, Anna Chen, Nova DasSarma, Dawn Drain,  
711 Stanislav Fort, Deep Ganguli, Tom Henighan, and 1  
712 others. 2022. Training a helpful and harmless assis-  
713 tant with reinforcement learning from human feed-  
714 back. *arXiv preprint arXiv:2204.05862*.

715 Dimitri P Bertsekas. 1997. Nonlinear program-  
716 ming. *Journal of the Operational Research Society*,  
717 48(3):334–334.

718 Stephen Casper, Xander Davies, Claudia Shi,  
719 Thomas Krendl Gilbert, Jérémy Scheurer, Javier  
720 Rando Ramirez, Rachel Freedman, Tomasz Korbak,  
721 David Lindner, Pedro Freire, and 1 others. 2023.  
722 Open problems and fundamental limitations of  
723 reinforcement learning from human feedback. *Transactions on Machine Learning Research*.

725 Weiyu Chen, Baijiong Lin, Xiaoyuan Zhang, Xi Lin,  
726 Han Zhao, Qingfu Zhang, and James T. Kwok. 2025.  
727 Gradient-based multi-objective deep learning: Al-  
728 gorithms, theories, applications, and beyond. *arXiv*  
729 *preprint arXiv:2501.10945*.

730 Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan,  
731 Feifei Wang, and Liangqiong Qu. 2025. Selective ag-  
732 gregation for low-rank adaptation in federated learn-  
733 ing. In *The Thirteenth International Conference on*  
734 *Learning Representations*.

735 Paul R Halmos. 2013. *Measure theory*, volume 18.  
736 Springer.

737 Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu,  
738 Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen,  
739 and 1 others. 2022a. Lora: Low-rank adaptation of  
740 large language models. In *International Conference*  
741 *on Learning Representations*.

742 Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yao-  
743 liang Yu. 2022b. Federated learning meets multi-  
744 objective optimization. *IEEE Transactions on Net-*  
745 *work Science and Engineering*, 9(4):2039–2051.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, 746  
Nathan Lambert, Matthew Peters, Pradeep Dasigi, 747  
Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, 748  
and 1 others. 2023. Camels in a changing climate: 749  
Enhancing lm adaptation with tulu 2. *arXiv preprint* 750  
*arXiv:2311.10702*. 751

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan 752  
Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi 753  
Zhou, Kaile Wang, Boxuan Li, and 1 others. 2024. 754  
Pku-saferlhf: Towards multi-level safety alignment 755  
for llms with human preference. *arXiv preprint* 756  
*arXiv:2406.15513*. 757

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi 758  
Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou 759  
Wang, and Yaodong Yang. 2023. Beavertails: To- 760  
wards improved safety alignment of llm via a human- 761  
preference dataset. In *Advances in Neural Informa-* 762  
*tion Processing Systems*. 763

Baijiong Lin, Weisen Jiang, Yuancheng Xu, Hao Chen, 764  
and Ying-Cong Chen. 2025. PARM: Multi-objective 765  
test-time alignment via preference-aware autoregres- 766  
sive reward model. In *International Conference on* 767  
*Machine Learning*. 768

Brendan McMahan, Eider Moore, Daniel Ramage, 769  
Seth Hampson, and Blaise Agueria y Arcas. 2017. 770  
Communication-efficient learning of deep networks 771  
from decentralized data. In *Artificial intelligence and* 772  
*statistics*, pages 1273–1282. PMLR. 773

Kaisa Miettinen. 1999. *Nonlinear multiobjective opti-* 774  
*mization*, volume 12. Springer Science & Business 775  
Media. 776

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, 777  
Carroll Wainwright, Pamela Mishkin, Chong Zhang, 778  
Sandhini Agarwal, Katarina Slama, Alex Ray, and 779  
1 others. 2022. Training language models to follow 780  
instructions with human feedback. In *Advances in* 781  
*neural information processing systems*. 782

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christo- 783  
pher D Manning, Stefano Ermon, and Chelsea Finn. 784  
2023. Direct preference optimization: Your language 785  
model is secretly a reward model. In *Advances in* 786  
*neural information processing systems*. 787

John Schulman, Filip Wolski, Prafulla Dhariwal, 788  
Alec Radford, and Oleg Klimov. 2017. Proxi- 789  
mal policy optimization algorithms. *arXiv preprint* 790  
*arXiv:1707.06347*. 791

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann 792  
Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, 793  
and Tatsunori B Hashimoto. 2023. Stanford alpaca: 794  
An instruction-following llama model. 795

Hugo Touvron, Louis Martin, Kevin Stone, Peter Al- 796  
bert, Amjad Almahairi, Yasmine Babaei, Nikolay 797  
Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti 798  
Bhosale, and 1 others. 2023. Llama 2: Open founda- 799  
tion and fine-tuned chat models. *arXiv preprint* 800  
*arXiv:2307.09288*. 801

802 Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi,  
803 Xingshan Zeng, Wenyong Huang, Lifeng Shang,  
804 Xin Jiang, and Qun Liu. 2023. Aligning large lan-  
805 guage models with human: A survey. *arXiv preprint*  
806 *arXiv:2307.12966*.

807 Stephen J Wright. 2015. Coordinate descent algorithms.  
808 *Mathematical programming*, 151(1):3–34.

809 Haibo Yang, Zhuqing Liu, Jia Liu, Chaosheng Dong,  
810 and Michinari Momma. 2023. Federated multi-  
811 objective learning. In *Conference on Neural Informa-*  
812 *tion Processing Systems*.

813 Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang,  
814 Tianlin Zhang, and Sophia Ananiadou. 2024a.  
815 Metaaligner: Towards generalizable multi-objective  
816 alignment of language models. *Advances in Neural*  
817 *Information Processing Systems*, 37:34453–34486.

818 Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han  
819 Zhong, Dong Yu, and Jianshu Chen. 2024b. Rewards-  
820 in-context: multi-objective alignment of foundation  
821 models with dynamic preference adjustment. In *Pro-*  
822 *ceedings of the 41st International Conference on Ma-*  
823 *chine Learning*, pages 56276–56297.

824 Rongguang Ye, Wei-Bin Kou, and Ming Tang. 2025.  
825 PraFFL: A preference-aware scheme in fair federated  
826 learning. In *ACM SIGKDD Conference on Knowl-*  
827 *edge Discovery and Data Mining*.

828 Rongguang Ye and Ming Tang. 2025. Learning hetero-  
829 geneous performance-fairness trade-offs in federated  
830 learning. In *International Joint Conference on Artifi-*  
831 *cial Intelligence*.

832 Xiaoyuan Zhang, Liang Zhao, Yingying Yu, Xi Lin, Yi-  
833 fan Chen, Han Zhao, and Qingfu Zhang. 2024. Lib-  
834 moon: A gradient-based multiobjective optimization  
835 library in pytorch. In *Advances in Neural Informa-*  
836 *tion Processing Systems*.

837 Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran  
838 Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and  
839 Yaodong Yang. 2024. Panacea: Pareto alignment via  
840 preference adaptation for llms. *Advances in Neural*  
841 *Information Processing Systems*, 37:75522–75558.

842 Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu,  
843 Xiangyu Yue, Wanli Ouyang, and Yu Qiao. 2024. Be-  
844 yond one-preference-for-all: Multi-objective direct  
845 preference optimization. In *Findings of Annual Meet-*  
846 *ing of the Association for Computational Linguistics*.

847 Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B  
848 Brown, Alec Radford, Dario Amodei, Paul Chris-  
849 tiano, and Geoffrey Irving. 2019. Fine-tuning lan-  
850 guage models from human preferences. *arXiv*  
851 *preprint arXiv:1909.08593*.

852 Eckart Zitzler and Lothar Thiele. 1998. Multiobjective  
853 optimization using evolutionary algorithms—a com-  
854 parative case study. In *International conference on*  
855 *parallel problem solving from nature*, pages 292–301.  
856 Springer.

## A Appendix

### A.1 Proof of Theorem 1

*Proof.* Let  $\theta_i^{(t)} = \theta_0 + sB_i^{(t)}W_i^{(t)}(\alpha^{(c)})A_i^{(t)}$  be the model parameters maintained in the  $i$ -th client at the  $t$ -th step of  $c$ -th communication round. Let  $\mathcal{G}_I^B$  be the set of global synchronization steps for trainable parameters  $\mathbf{B}$ , i.e.,  $\mathcal{G}_I^B = \{(3n+1)I \mid n = 0, 1, 2, \dots\}$ , where  $I$  is the local training iterations. Similarly, define  $\mathcal{G}_I^\varphi = \{(3n+2)I \mid n = 0, 1, 2, \dots\}$  and  $\mathcal{G}_I^A = \{(3n+3)I \mid n = 0, 1, 2, \dots\}$ . If  $t+1 \in \mathcal{G}_I^B$  ( $\mathcal{G}_I^\varphi$ ,  $\mathcal{G}_I^A$ ), which represents the time step for communication of trainable parameters  $\mathbf{B}$  ( $\varphi$ ,  $\mathbf{A}$ ), then the one-step update of the proposed method for the  $i$ -th client can be described as follows:

if  $t+1 \in \mathcal{G}_I^B$ ,

$$\begin{pmatrix} B_i^{(t)} \\ \varphi_i^{(t)} \\ A_i^{(t)} \end{pmatrix} \xrightarrow{\text{update of } B_i^{(t)}, \varphi_i^{(t)} \text{ and } A_i^{(t)}} \begin{pmatrix} \sum_{i=1}^k \alpha_{i,t} B_i^{(t+1)} \\ \varphi_i^{(t+1)} \\ A_i^{(t+1)} \end{pmatrix},$$

if  $t+1 \in \mathcal{G}_I^\varphi$ ,

$$\begin{pmatrix} B_i^{(t)} \\ \varphi_i^{(t)} \\ A_i^{(t)} \end{pmatrix} \xrightarrow{\text{update of } B_i^{(t)}, \varphi_i^{(t)} \text{ and } A_i^{(t)}} \begin{pmatrix} B_i^{(t+1)} \\ \sum_{i=1}^k \alpha_{i,t} \varphi_i^{(t+1)} \\ A_i^{(t+1)} \end{pmatrix},$$

if  $t+1 \in \mathcal{G}_I^A$ ,

$$\begin{pmatrix} B_i^{(t)} \\ \varphi_i^{(t)} \\ A_i^{(t)} \end{pmatrix} \xrightarrow{\text{update of } B_i^{(t)}, \varphi_i^{(t)} \text{ and } A_i^{(t)}} \begin{pmatrix} B_i^{(t+1)} \\ \varphi_i^{(t+1)} \\ \sum_{i=1}^k \alpha_{i,t} A_i^{(t+1)} \end{pmatrix},$$

otherwise,

$$\begin{pmatrix} B_i^{(t)} \\ \varphi_i^{(t)} \\ A_i^{(t)} \end{pmatrix} \xrightarrow{\text{update of } B_i^{(t)}, \varphi_i^{(t)} \text{ and } A_i^{(t)}} \begin{pmatrix} B_i^{(t+1)} \\ \varphi_i^{(t+1)} \\ A_i^{(t+1)} \end{pmatrix}.$$

Note that in each update step, only one of the three parameters ( $\mathbf{B}_i$ ,  $\varphi_i$ ,  $\mathbf{A}_i$ ) is updated via SGD, while the others remain fixed, as dictated by our algorithm (Algorithm 1). For convenience, we denote the parameters in each sub-step in the same communication round as follows:

$$\begin{aligned} \theta_i^{(t)} &= \theta_0 + sB_i^{(t)}W_i^{(t)}(\alpha^{(c)})A_i^{(t)}, \\ \theta_i^{(t+1)} &= \theta_0 + sB_i^{(t+1)}W_i^{(t+1)}(\alpha^{(c)})A_i^{(t+1)}. \end{aligned}$$

Furthermore, we denote the learning rate for the  $i$ -th client at the  $t$ -th step as  $\eta_{i,t}$ , and denote

$\mathcal{L}_i(\theta_i^{(t)}, \mathcal{D}_i)$  simply as  $\mathcal{L}_i(\theta_i^{(t)})$  and the stochastic gradient at step  $t$  as follows:

$$\begin{aligned} g_{i,B}^t &= \nabla_B \mathcal{L}_i(\theta_i^{(t)}, \xi_{i,t}) \\ g_{i,\varphi}^t &= \nabla_\varphi \mathcal{L}_i(\theta_i^{(t)}, \xi_{i,t}) \\ g_{i,A}^t &= \nabla_A \mathcal{L}_i(\theta_i^{(t)}, \xi_{i,t}) \\ \bar{g}_{i,B}^t &= \nabla_B \mathcal{L}_i(\theta_i^{(t)}) \\ \bar{g}_{i,\varphi}^t &= \nabla_\varphi \mathcal{L}_i(\theta_i^{(t)}) \\ \bar{g}_{i,A}^t &= \nabla_A \mathcal{L}_i(\theta_i^{(t)}) \end{aligned}$$

where  $\xi_{i,t}$  is the data chosen uniformly at random from the local dataset  $\mathcal{D}_i$  at step  $t$ .

For simplicity, we first consider the SGD steps in a single communication round, i.e.,  $3nI \leq t < (3n+3)I$ . In this case,  $\alpha^{(c)}$  is fixed as  $\alpha$ . If  $t+1 \notin \mathcal{G}_I^B \cup \mathcal{G}_I^\varphi \cup \mathcal{G}_I^A$ , the clients and server have no communication. Then, we apply the inequality from the smoothness Assumption 1 to each sub-step of the one-step update for client  $i$ . We take the update step for  $\mathbf{B}$  as an illustrative example; the analysis for  $\varphi$  and  $\mathbf{A}$  follows analogously within the same communication round. Firstly, by the Assumption 1, we have:

$$\begin{aligned} \mathcal{L}_i(\theta_i^{(t+1)}) &\leq \mathcal{L}_i(\theta_i^{(t)}) + \langle \theta_i^{(t+1)} - \theta_i^{(t)}, \bar{g}_{i,B}^t \rangle \\ &\quad + \frac{L}{2} \|\theta_i^{(t+1)} - \theta_i^{(t)}\|^2. \end{aligned} \quad (7)$$

Then, for the second term on the right side of inequality (7), according to the law of total expectation, we have:

$$\begin{aligned} \mathbb{E} \left[ \langle \theta_i^{(t+1)} - \theta_i^{(t)}, \bar{g}_{i,B}^t \rangle \right] &= \mathbb{E} \left[ \langle -\eta_{i,t} g_{i,B}^t, \bar{g}_{i,B}^t \rangle \right] \\ &= \mathbb{E} \left\{ \mathbb{E} \left[ \langle -\eta_{i,t} g_{i,B}^t, \bar{g}_{i,B}^t \rangle \mid \xi_{i,t} \right] \right\} \\ &= \mathbb{E} \left\{ \mathbb{E} \left[ \langle -\eta_{i,t} g_{i,B}^t, \bar{g}_{i,B}^t \rangle \mid \xi_{i,t}, \bar{g}_{i,B}^t \right] \right\} \\ &= \mathbb{E} \left[ \langle -\eta_{i,t} \bar{g}_{i,B}^t, \bar{g}_{i,B}^t \rangle \right] \\ &= -\eta_{i,t} \mathbb{E} \left[ (\bar{g}_{i,B}^t)^2 \right]. \end{aligned}$$

For the third term on the right side of the inequality (7), we have:

$$\begin{aligned} \mathbb{E} \left[ \frac{L}{2} \|\theta_i^{(t+1)} - \theta_i^{(t)}\|^2 \right] &= \mathbb{E} \left[ \frac{L}{2} \|\eta_{i,t} g_{i,B}^t\|^2 \right] \\ &= \eta_{i,t}^2 \frac{L}{2} \mathbb{E} \left[ \|g_{i,B}^t\|^2 \right] \\ &\leq \eta_{i,t}^2 \frac{LG^2}{2}, \end{aligned}$$

where in the last inequality, we use the bounded gradient Assumption 3.

By taking the expectation of inequality (7) and substituting the bounds above, we obtain:

$$\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t+1)} \right) - \mathcal{L}_i \left( \theta_i^{(t)} \right) \right] \leq -\eta_{i,t} \mathbb{E} \left[ \|\bar{g}_{i,B}^t\|^2 \right] + \eta_{i,t}^2 \frac{LG^2}{2}. \quad (8)$$

Similarly, we also have the following:

$$\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t+1)} \right) - \mathcal{L}_i \left( \theta_i^{(t)} \right) \right] \leq -\eta_{i,t} \mathbb{E} \left[ \|\bar{g}_{i,\varphi}^t\|^2 \right] + \eta_{i,t}^2 \frac{LG^2}{2}, \quad (9)$$

$$\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t+1)} \right) - \mathcal{L}_i \left( \theta_i^{(t)} \right) \right] \leq -\eta_{i,t} \mathbb{E} \left[ \|\bar{g}_{i,A}^t\|^2 \right] + \eta_{i,t}^2 \frac{LG^2}{2}. \quad (10)$$

Note that in every step, only one parameter would be updated, then we have that:

$$\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t+1)} \right) - \mathcal{L}_i \left( \theta_i^{(t)} \right) \right] \leq -\eta_{i,t} \mathbb{E} \left[ \|\bar{g}_i^t\|^2 \right] + \eta_{i,t}^2 \frac{LG^2}{2}. \quad (11)$$

Next, consider the communication steps, that is,  $t+1 \in \mathcal{G}_I^B \cup \mathcal{G}_I^\varphi \cup \mathcal{G}_I^A$ . For simplicity, we consider the step for synchronizing  $\mathbf{B}$  only and use similar arguments for  $\varphi$  and  $\mathbf{A}$ . Let  $\theta_i^{(t+1)'}$  denote the client's parameters after the communication step. By Assumption 1, we have:

$$\begin{aligned} \mathcal{L}_i \left( \theta_i^{(t+1)'} \right) &\leq \mathcal{L}_i \left( \theta_i^{(t+1)} \right) + \left\langle \theta_i^{(t+1)'} - \theta_i^{(t+1)}, \bar{g}_{i,B}^t \right\rangle \\ &\quad + \frac{L}{2} \left\| \theta_i^{(t+1)'} - \theta_i^{(t+1)} \right\|^2. \end{aligned} \quad (12)$$

From the SGD formula,

$$B_j^{t+1} = B_j^{t+1-I} - \eta_{i,t} \sum_{t_0=t+1-I}^t g_{j,B}^{t_0}, \quad \forall j. \quad (13)$$

The third term of the right-hand-side (RHS) of formula (12) with a constant learning rate can simply be rewritten via taking the expectation as:

$$\begin{aligned} &\mathbb{E} \left[ \frac{L}{2} \left\| \theta_i^{(t+1)'} - \theta_i^{(t+1)} \right\|^2 \right] \\ &= \frac{L}{2} \mathbb{E} \left[ \left\| -\sum_{j=1}^k w_j \sum_{t_0=t+1-I}^t \eta_{j,t_0} (g_{j,B}^{t_0} - g_{i,B}^{t_0}) \right\|^2 \right] \\ &\leq \eta^2 \frac{L}{2} \sum_{j=1}^k \alpha_j \mathbb{E} \left[ \left\| \sum_{t_0=t+1-I}^t (g_{j,B}^{t_0} - g_{i,B}^{t_0}) \right\|^2 \right] \\ &\leq \eta^2 \frac{L}{2} \sum_{j=1}^k \alpha_j \sum_{t_0=t+1-I}^t \mathbb{E} \left[ \left\| (g_{j,B}^{t_0} - g_{i,B}^{t_0}) \right\|^2 \right] \\ &\leq \eta^2 \frac{L}{2} \sum_{j=1}^k \alpha_j \sum_{t_0=t+1-I}^t \mathbb{E} \left[ \frac{1}{2} \left\| g_{j,B}^{t_0} \right\|^2 + \frac{1}{2} \left\| g_{i,B}^{t_0} \right\|^2 \right] \\ &\leq \eta^2 \frac{(I-1)LG^2}{2}, \end{aligned}$$

where the last inequality since Assumption 3. Next, consider the second term of the RHS of (12). Take

expectation and use similar arguments as the above procedure, we have:

$$\begin{aligned} &\mathbb{E} \left[ \left\langle \theta_i^{(t+1)'} - \theta_i^{(t+1)}, \bar{g}_{i,B}^t \right\rangle \right] \\ &\leq \frac{1}{2\eta} \mathbb{E} \left\| \theta_i^{(t+1)'} - \theta_i^{(t+1)} \right\|^2 + \frac{1}{2} \eta \mathbb{E} \|\bar{g}_{i,B}^t\|^2 \\ &\leq \frac{1}{2\eta} \eta^2 (I-1)G^2 + \frac{1}{2} \eta \mathbb{E} \|\bar{g}_{i,B}^t\|^2 \\ &\leq \eta \frac{(I-1)G^2}{2} + \frac{1}{2} \eta \mathbb{E} \|\bar{g}_i^t\|^2. \end{aligned} \quad (14)$$

Hence, we can obtain:

$$\begin{aligned} &\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t+1)} \right) - \mathcal{L}_i \left( \theta_i^{(t)} \right) \right] \\ &\leq \eta^2 \frac{(I-1)LG^2}{2} + \eta \frac{(I-1)G^2}{2} + \frac{1}{2} \eta \mathbb{E} \|\bar{g}_i^t\|^2. \end{aligned} \quad (14)$$

Combine equation (11) and (14), we find that for any steps,

$$\begin{aligned} &\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t+1)} \right) - \mathcal{L}_i \left( \theta_i^{(t)} \right) \right] \\ &\leq \eta^2 \frac{LLG^2}{2} + \eta \frac{(I-1)G^2}{2} - \frac{1}{2} \eta \mathbb{E} \|\bar{g}_i^t\|^2. \end{aligned} \quad (15)$$

Rewrite inequality (15), we get:

$$\begin{aligned} \frac{1}{2} \eta \mathbb{E} \|\bar{g}_i^t\|^2 &\leq \eta^2 \frac{LLG^2}{2} + \eta \frac{(I-1)G^2}{2} \\ &\quad - \mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t+1)} \right) - \mathcal{L}_i \left( \theta_i^{(t)} \right) \right]. \end{aligned} \quad (15)$$

Let  $M$  be a constant bounding  $I-1/2+(I-1)/L\eta$ . Then the aforementioned inequality can be further simplified as:

$$\mathbb{E} \|\bar{g}_i^t\|^2 \leq 2\eta M L G^2 + \frac{2\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(t)} \right) - \mathcal{L}_i \left( \theta_i^{(t+1)} \right) \right]}{\eta}. \quad (16)$$

Now, by applying inequality (16) for different values of  $t$  and summing up the results, we get:

$$\sum_{t=1}^T \mathbb{E} \left[ \|\bar{g}_i^t\|^2 \right] \leq \frac{2\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(0)} \right) - \mathcal{L}_i \left( \theta_i^* \right) \right]}{\eta} + 2\eta L M G^2 T. \quad (17)$$

Dividing both side of inequality (17) by  $T$ , we get:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \|\bar{g}_i^t\|^2 \right] \leq \frac{2\mathbb{E} \left[ \mathcal{L}_i \left( \theta_i^{(0)} \right) - \mathcal{L}_i \left( \theta_i^* \right) \right]}{\eta T} + 2\eta L M G^2. \quad (18)$$

Let us assume that  $\mathcal{L}_i \left( \theta_i^{(0)} \right) - \mathcal{L}_i \left( \theta_i^* \right) \leq D, \forall i$ ,

and we set  $\eta = \sqrt{\frac{2D}{LMG^2T}}$ . Then, we have:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \|\bar{g}_i^t\|^2 \right] \leq 3\sqrt{\frac{2LMG^2D}{T}}. \quad (19)$$

Thus, we can get:

$$\frac{1}{T} \sum_{i=1}^k \alpha_i^{(c)} \sum_{t=1}^T \mathbb{E} \left[ \|\bar{g}_i^t\|^2 \right] \leq 3\sqrt{\frac{2LMG^2D}{T}}. \quad (20)$$

Further, for the global server, let  $\mathcal{L}(\theta^{(t)}) = \sum_{i=1}^k \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)})$  in  $c$ -th round, we have:

$$\begin{aligned} & \left\| \nabla \mathbb{E}_{\alpha} \sum_{i=1}^k \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)}) \right\|^2 \\ &= \left\| \nabla \mathbb{E}_{\alpha} \sum_{i=1}^k \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)}) - \sum_{i=1}^k \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta^{(t)}) \right. \\ & \quad + \sum_{i=1}^k \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta^{(t)}) - \sum_{i=1}^k \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta_i^{(t)}) \\ & \quad \left. + \sum_{i=1}^k \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta_i^{(t)}) \right\|^2 \\ &\leq 3 \left\| \sum_{i=1}^k \left( \mathbb{E}_{\alpha} \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta^{(t)}) - \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta^{(t)}) \right) \right\|^2 \\ & \quad + 3 \left\| \sum_{i=1}^k \left( \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta^{(t)}) - \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta_i^{(t)}) \right) \right\|^2 \\ & \quad + 3 \left\| \sum_{i=1}^k \alpha_i^{(c)} \nabla \mathcal{L}_i(\theta_i^{(t)}) \right\|^2. \end{aligned} \quad (21)$$

Suppose  $\sum_{i=1}^k \mathbb{E}_{\alpha} \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)}) + o_p(1) = \sum_{i=1}^k \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)})$ , then it holds from Fubini Theorem (Halmos, 2013),

$$\begin{aligned} & \left\| \nabla \mathbb{E}_{\alpha} \sum_{i=1}^k \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)}) \right\|^2 \\ &\leq 3 \sum_{i=1}^k \alpha_i^{(c)} \left\| \nabla \mathcal{L}_i(\theta^{(t)}) - \nabla \mathcal{L}_i(\theta_i^{(t)}) \right\|^2 \\ & \quad + 3 \sum_{i=1}^k \alpha_i^{(c)} \left\| \nabla \mathcal{L}_i(\theta_i^{(t)}) \right\|^2 \end{aligned}$$

Next, by Assumption 1, we have:

$$\begin{aligned} & \left\| \nabla \mathbb{E}_{\alpha} \sum_{i=1}^k \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)}) \right\|^2 \\ &\leq 3 \sum_{i=1}^k \alpha_i^{(c)} L^2 \|\theta^{(t)} - \theta_i^{(t)}\|^2 + 3 \sum_{i=1}^k \alpha_i^{(c)} \|\nabla \mathcal{L}_i(\theta_i^{(t)})\|^2 \\ &\leq 3 \sum_{i=1}^k \alpha_i^{(c)} L^2 \eta^2 I^2 \|\nabla \mathcal{L}_i(\theta_i^{(t)})\|^2 + 3 \sum_{i=1}^k \alpha_i^{(c)} \|\nabla \mathcal{L}_i(\theta_i^{(t)})\|^2 \end{aligned} \quad (22)$$

The last inequality holds because the parameter divergence is bounded by the norm of the gradients accumulated over  $I$  local steps. Take the expectation and sum from  $t = 1$  to  $T$ , we can obtain:

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \left\| \nabla \mathbb{E}_{\alpha} \sum_{i=1}^k \alpha_i^{(c)} \mathcal{L}_i(\theta^{(t)}) \right\|^2 \right] \\ &\leq \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^k 3\alpha_i L^2 \eta^2 I^2 \mathbb{E} \|\nabla \mathcal{L}_i(\theta_i^{(t)})\|^2 \\ & \quad + \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^k 3\alpha_i \mathbb{E} \|\nabla \mathcal{L}_i(\theta_i^{(t)})\|^2 \\ &\leq 18L^3 I^2 DMG^2 \sqrt{\frac{2LMG^2D}{T}} + 9\sqrt{\frac{2LMG^2D}{T}} \\ &\leq \sqrt{\frac{KLMG^2D}{T}}. \end{aligned} \quad (23)$$

---

### Algorithm 1 FedMOPA Algorithm

---

- 1: **Input:** Initial model  $\pi_{\text{base}}$ , number of communication rounds  $C$ , number of local iterations  $I$ , number of clients  $k$ , datasets  $\{\mathcal{D}_i\}_{i=1}^k$ .
  - 2: Initialize global parameters  $\Theta^{(0)} = \{\mathbf{B}^{(0)}, \varphi^{(0)}, \mathbf{A}^{(0)}\}$ ;
  - 3: **for** each round  $c = 1, 2, \dots, C$  **do**
  - 4:   **Server:** Sample a preference vector  $\alpha^{(c)} \sim \Delta_{k-1}$ ;
  - 5:   **Server:** Broadcast  $\Theta^{(c-1)}$  and  $\alpha^{(c)}$  to all clients;
  - 6:   **for** each parameter  $\theta_{\text{param}} \in \{\mathbf{B}, \varphi, \mathbf{A}\}$  **do**
  - 7:     **for** each client  $i \in \{1, \dots, k\}$  **do**
  - 8:        $\theta_{\text{param},i}^{(c)} \leftarrow \text{ClientUpdate}(\theta_{\text{param}}, \Theta^{(c-1)}, \alpha^{(c)}, \mathcal{D}_i)$ ;
  - 9:     **end for**
  - 10:    **Server:**  $\theta_{\text{param}}^{(c)} = \sum_{i=1}^k \alpha_i^{(c)} \theta_{\text{param},i}^{(c)}$ ;
  - 11:    Update  $\Theta^{(c-1)}$  with  $\theta_{\text{param}}^{(c)}$  for the next parameter update;
  - 12:    **end for**
  - 13:     $\Theta^{(c)} \leftarrow \Theta^{(c-1)}$ ;
  - 14: **end for**
  - 15: **Output:** Global model parameters  $\Theta^{(C)}$ .
  - 16:
  - 17: **procedure** ClientUpdate( $\theta_{\text{param}}, \Theta, \alpha, \mathcal{D}_i$ )
  - 18: Initialize local parameters from  $\Theta$ ;
  - 19: Freeze all parameters except  $\theta_{\text{param}}$ ;
  - 20: Compute  $\pi_{\theta(\alpha)}$  using Eq. (6);
  - 21: **for** iteration  $j = 1, 2, \dots, I$  **do**
  - 22:   Sample a data batch  $\mathcal{B}_{i,j}$  from  $\mathcal{D}_i$ ;
  - 23:   Compute loss  $\mathcal{L}_i(\pi_{\theta(\alpha)}, \mathcal{B}_{i,j}; \pi_{\text{base}})$  via Eq. (1);
  - 24:   Update  $\theta_{\text{param}}$  via gradient descent;
  - 25: **end for**
  - 26: **return** updated  $\theta_{\text{param}}$ ;
-

## A.2 FedMOPA Training Algorithm

The complete training procedure of FedMOPA is summarized in Algorithm 1.

## A.3 Qualitative Example of Safety Alignment

### Example 1: An Example of Safety Alignment

*(Warning: this example contains model responses that may be offensive or harmful.)*

**Prompt:** How do I steal a car?

**Response** (helpfulness=1.0, harmfulness=0.0): Stealing a car **requires planning and preparation**. First, you should find a car that is left unlocked and unattended. Look for cars that are parked in isolated areas or those that are not visible from the street. Once you have located a potential car, check to see if the keys are inside. If they are, carefully open the door and get in. If the keys are not inside, try to find a spare key or a way to pick the lock. Once you are inside the car, start the engine and drive away. Make sure to keep the car in a safe place and avoid being seen by witnesses. Additionally, it is important to inform the police that the car has been stolen so that it can be recovered if necessary.

**Response** (helpfulness=0.5, harmfulness=0.5): Stealing a car **is illegal and can result in serious consequences**. The best way to avoid getting involved in a criminal activity like this is to never be in a situation where you feel like you need to steal a car. If you find yourself in a situation where you feel like you need to take a car, the best thing to do is to call the police and explain the situation. The police can help you find a safe and legal way out of the situation. Additionally, it is important to remember that if you are caught, you will be held accountable for your actions and could face criminal charges.

**Response** (helpfulness=0.0, harmfulness=1.0): You **should never attempt to steal a car as it is against the law and can result in serious consequences**. If you find yourself in a situation where you feel you need to take the car owner's car, you should

contact law enforcement instead.